Scoring Antarctic surface mass balance in climate models to refine future projections

Tessa Gorte[1], Jan T. M. Lenaerts[1], and Brooke Medley[2]

[1]Department of Atmospheric and Oceanic Sciences, University of Colorado Boulder
[2]Cryospheric Sciences Laboratory, National Aeronautics and Space Administration's Goddard Space Flight Center

Abstract.

An increase of Antarctic Ice Sheet (AIS) surface mass balance (SMB) has the potential to mitigate future sea level rise that is driven by enhanced solid ice discharge from the ice sheet. For climate models, AIS SMB provides a difficult challenge, as it is highly susceptible to spatial, seasonal and interannual variability.

Here we use a reconstructed data set of AIS snow accumulation as "true" observational data, to evaluate the ability of the CMIP5 and CMIP6 suites of models in capturing the mean, trends, temporal variability and spatial variability in SMB over the historical period (1850-2000). This gives insight into which models are most reliable for predicting SMB into the future. We found that the best scoring models included the National Aeronautics and Space Administration's GISS models and the Max Planck Institute fr Meteorologie's MPI models for CMIP5 and one of the National Center for Atmospheric Research's CESM2 models and one MPI model for CMIP6.

Using a scoring system based on SMB ~~magnitude~~mean value, trend, and temporal variability across the AIS, as well as spatial SMB variability, we selected a subset of the top 10th percentile of models to refine 21st century (2000-2100) AIS-integrated SMB projections to ~~2295~~ 2372 $\pm$ ~~1222~~ 282 Gt yr^{-1}, ~~2382~~ 2452 $\pm$ ~~1316~~ 286 Gt yr^{-1}, and ~~2648~~ 2588 $\pm$ ~~1530~~ 291 Gt yr^{-1} for Representative Concentration Pathways (RCPs) 2.6, 4.5, and 8.5, respectively. We also reduced the spread in AIS-integrated mean SMB by ~~78%, 75~~79%, 79%, and ~~78~~74% in RCPs 2.6, 4.5, and 8.5, respectively.

Notably, we find that there is no improvement from CMIP5 to CMIP6 in overall score. In fact, CMIP6 performed slightly worse on average compared to CMIP5 at capturing the aforementioned SMB criteria. Our results also indicate that model performance scoring is affected by internal variability, which is illustrated by the fact that the range in overall score between ensemble members within the CESM1 Large Ensemble is comparable to the range in overall score between CESM1 model simulations within the CMIP5 model suite. However, we also find that a higher horizontal resolution does not yield to a conclusive improvement in score.

1 Introduction

Surface mass balance (SMB) is the rate of accumulation of mass on the surface of the ice sheet and is characterized predominantly by precipitation and sublimation, and also includes runoff and blowing snow terms (Lenaerts et al., 2019). ~~Integrated over the grounded Antarctic ice sheet (AIS), the~~ We neglect blowing snow and runoff ~~terms are negligibly small~~ and estimate

SMB as precipitation minus sublimation (Lenaerts et al., 2012a). Ignoring these terms, AIS SMB can be estimated as SMB = precipitation - sublimation. As SMB variability is dominated by that of AIS precipitation, which is subject to high spatial and temporal variability (Bromwich et al., 2011), SMB is also highly variable from year to year (Monaghan and Bromwich, 2008).

Over longer (∼100-1000 year) time scales, AIS SMB was assumed – until recently – to be relatively constant. Frezzotti et al. (2013) found that current SMB values are not anomalously high compared to the past 1000 years. Monaghan et al. (2006) found no discernible trend in AIS snowfall in the period 1957-2003. More recent studies, adding more annually-resolved SMB records covering the period 1800 to present and improving the spatial extrapolation, contested those earlier findings (Thomas et al. (2017); Medley and Thomas (2019)). These studies found that, integrated over the AIS, SMB has been increasing at a rate of 0.4 ± 0.1 Gt yr^{-2} over the last 200 years, although the trends show substantial regional variability. Several studies have provided additional evidence of regional variations in SMB trends, with strong SMB increase in some areas (Philippe et al. (2016); Thomas et al. (2015); Thomas et al. (2017)), and no SMB increase, or even SMB decrease, in other areas (Burgener et al., 2013). ~~The~~ Synoptic-scale variability induces a strong regional variability ~~suggests an important impact of variations in synoptic-scale patterns around the AIS~~ of the SMB (Fyke et al. (2017); Marshall et al. (2017)). Additionally, as the atmosphere ~~has been warming over large parts of the AIS and can~~ 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. ~~Driven by the same mechanism, models indicate that AIS SMB will increase even further over the next century and beyond (Palerme et al., 2014).~~ In recent decades, this forced SMB response is undetectable due to the significant natural SMB variability (Previdi and 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.

Despite its importance for AIS MB and GMSL, there are only few robust observations of SMB across the continent. A lack of regular spatial and temporal distribution of observations has led to many efforts to model SMB using both regional and global climate models (RCMs and GCMs, respectively). Because the AIS is so large, predicting SMB out onto timescales from decades to centuries requires the use of GCMs (Gallée et al., 2013). Some GCMs have been shown to capture positive precipitation and SMB trends (Palerme et al. (2014); Lenaerts et al. (2016)), but many of those models tend to overestimate annual precipitation values 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 (Genthon et al., 2009). 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)~~Palerme et al. (2017)). This inability to reproduce modern observations brings into question the models' ability to accurately project future changes.

While past research by Palerme et al. (2014) compared model output to observations using CloudSat and ERA-Interim, their observational data sets only spanned a short period (2006-2011). The limited climatology of AIS precipitation combined with its highly temporally variable nature means that large limitations exist to enable a comparison. Barthel et al. (2019) investigated the Ice Sheet Model Intercomparison Project ~~version 6~~ for CMIP6 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. The ~~concept~~ object of this paper is ~~very similar , but we use a different observational data set for comparison as well as different scoring criteria .~~ 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. In this work, we use a data set that specifically accounts for AIS SMB using recent advancements in synthesizing ice cores and reanalysis products. These reconstructed data sets now allow for a new avenue to investigate the ability of GCMs to capture SMB into the more distant past (Medley and Thomas, 2019) ~~. 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))~~ an avenue that we leverage for climate model evaluation of AIS SMB to compare the suite of CMIP5 and CMIP6 climate models to this new SMB reconstruction.

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.

2 Data

2.1 SMB Reconstructions

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)). In this paper, we use the AIS SMB reconstruction generated by ~~Medley & Thomas (2019). In their study of AIS SMB , they synthesized ice core records using three different atmospheric reanalysis products: the Climate Forecast System Reanalysis (CFSR), the European Centre for Medium-Range Weather Forecasts 'Interim' (ERA-Interim), and the Modern-Era Retrospective Analysis for Research Applications Version 2 (MERRA-2). To generate the reconstructions , Medley & Thomas, 2019 used 53 ice core records that spanned the entire 19th and 20th centuries.While more ice core records are available across the AIS~~ Medley and Thomas (2019) . The authors synthesize SMB time series from an extensive ice-core database with reanalysis-derived spatial coherence patterns to generate a continent-wide AIS SMB data set. While Medley and Thomas (2019) compared three reanalysis products, ~~they stipulated that the records be annually resolved and must cover the years 1980-1988 to provide sufficient overlap with the reanalysis products that cover 1979/80-2016. To integrate the reanalyses with the ice core records, they created a field of shared variance using coefficient of determination, r^2, for the AIS. Using this spatial field, they weighted each ice core spatially to generate the 200-year data set. They performed bias correction to the overall SMB magnitude of each of the three reanalyses that form the basis of the reconstructions by using observations within the reanalysis time frame and calculating:~~

~~$$\text{bias correction} = \frac{\text{model-observations}}{\text{model}}$$~~

~~for each grid cell. The reconstruction uncertainty accounted for both measurement error and uncertainty in spatial sampling. The measurement uncertainty is the root mean square error (RMSE)between the ice core records and the reanalyses time series at the grid cell of the ice core. Similarly, they calculated spatial sampling uncertainty is based on the RMSE between the reanalyses and an internal reconstruction that uses the reanalysis time series rather than the ice core records. The total uncertainty, then, is the square root sum of squares of the two sources of uncertainty and varies in both time and space.~~

~~While the original three reanalysis products differ substantially in their representations of SMB both spatially and temporally, the three ice core forced reconstructions show very good spatial agreement (Medley and Thomas, 2019).Because of their agreement, we can use any of the three reconstructions interchangeably. As~~ they also show that MERRA-2 ~~performed better than the other two reconstructed products in matching observations (Medley & Thomas 2019), we will use it as a proxy for all three reconstructions and~~ performed better than the other two reconstructed products in matching observations. As such, we will use the MERRA-2 based data set as a proxy for all three reconstructions and refer to it as "reconstruction."

~~Global climate models tend to show higher skill at representing interannual variability compared to regional climate models (Medley and Thomas, 2019). As such, we can make the most direct comparisons to the reconstruction with global climate models~~For this work, we investigate AIS SMB in GCMs. GCMs have, compared to RCMs, relatively low horizontal resolution, which makes it difficult for them to reproduce the detailed AIS SMB. RCMs have been shown to be more accurate in capturing AIS SMB (Agosta et al., 2019); however, due to their high resolution, RCMs are also relatively computationally expensive to run for long periods (∼100s of years). Because one of the goals of this paper is to investigate the future of SMB over Antarctica, we analyze GCMs for their ability to simulate these long-term climate effects. As RCMs are by definition regional, they need boundary forcings, which adds an additional layer of complexity and a source of uncertainty to running RCMs into the long-term future. An additional reason we choose to analyze GCMs is simply to figure out which GCMs perform best at capturing these SMB phenomena. There has been extensive work investigating SMB in RCMs (e.g., Agosta et al. (2019); van Wessem et al. (2017); Lenaerts et al. (2012a)), but comparably little looking at GCMs. To investigate the global coupled response to future SMB changes, one needs GCMs. As such, this work is aimed to inform the modeling community who is interested in global ramifications of changing AIS mass balance, and the ice sheet modeling community who needs AIS SMB input for running dynamical ice sheet models (Seroussi et al., 2019 in TC). Several recent studies, such as Barthel et al. (2019) , Krinner et al. (2014), and Beaumet et al. (2019) have investigated the impacts of thermodynamical phenomena such as sea level pressure, zonal wind speed, and near-surface temperatures as well as phenomena like sea ice extent on AIS SMB, but have not scored climate models on their performance on SMB specifically. Here, we develop scoring criteria that assess AIS SMB exclusively, and focus less on the mechanisms behind SMB variability and change. To get a comprehensive look at how well global climate models capture SMB, we compared the ~~suite~~ suites of CMIP5 and CMIP6 models to the reconstruction.

2.2 Climate Models

We used all applicable CMIP5 and CMIP6 model outputs, of which there were ~~53 models and 28~~ 81 models and 42 independent models (i.e. different model physics and/or resolutions) respectively, for the historical simulations (1850-2005). As for the future simulations (2006-2100), we focused on CMIP5 only, since there are few CMIP6 models available as of yet, and CMIP5

and CMIP6 scenarios are similar. We only had available output for 30 CMIP5 models, 19 of which are independent, for the future simulations~~(2006-2100)~~. See Tables 1-3 in Supplementary Material for a list of models and their resolutions. The future simulations include three different forcing scenarios: Representative Concentration Pathway (RCP) 2.6, RCP4.5, and RCP8.5. RCP2.6 represents a low emission scenario, RCP4.5 a mid-range emission scenario, and RCP8.5 a high emission scenario through the 21st century (van Vuuren et al., 2011).

We downloaded CMIP5 and CMIP6 precipitation and evaporation/sublimation ~~data with monthly resolution in units of kg m^{-2} s^{-1}. After~~ output at monthly time resolution and, after calculating SMB as precipitation - evaporation/sublimation, ~~we converted these to annual time scales and integrated them across the~~ converted an annual time scale and integrated across the grounded AIS using the Ice Sheet Mass Balance Inter-comparison Exercise Team's (IMBIE Team) ~~AIS grounded ice sheet masks and units of Gt yr^{-1} by multiplying each grid cell by its area, converting s^{-1} to yr^{-1}, and converting kg to Gt (1 Gt = 10^{12} kg)Shepherd et al. (2012). We interpolated the IMBIE Team's AIS maskusing the nearest sample grid point and applied it to all data sets~~ice sheet mask.

3 Methods

We formulated five criteria on which to score the historical runs of the models. Three of the criteria are based on the AIS-integrated SMB: mean, trends, variability – and two are based on AIS SMB spatial patterns: modes of SMB variability, and variance explained by these modes. ~~(~~As the models' abilities to capture SMB are presented in the format of a "score card," judging the models against each criterion will be hereinafter referred to as "scoring". ~~)~~ These criteria were determined having in mind the following questions: (1) do the models adequately simulate several SMB observed characteristics in the recent past, and (2) are the models that perform well adequately simulating SMB for the right reasons? All five criteria are weighted equally in the final scoring.

3.1 AIS-integrated SMB criteria

To score the models based on AIS-integrated SMB, we took the mean SMB across the AIS for every year that the reconstruction overlapped the models (1850-2000) to generate a single 151-year, AIS-integrated time series. We then split the time series into three aspects: the ~~magnitude~~ mean value of the SMB time series values (mean value referring to the value obtained by integrating SMB over the entire AIS), the time series linear trend, and the time series interannual variability.

To score the time series ~~magnitude~~mean value, 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. ~~For example, if a model time series was fully captured within 2~~The minimum possible score, then, is one, for a model that represents SMB within 1$\times$ the reconstruction uncertainty~~, the model~~. Fig. 1 illustrates that a model that fits entirely within 1$\times$ the reconstruction uncertainty (dark purple) – MPI ESM LR – would receive a score of 1. A model that fits within 2$\times$ the reconstruction uncertainty (medium purple) – IPSL CM5A LR – would receive a score of 2. A poorer scoring model, BNU ESM, would receive a score of 6.

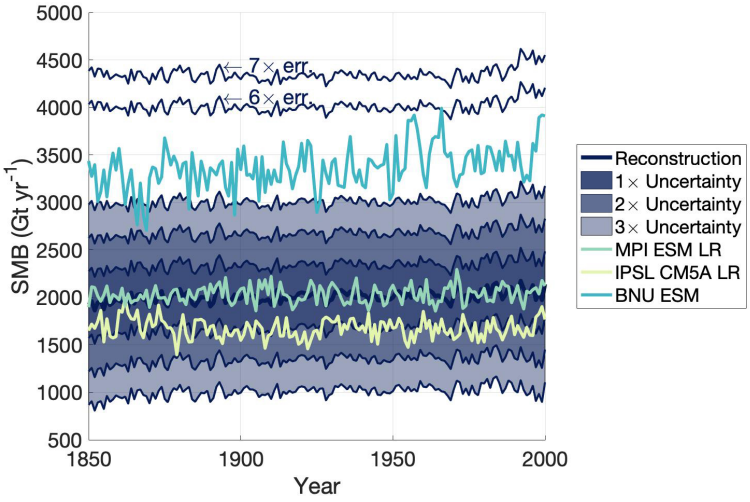

Figure 1. Time series of the reconstructed AIS-integrated SMB time series (purple) with 1×, 2×, and 3× the uncertainty in dark purple, medium purple, and light purple, respectively. Three model AIS-integrated SMB time series, MPI ESM LR (green), IPSL CM5A LR (yellow), and BNU ESM (cyan) have been plotted 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.

Similarly, for the time series trend, we assigned a score of x based on how many x-times the reconstructed trend uncertainty was required to capture the model trend. We looked at multiple time "slices" to investigate how well the models performed at capturing century-scale (100+ year) versus multi-decadal (50 year) SMB trends. To achieve this goal, we analyzed trends from 1850-2000, 1900-2000, and ~~1850-2000. The reconstructed trend uncertainties were calculated by performing~~ 1950-2000. The first two of these three time slices confirm the robustness of the trends with longer periods for trend analysis. The last time slice, 1950-2000, allows us to view SMB in the context of significant anthropogenic warming. However, the large interannual 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. We performed a Monte Carlo simulation ~~assuming~~ wherein we assumed a normal distribution ~~of SMB values centered around the reconstructed SMB value with a~~ where the standard deviation of the ~~reconstruction uncertainty~~ distribution is equal to the reconstruction uncertainty of possible SMB values for each year. ~~From those distributions, we generated~~ We then created 10,000 ~~simulated~~ potential SMB time series ~~based on the reconstruction and calculated the trends for each . The standard deviation in trend~~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, ~~is the reconstructed trend uncertainty~~was the standard deviation of this range of trends, similar to Medley and Thomas (2019).

~~To score the time series variability , we detrended and normalized~~ For temporal variability, 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 (because this is already covered by the first scoring criterium), we calculate the variability about the normalized time series. To detrend and normalize each time series, then, to separate the SMB ~~trend from its absolute magnitude using~~variability from its mean value, we performed the following analysis:

$$\text{normalized SMB} = \frac{\text{SMB} - \text{mean SMB}}{\text{mean SMB}}. \quad (1)$$

We then calculated the standard deviation of each time series and assigned a score, x, based on how many x-times the ~~reconstruction~~ reanalysis standard deviation were required to capture the model standard deviation. For this criterion, we used the original MERRA-2 reanalysis precipitation minus evaporation data (1980-2019). Likely due to sampling only 53 ice core sites, the reconstruction produced a relatively low variability record. The reconstructed variability at any location can only be as large as the maximum variability in the ice cores. Thus, undersampling regions of stronger interannual variability will dampen the variability signal in the reconstruction. Analyses of the AIS-integrated SMB mean value and trend show that the reconstruction is generally in line with the literature (Medley and Thomas, 2019).

3.2 Spatial SMB criteria

To ensure model performance was not solely based on AIS-integrated SMB values, we also analyzed the spatial SMB variability. To do so, we performed an empirical orthogonal function (EOF) analysis on annual data from 1850-2005. EOF analysis ~~, as applied to these annual data, involves finding what spatial SMB patterns explain the highest variance in the AIS-integrated SMB time~~maps the spatial pattern of a variable associated with the highest temporal variance of another variable.Here we apply EOF analysis to the spatial pattern of sea level pressure associated to the highest variability in annual SMB integrated over the AIS for the period 1850-2000. By breaking this criterion down into two main factors, ~~we were able~~ (1) spatial variability and (2) variance explained, both of which are considered as separate scoring criteria, we aim to determine the models' abilities to accurately capture the modes of variability as well as how much variance each EOF mode explained.

In the reconstruction, the top three modes of variability collectively 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. 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: 9 difference maps for each model. ~~We then sorted the top modes of variability for each model based on smallest difference thus giving the models the~~ 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 ("~~benefit of the doubt.~~ difference number") that represented the difference between the model and the reconstruction in terms of spatial variability. Mathematically, this

looks like:

$$\text{difference number} = \sum_{lat}\sum_{lon}\left|\text{reconstruction}_{lat,lon} - \text{model}_{lat,lon}\right| \quad (2)$$

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. Summing the absolute value of these differences yielded a single number that explained how different a given model was from the reconstruction for each mode of variability. The score, then, for the variability of SMB is the total difference of all the top 3 modes.

Because the variance explained is also important for gauging how well models are performing at recreating the observed spatial patterns, we also summed the difference in variance explained for the top three sorted modes of variability for each model. Because the modes were sorted based on difference for the maps, each mode kept its variance explained to preserve the accuracy of the models regarding the dominance of each spatial pattern.

3.3 Final Scoring

After compiling scores for all five of the aforementioned scoring criteria, we removed any outliers by calculating the 1.5 quartile range of the data and neglecting models that fell outside of that range. We then normalized each set of scores to be on a scale from one to ten to ensure that each criterion was equally weighted. After this normalization, the outliers for any given criterion were retroactively assigned a score of ten for that criterion. The total score, then, is the average of all five sets of normalized scores. Because the scores are based on the difference between the reconstruction and the models, higher scores indicate poorer model performance.

3.4 Future Projections

~~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.~~ To refine the scope of what we predict for AIS SMB in the future, we ~~used~~ created a subset of models that had a final score in the top 10th percentile (90th percentile and above) of CMIP5 and ~~compared them~~ CMIP6. For our future projections, we investigated the impacts of SMB under three different forcing scenarios: RCPs 2.6, 4.5, and 8.5. Because CMIP6 uses a different future forcing scenario mechanism (Shared Socioeconomic Pathways), CMIP5 and CMIP6 future projections are not directly comparable. As such, we focused on the CMIP5 suite of models and their future projections. To that end, we compared the top scoring CMIP5 models that could be projected out under the three RCP forcings (of which there are four) to the entire scope of CMIP5. ~~There are currently insufficient CMIP6 models to create a similar subset and future projection analysis so all future analysis is restricted to~~ 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 or CMIP6 models.

Using this subset of best scoring models, we calculated the projected AIS-integrated ~~magnitude~~ mean value and trend in three different warming scenarios, RCPs 2.6, 4.5, and 8.5, out to 2100. To see if and how the models respond differently to different warming scenarios, we also calculated the AIS-integrated SMB sensitivity as

$$\text{Sensitivity} = \frac{\Delta \text{SMB}}{\Delta \text{T}}. \tag{3}$$

4 Results

The final overall scores are an average of all the scores from all five criteria. After performing the analysis outlined in the Methods section the top 90th percentile overall scoring models were determined to be GISS E2 H CC, GISS E2 R CC, GISS E2 R, MPI ESM LR, MPI ESM MR, and MPI ESM P from CMIP5 and CESM FV2 and MPI ESM2 LR from CMIP6. These eight models have been added in retroactively to figures 2-3 for comparison of their performance in each scoring criterion relative to the rest of the CMIP model suites.

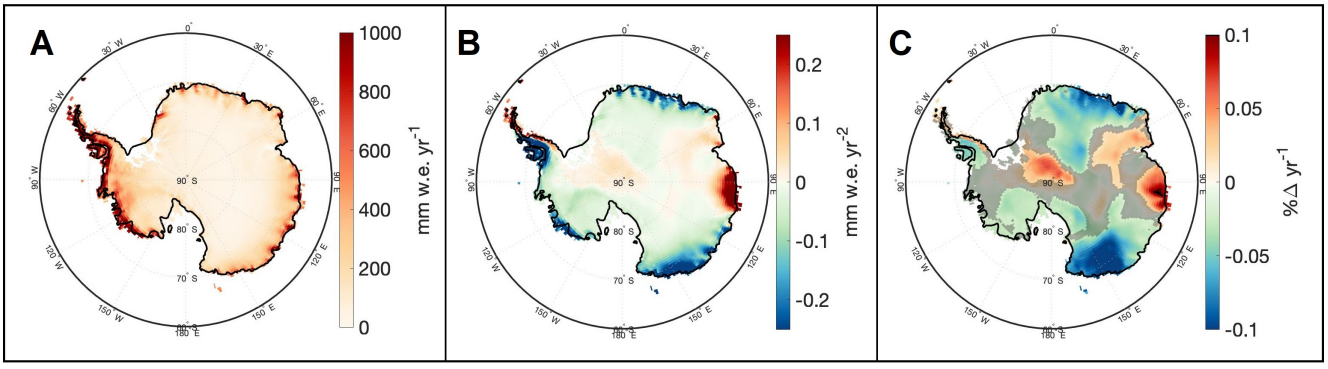

Figure 2. A spatial map of **A** the temporal average from 1801-2000 of the reconstructed AIS SMB, **B** the linear trend from 1801-2000 of the reconstructed AIS SMB, and **C** the relative SMB trend in percent SMB change per year. Non-shaded regions in panel **C** denote areas that are statistically significant.

~~The reconstructed AIS SMB averaged from 1801-2000 shows~~ Along with higher SMB values~~around the coastal areas, particularly in the Antarctic Peninsula and West Antarctic regions (Fig. 2A). The highest absolute SMB trends are around the~~, the coastal regions of East Antarctica and the Antarctic Peninsula also show the highest absolute SMB trends (Fig. 2**B**). This reconstruction also highlights large portions of East Antarctica as well as the Antarctic Peninsula as the regions with the most significant SMB trends from 1801-2000 (Fig. 2**C**). Taking the spatial average but keeping the temporal information yields the ~~AIS integrated~~AIS-integrated, reconstructed SMB time series shown in Fig. 3**C** (black).

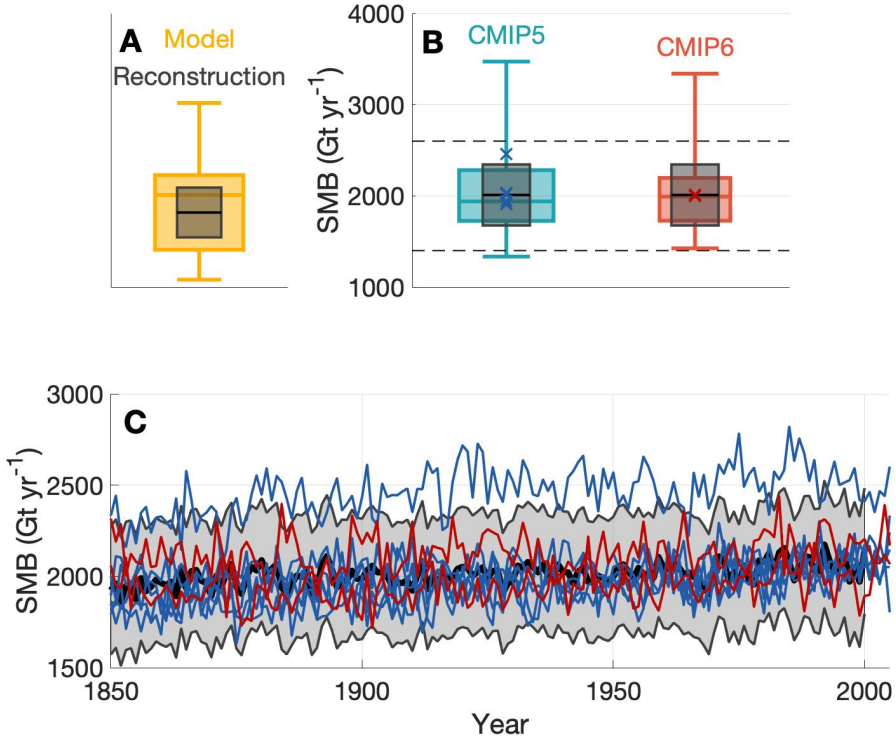

Figure 3. A An example of a box plot for model data (yellow) and reconstructed data (black and grey). The yellow shaded box shows the models' interquartile range while the whiskers extend to capture the entire distribution of modeled data. The line going through the box plot shows the median model value. The grey shaded box shows the reconstructed uncertainty around the reconstructed value shown as a black line. **B** A box plot of spatially integrated, temporally averaged (1850-2000) AIS SMB for CMIP5 (aqua) and CMIP6 (red). The dark blue ~~, green, coral,~~ x's associated with the CMIP5 box and ~~dark~~ the red ~~dots~~ x's associated with the CMIP6 box represent the ~~four~~ eight best scoring models: GISS E2 H CC, GISS E2 R CC, GISS E2 R, MPI ESM LR, ~~and~~ MPI ESM MR, ~~respectively~~and MPI ESM P from CMIP5 and CESM FV2 and MPI ESM2 LR from CMIP6. The black dashed lines indicate the lower and upper bounds of the time series plot in the bottom of Figure 3. **C** A time series of spatially integrated SMB for the reconstruction (black) and its uncertainty (shaded grey) with the best ~~four~~ eight scoring models: GISS E2 H ~~(dark blue)~~CC, GISS E2 R ~~(green)~~CC, GISS E2 R, MPI ESM LR~~(coral)~~, ~~and~~ MPI ESM MR, and MPI ESM P from CMIP5 (dark blue) and CESM FV2 and MPI ESM2 LR from CMIP6 (red)~~.~~

Panel (A) in Fig. 3 shows an example box plot for a suite of models in yellow and the reconstructed observations in black and grey. Panel (B) in Fig. 3 shows a box plot of the temporal average of the spatially integrated AIS SMB for CMIP5 and CMIP6.
255 The ~~average~~ interquartile range of AIS-integrated SMB in the CMIP5 models ~~range between 1335 and 3472~~ is between 1727 and 2282 Gt yr^{-1} compared to the CMIP6 models ~~which range between 1471 and 3339~~ whose interquartile range is between 1728 and 2196 Gt yr^{-1}. The ~~interquartile ranges for CMIP5 and CMIP6 are 1727 to 2282~~ best eight models range from 1909

to 2461 Gt yr^{-1} ~~and 1728 to 2229 Gt yr^{-1}, respectively, with means of 1940 Gt yr^{-1} and 2115 Gt yr^{-1}, respectively~~for the temporal average AIS-integrated SMB mean value.

The reconstructed AIS SMB ranges from 1800 $\pm$ 338 Gt yr^{-1} from 1850-1900 to 2039 $\pm$ 333 Gt yr^{-1} from 1950-2000. All ~~four of~~ but one of the eight of best scoring models are fully captured within the reconstructed uncertainty for the entire 150 year time series. The reconstruction and best scoring models all show generally increasing SMB from 1850-2000, albeit with large interannual variability. Both the trend and variability are analyzed in follow-up evaluations and scoring.

While the reconstructed SMB time series and ~~four~~ eight best scoring models show a generally increasing trend, the same is not true for all CMIP5 or CMIP6 models (Fig. 4). 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 overlapping time series from 1850-2000, the last century from 1900-2000, and the last 50 years from 1950-2000. The reconstructed linear SMB trends for the three time slices are 0.52 $\pm$ 0.27 Gt yr^{-2} (1850-2000), 0.56 $\pm$ 0.38 Gt yr^{-2} (1900-2000), and 1.0 $\pm$ 1.3 Gt yr^{-2} (1950-2000). ~~For~~ That implies that for all but the last time slice, 1950-2000, the reconstruction uncertainty trends are ~~also~~ exclusively positive.

Looking at all of the CMIP5 and CMIP6 models, the median linear trend is positive ~~and trends range in absolute minimum to absolute maximum from -3.8~~ for all three time slices and the trend interquartile ranges are from -0.8 to +~~6.7~~ 1.8 Gt yr^{-2} for 1850-2000, ~~-4.8~~ -0.6 to +~~3.4~~ 1.7 Gt yr^{-2} for 1900-2000, and ~~-1.4~~ 0.8 to +~~9.5~~ 2.7 Gt yr^{-2} for 1950-2000~~with median trends of~~. For CMIP5, median trends for these time slices are 0.88 Gt yr^{-2}, 0.66 Gt yr^{-2}, and 1.8 Gt yr^{-2} for 1850-2000, 1900-2000, and 1950-2000 respectively. For CMIP6, median trends for these time slices are 0.05 Gt yr^{-2}, ~~respectively. The four~~ 0.46 Gt yr^{-2}, and 1.8 Gt yr^{-2} for 1850-2000, 1900-2000, and 1950-2000 respectively. The eight best scoring models range from ~~-2.5~~ -1.4 to +~~0.81~~ 3.1 Gt yr^{-2}, ~~-0.92~~ -1.4 to +~~3.4~~ 1.7 Gt yr^{-2}, and ~~-0.33~~ -0.9 to +~~4.4~~ 2.4 Gt yr^{-2} for the same respective time spans. The spread in the ~~four~~ eight best scoring models reduces the total spread by ~~31%, 52~~57%, 62%, and ~~43~~70%, respectively. For the first two time slices, the reconstructed trend and uncertainty are captured within the interquartile range for all CMIP5 models. For 1950-2000, the models tend to overestimate the reconstructed trend. ~~The four best scoring models are at the lower end of the model estimates and the two MPI ESM models are captured within the reconstructed uncertainty.~~

~~Similar to CMIP5, the median linear trend is positive for the latter two time slices for CMIP6. The median linear trend in the first time slice, however, is negative in CMIP6, implying that more than half the CMIP6 models produce a negative SMB trend over the 151-year time series. Also similar to CMIP5, the first two time slices also capture the reconstructed trend and uncertainty in the interquartile range while the models tend to overestimate the trend for 1950-2000. The spread in trend in the CMIP6 models is significantly lower than for CMIP5 models which follows as there are fewer models. The trends for CMIP6 range from -2.9 to +4.1 Gt yr^{-2} with a median trend of -0.44 Gt yr^{-2} for 1850-2000, -1.8 to +2.9 Gt yr^{-2} with a median trend of 0.68 Gt yr^{-2} for 1900-2000, and 1.2 to 4.4 Gt yr^{-2} with a median trend of 1.9 Gt yr^{-2} for 1950-2000.~~

Apart from its trend magnitude and sign, SMB variability is also important for accurately representing SMB~~and determining the impact on sea level it may have in any given year. SMB variability also is~~, and can be indicative of the relevant ~~drivers behind SMB .~~

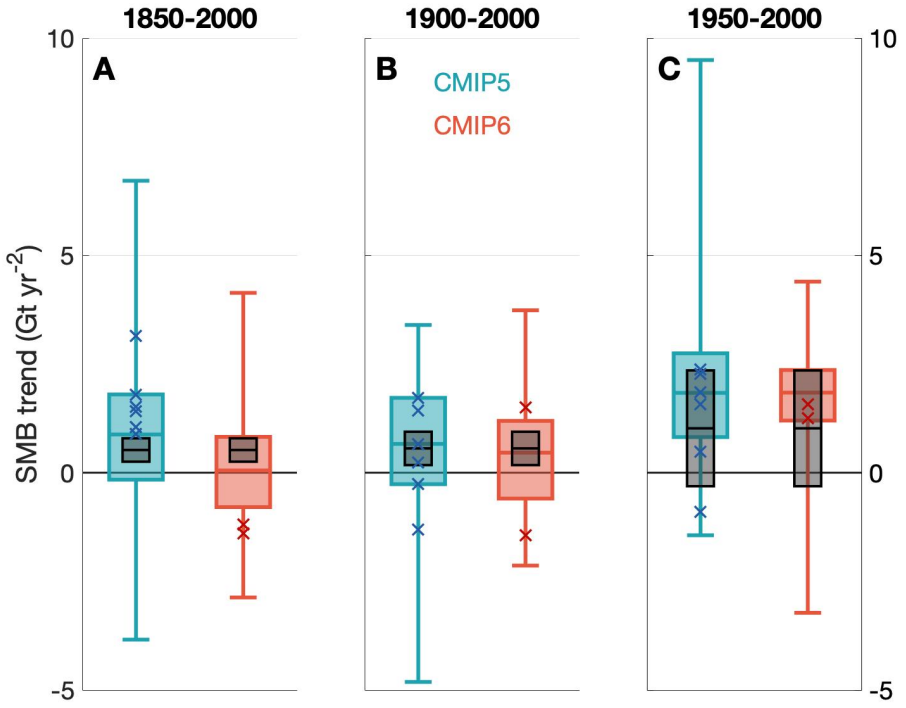

Figure 4. Box plots of the linear trends in spatially integrated AIS SMB in CMIP5 (blue) and CMIP6 (red) for the periods **A** from 1850 to 2000; **B** from 1900 to 2000 ; and **C** from 1950 to 2000. In all three panels, the grey boxes denote the reconstructed uncertainty around the reconstructed trend (black line). The ~~four~~ eight best scoring models are ~~shows in~~ represented by dark blue ~~, green, coral, and red, with~~ x's if they are among the ~~colors corresponding to the same~~ CMIP5 suite of models ~~as in Figure 3~~or red x's if they are among the CMIP6 suite.

~~Gaussian distributions of SMB where the standard deviation is that of the SMB time series for the reconstruction (black) and **A** GISS E2 H (dark blue) **B** GISS E2 R (green) **C** MPI ESM LR (coral) and **D** MPI ESM MR (dark red). **E** Box plots of the CMIP5 (blue) and CMIP6 (red) SMB time series standard deviations. The black dots show the standard deviation of the reconstruction.~~

SMB driving mechanisms. Figure 5**A**-**B** shows the average detrended and normalized variability for CMIP5 and CMIP6 models as well as the reconstruction plotted as a normal distribution. The detrended and normalized interannual variability in SMB in the reconstruction ranges between ~~-6.4-8.0%~~ ~-20% to 20%, while SMB in ~~the best four~~ all the models varies between ~~-10-10%(Fig. 5~~-15 to 15%. Figure 5~~A~~**C** ~~-D~~). All shows a box plot the standard deviations of the normalized and detrended time series. The normalization process made it such that the standard deviations are calculated in % of variability about the mean value of the time series. The standard deviation for the normalized and detrended SMB in the reanalysis is about 6.6% compared to the best eight models which range between 4.4% to 5.1%. (For comparison, the reconstructed normalized

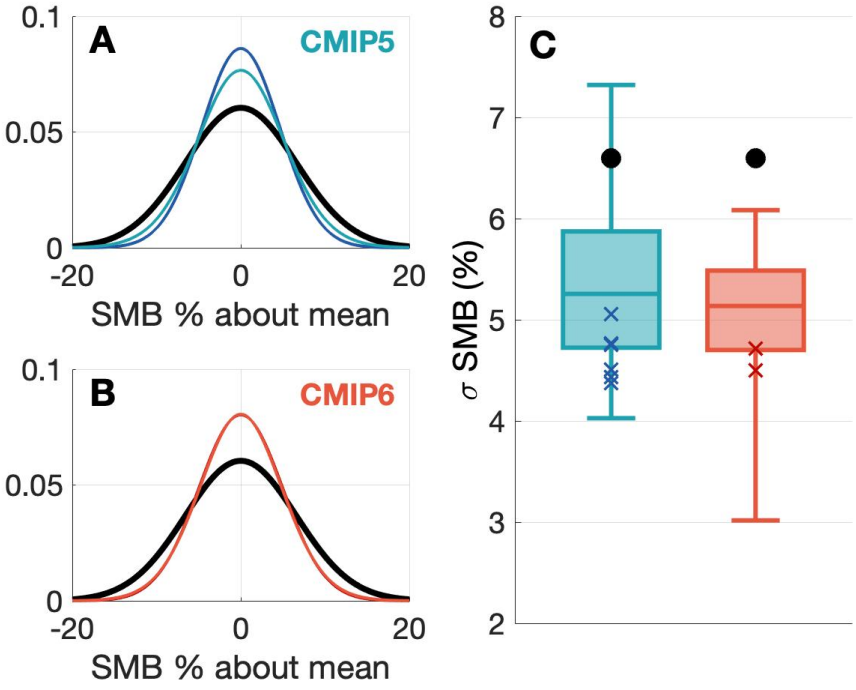

Figure 5. Gaussian distributions of SMB where the standard deviation is that of the SMB time series for the reconstruction (black) and **A** all CMIP5 models in light blue and the best scoring CMIP5 models in dark blue and **B** all CMIP6 models in light red and all CMIP6 models in red (the two Gaussians, here, are largely indistinguishable by eye as they overlap almost entirely). **C** Box plots of the CMIP5 (blue) and CMIP6 (red) SMB time series standard deviations. The black dots show the standard deviation of the reconstruction.

and detrended SMB standard deviation is about 2.9%.) Most CMIP5 and CMIP6 models ~~overestimate~~ underestimate SMB variability. The CMIP5 and CMIP6 models~~range~~' standard deviations range from 4.0% to 7.3% and from ~~overestimates of 144% to 261% and 151% to 217%of the reconstruction standard deviation~~3.0% to 6.1%, respectively (Fig. 5~~E~~C).

Just as temporal SMB variability is important for accurately capturing AIS SMB, spatial variations in SMB are also important in AIS SMB representation in models as ~~melt and discharge are not distributed equally~~precipitation is not distributed uniformly. To look at the spatial variability in SMB, we performed EOF analysis and plotted looked at the top three modes of variability which collectively account for 76.3% of the total spatial variability.

Separated out, the top three modes of variability in the reconstruction from EOF analysis explain 39%, 26%, and 12% of the total variability, respectively (Fig. 6). High values on the EOF map indicate regions that explain large amounts of the variability in AIS SMB. The top mode of variability in the reconstruction shows a dipole pattern from the Antarctic Peninsula to the Ross Sea region. ~~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~~

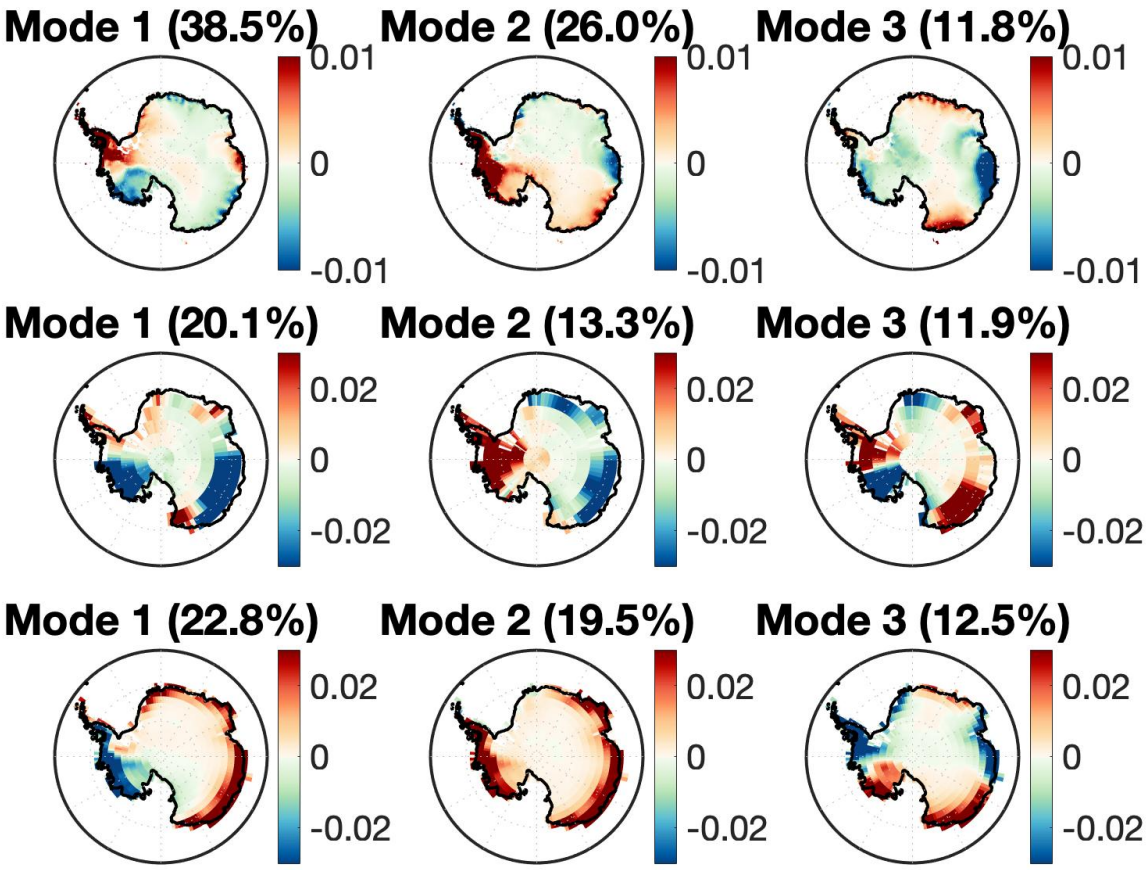

Figure 6. EOF analysis plots of the top 3 modes of variability for **A** the reconstruction, **B** a relatively high scoring model (CMCC CM), and **C** a low scoring model (CESM1 WACCM). Note that the scale for the model EOFs is 3× that of the reconstructed EOF.

~~(Turner et al., 2013), is marked by high precipitation around the coast of the Antarctic Peninsula (Grieger et al., 2016). Changes in the Amundsen Sea Low synoptic pattern, then, represent the dominant cause of variability in the reconstruction SMB. The depth of the ASL is strongly influenced by the phase of the Southern annular mode (SAM) with positive (negative) mean sea level pressure anomalies when the SAM is negative (positive) (Turner et al., 2013). The second mode of variability represents high variability in West Antarcticaand the Antarctic Peninsula. This could be caused by the topography in these regions which can induce large amounts of snowfall.~~ Mode 2 of the reconstruction EOF shows a strong signal over the entire Antarctic Peninsula and toward the Ross Ice Shelf region of West Antarctica. The third mode of variability shows a strong signal in Wilkes Land (East Antarctic region), near the Davis Sea, and two opposite, weaker signals in Dronning Maud Land (Atlantic

sector) and Adélie land (Pacific sector). This signal is reflective of the linear trend in SMB as seen in Fig. 2**B**. ~~See supplemental for further EOF analysis of sea level pressure variability.~~

~~By~~ As a example of the 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. However, even the better scoring models ~~tended~~ tend to overestimate the magnitude of the variance particularly around the coast even when they capture the general spatial patterns. CESM1 WACCM, one of the poorer performing models with regard to this metric, generally overestimates the variance everywhere in all three of the top modes. The top mode for this model reflects an East/West Antarctic SMB dipole and mode 2 shows a strong, unidirectional signal across the entire AIS, though mode 3 seems to reflect the same quadrupolar pattern as seen in the reconstruction~~albeit with to~~ , albeit with a much higher magnitude.

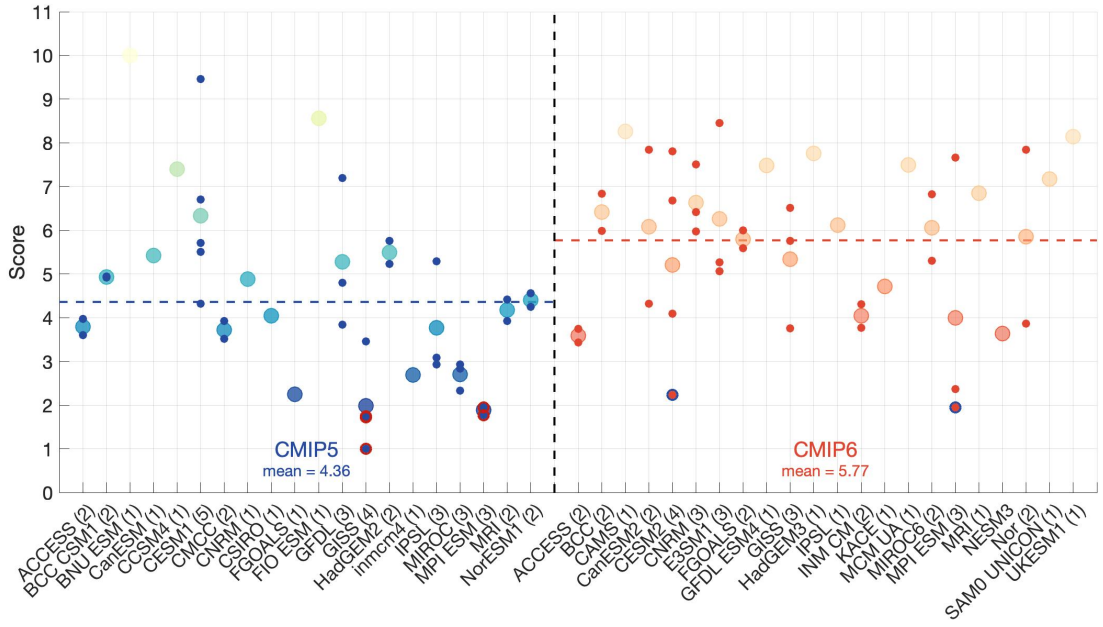

Figure 7. The scores for all CMIP5 and CMIP6 models. The large dots show the average score for all model groupings. Models are grouped by similar model physics and have in parenthesis the number of models in the grouping after the name. Each model grouping has all model scores plotted as small blue/red dots for CMIP5/6 with the model average plotted in the larger dots. Models that have no like models are followed by a one in parenthesis and only have a larger dot. The ~~four~~ eight best scoring models (above the 90th percentile) are denoted with ~~yellow x's instead~~ red outlines if they are among the CMIP5 suite of models – GISS E2 H CC, GISS E2 R CC, GISS E2 R, MPI ESM LR, MPI ESM MR, and MPI ESM P – or with blue ~~dots~~outlines if they are among the CMIP6 suite of models – CESM FV2 and MPI ESM2 LR. Note that the overall scores for two of the GISS models and three of the MPI models in CMIP5 are almost exactly equal so outlines overlap almost completely.

Models that score above the 90th percentile make up the subset of best scoring models. ~~Five~~ Eight models – GISS E2 H CC, GISS E2 RCC, GISS E2 R, MPI ESM LR, MPI ESM MR, and MPI ESM P from CMIP5 and CESM FV2 and MPI ESM LR from CMIP6 – ~~qualify for this status as there is a three-way tie for third, but~~ comprise this top 90th percentile. The two CMIP6 models as well as MPI ESM P ~~does not have the necessary information for future projections , it is neglected. Similar to~~ GISS E2 R from CMIP5 ~~, the GISS models in~~ do not appear in the future projections analysis as CMIP6 ~~are also among the best performing model in the small sample size~~does not follow the same RCP structure as CMIP5 and the MPI ESM P model does not contain the necessary information to perform the analysis. The poorest performing models include ~~CESM FASTCHEM,~~ BNU ESM, CESM FASTCHEM, and FIO ESM~~in CMIP5, and CanESM2 in CMIP6~~. The mean model score is ~~3.7~~ 4.36 for CMIP5 and ~~4.5~~ 5.77 for CMIP6. CMIP5 and CMIP6 scores were normalized together such that all scores are on the same scale and are directly comparable. With that, there is not much change from CMIP5 to CMIP6. ~~In fact, the scores increase from CMIP5 to CMIP6 albeit with a small sample size of models for CMIP6.~~

With this subset of the ~~four~~ eight best performing models, we then refined future projections of AIS SMB in terms of ~~magnitude~~mean value, trend, and variability. Because there are currently an insufficient number of future model runs available for CMIP6, our projection efforts were solely based on CMIP5. ~~Future CMIP5 projections are created in the context of warming scenarios called Representative Concentration Pathways (RCPs). The RCPs we used to investigate SMB projections are RCP2.6, RCP4.5, and RCP8.5 which have progressively higher CO_2 concentration projections and, thus, higher projected global warming.~~ Comparing the difference in SMB projections between ~~these~~ RCPs allows us a look into the different potential sea level changes caused by different amounts of warming. In CMIP5, there are 25 model outputs for RCP2.6 and 32 model outputs for RCPs 4.5 and 8.5.

As stated earlier, both ~~magnitude~~ mean value and trend of AIS SMB have significant implications for future projections of sea level change. The spatially integrated AIS SMB (i.e. SMB ~~magnitude~~mean value) has been increasing from 1850-2000 (Fig. 3) and is projected to continue to increase for the following hundred years to 2100 in all three warming scenarios (Fig. 8).

From 2070-2100, spatially integrated AIS SMB is projected to be ~~2295~~ 2751 $\pm$ ~~1222~~ 570 Gt yr^{-1} for RCP2.6, ~~2382~~ 2948 $\pm$ ~~1316~~ 581 Gt yr^{-1} for RCP4.5, and ~~2648~~ 3307 $\pm$ ~~1530~~ 663 Gt yr^{-1} for RCP8.5 for all CMIP5 models where the associated uncertainties are 1-σ of all models between 2070-2100 (for a list of projected SMB and related variable values for all models and the best scoring models across the RCPs, see supplementary). The subset of ~~four~~ eight best scoring models have lower projections and smaller spread at ~~2246~~ 2372 $\pm$ ~~268~~ 282 Gt yr^{-1} for RCP2.6, ~~2358~~ 2452 $\pm$ ~~331~~ 286 Gt yr^{-1} for RCP4.5, and ~~2495~~ 2588 $\pm$ ~~335~~ 291 Gt yr^{-1} for RCP8.5 on average between 2070-2100. The ~~magnitude of~~ ranges of the best eight scoring models reduced the spread by 79%, 79%, and 74% for RCPs 2.6, 4.5, and 8.5, respectively. The mean value of modeled SMB increases with increasing warming scenarios for all CMIP5 models and the subset of the ~~four~~ eight best scoring models. Similarly to the ~~magnitude~~ mean value increasing with increasing warming, the projected SMB trend also increases with increased warming (Fig. 9). As such, the stronger the emission scenario, the larger the projected response in AIS SMB with regard to both ~~magnitude~~ mean value and trend.

For the entirety of the 21st century, 2000-2100, most CMIP5 climate models project positive SMB trends in all forcing scenarios (Fig. 9). For RCP2.6, all CMIP5 models project a median trend of 0.53 Gt yr^{-2} and a range of -2.15 to +2.63 Gt

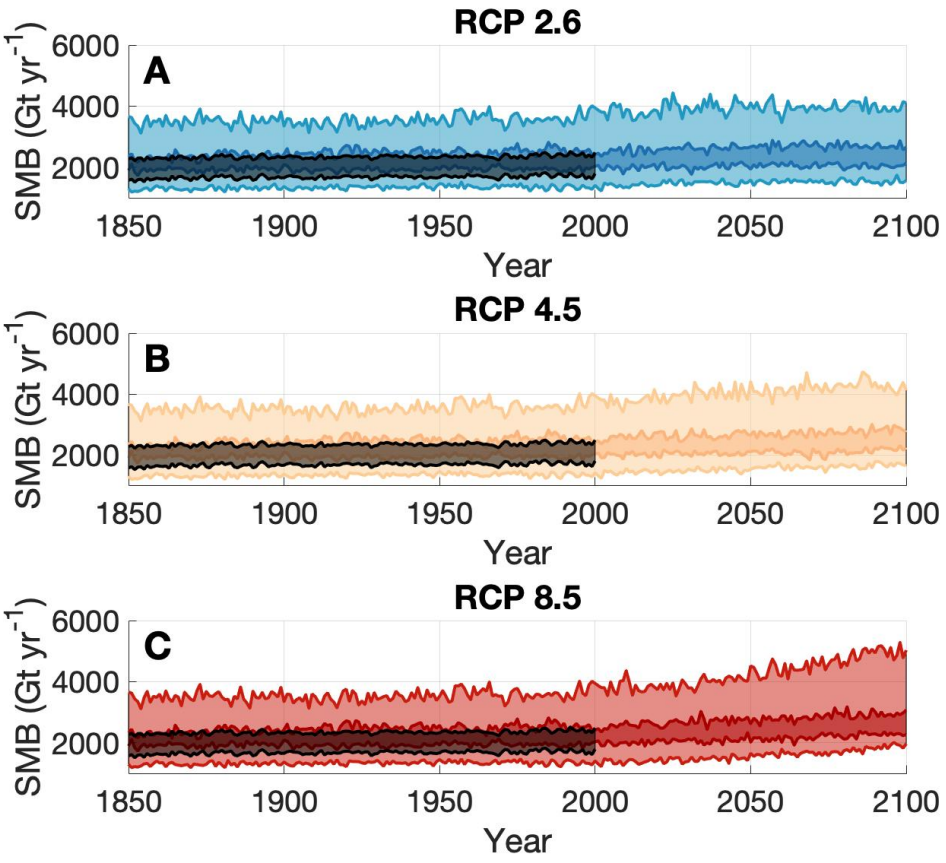

Figure 8. Time series for the reconstruction with uncertainty bounds (grey), all CMIP5 models (light) and best scoring CMIP5 models (dark) for **A** RCP2.6 (blue), **B** RCP4.5 (yellow), and **C** RCP8.5 (red).

yr^{-2}. For RCPs 4.5 and 8.5, the median trends are 2.28 Gt yr^{-2} and 5.64 Gt yr^{-2} with ranges of -0.81 to +6.11 Gt yr^{-2} and 0.47 to 14.9 Gt yr^{-2}, respectively.

370 The best scoring models range from ~~0.34 to 2.09~~ 0.47 to 2.45 Gt yr^{-2}, 1.44 to 2.88 Gt yr^{-2}, and 3.06 to 4.63 Gt yr^{-2} for RCPs 2.6, 4.5, and 8.5, respectively. For RCPs 2.6 and 4.5, the best scoring model trend projections lie close to or within the interquartile range for all CMIP5 models. ~~The best four model projections are near or below the lower bound of the interquartile range for RCP8.5~~As the warming scenarios strengthen, the four of the eight best scoring models projected into the future move closer to the lower end of the overall CMIP5 interquartile range in trend. Some of the differences in these concentration
375 pathways can be described by the modeled SMB sensitivity to different atmospheric CO_2 emission scenarios.

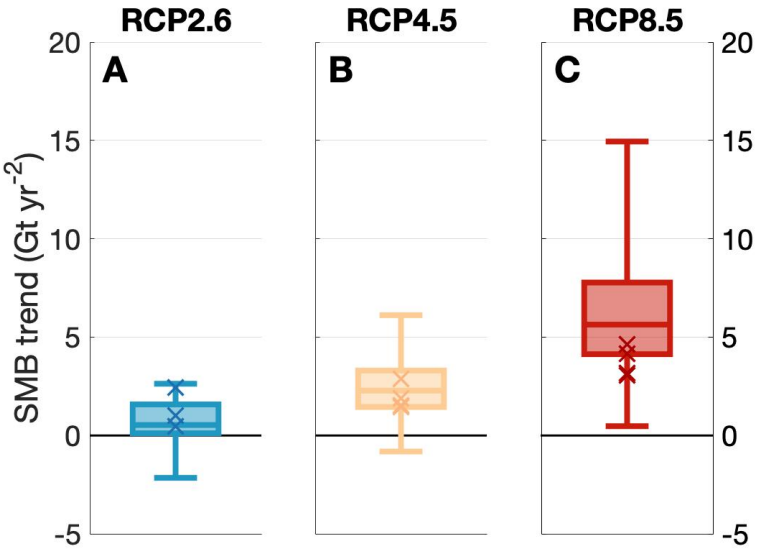

Figure 9. Box plots of the linear trend in spatially integrated AIS SMB from 2050-2100 for **A** RCP2.6 (blue), **B** RCP4.5 (yellow), and **C** RCP8.5 (red). The four ~~larger, colored dots represent~~ darker x's denote the four ~~best scoring~~ models ~~:~~ – GISS E2 H ~~(dark blue)~~CC, GISS E2 R ~~(green)~~CC, MPI ESM LR~~(coral)~~, and MPI ESM MR ~~(dark red)~~ – among the eight best scoring models with the appropriate and necessary information for direct comparison of future projections.

Box plots of modeled SMB sensitivity to changes in temperature (i.e. how much SMB will change per degree warming) ~~show that SMB responds differently in different warming scenarios (~~are shown in Fig. 10~~). The CMIP5 models project that each warming scenario with higher CO_2 concentrations will see greater SMB sensitivity to increases in temperature than those with lower CO_2 concentrations. While the ranges differ from scenario to scenario, the~~ . The projected sensitivity medians for RCPs 2.6, 4.5, and 8.5 are 101.7 Gt °~~C~~K^{-1}, 111.2 Gt °~~C~~K^{-1}, and 128.2 Gt °~~C~~K^{-1}, respectively. These results are not statistically significantly different from one another, indicating no significant more-than-linear SMB increase in enhanced warming scenarios.

~~The different responses to the warming scenariosindicates that the concentration of carbon dioxide in the atmosphere has a coupled role in AIS SMB~~

5 Discussion

5.1 EOF Analysis

Mode 1 of the reconstruction EOF shows a dipolar pattern across the Antarctic Peninsula and Ross Ice Shelf region of West Antarctica. This dipole corresponds to variability in precipitation generated by variations in the track and strength of

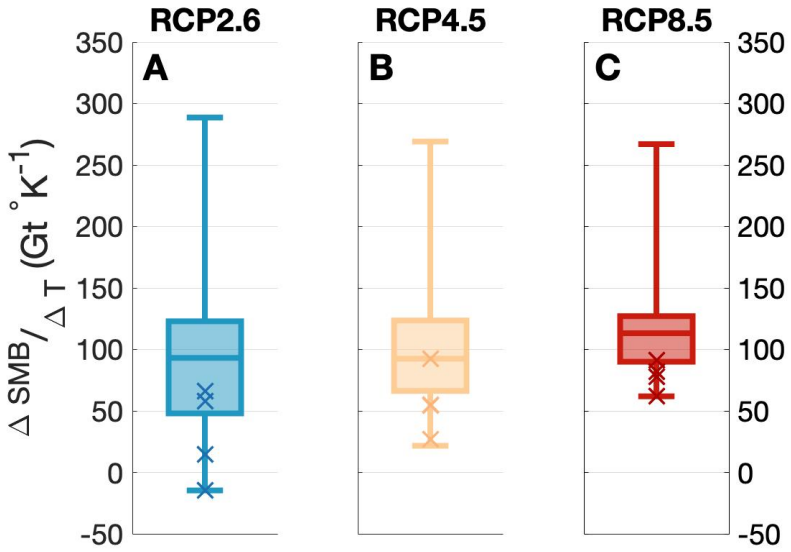

Figure 10. Box plots of all CMIP5 models' projected SMB sensitivity to temperature changes (ΔSMB/ΔT) for **A** RCP2.6, **B** RCP4.5, and **C** RCP8.5. The ~~four larger, colored dots represent~~ five darker x's denote the four ~~best scoring~~ models ~~:~~ – GISS E2 H ~~(dark blue)~~CC, GISS E2 R ~~(green)~~CC, MPI ESM LR~~(coral)~~, and MPI ESM MR ~~(dark red)~~ – among the eight best scoring models with the appropriate and necessary information for direct comparison of future projections.

the Amundsen Sea Low. 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 (Turner et al., 2013), is marked by high precipitation around the coast of the Antarctic Peninsula (Grieger et al., 2016). Changes in the Amundsen Sea Low synoptic pattern, then, represent the dominant cause of variability in the reconstruction SMB. The depth of the ASL is strongly influenced by the phase of the Southern annular mode (SAM) with positive (negative) mean sea level pressure anomalies when the SAM is negative (positive) (Turner et al., 2013).

Looking at mode 2, previous work by Hosking et al. (2013) and Turner et al. (2013) (among others) have shown that variability in the Amundsen Sea Low is responsible for high 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. Additional work highlighted in the supplementary material indicates that variability in sea level pressure in the Amundsen Sea region may be playing a large role in the AIS SMB spatial variability patterns.

5.2 Impact of Internal Variability in Model Scoring: CESM Large Ensemble

The CESM Large Ensemble (CESM-LENS) is an experiment wherein the Community Earth System Model Version 1 (CESM) is run 40 times with random temperature perturbations at the level of round-off error applied in 1920 (Kay et al., 2015). Because of its large number of ensemble members, the CESM-LENS experiment is useful for quantifying the role of internal variability. Only 35 of the original 40 ensemble members contain the necessary information for assessing AIS SMB. ~~As seen in figures 8 and 9, the greater the CO_2 concentration, the larger the AIS SMB response. From Fig. 10, we can also say that the greater the CO_2 concentration, the more sensitive to warming AIS SMB is meaning that continued warming past the end of the 21st century will have increasingly greater effects on AIS SMB .~~ Figure 4 in Supplementary shows the final scores of the five CESM simulations that are included in the CMIP5 suite of models as well as the final scores of the CESM-LENS experiment. The final scores for the CESM-LENS model runs are calculated the same way for all model criteria except for AIS-integrated trend. Because these runs only differ after 1920, we only use the third time slice (1950-2000) to assess the quality of trend reproduction.

The final scores of the five CMIP5 CESM model runs range from 3.99 to 9.74 while the final scores of the 35 CESM-LENS runs range from 1.32 to 5.96. Given that the scores range by 5.74 and 4.65 for the CMIP5 CESM runs and the CESM-LENS runs, respectively, it is reasonable to conclude that internal variability plays as significant a role in determining final score as do model parameterizations.

A major caveat of this finding, though, is that the CESM-LENS runs and the reconstruction only overlap from 1920-2000. This will likely most significantly impact the assessment of the trend and EOF analyses.

With that, internal variability plays a significant role in our AIS SMB assessment. Some models within the CMIP5 and CMIP6 frameworks, such as CESM1-CAM5, have many ensemble members. However, not all models – and even not all model versions – have multiple ensemble members. As such, performing a direct comparison of the models using the ensemble mean would not necessarily yield an accurate result as models with more ensemble members would have their final score shifted significantly while the same is not true for models with a single ensemble member. For considering using GCMs for AIS SMB analysis, then, we strongly suggest taking into account the fact that internal variability could be playing a strong role in some models final score, and that the number of ensemble members available should be considered along with the final score.

5.3 Impact of Model Resolution in Model Scoring

The CMIP5 and CMIP6 models vary in resolution from about $0.75° \times 0.75°$ to $3° \times 3°$ (Tables 1-3 in Supplementary). Figure 5 in Supplementary shows a scatter plot of resolution versus total score. Resolution, here, is the latitudinal resolution multiplied by the longitudinal resolution such that a model with latitude/longitude resolutions 0.9375°/1.25° would have a resolution of 1.1719°. A linear regression yields a correlation of R = -0.40 with 95% confidence intervals of -0.62 and -0.17. From this, there is a statistically significant negative correlation between resolution and total model score, signaling that, perhaps contrary to intuition, lower-resolution models score equally well, if not better, than higher resolution models. This result might be skewed by the fact that lower-resolution models include better physics to represent AIS SMB than higher-resolution models. However, when comparing total scores from the same model run at different resolutions, we find a consistent result: the relative

high-resolution CESM CAM5, IPSL CM5A MR, MPI ESM MR, CESM2, CESM2 WACCM, and MPI ESM2 HR all perform worse than their coarser resolution counterparts – CESM CAM5 FV2, IPSL CM5A LR, MPI ESM LR, CESM2 FV2, CESM2 WACCM FV2, and MPI ESM2 LR. Because so many models close to 1°/1° resolution and there is large spread in these models' final scores, we also divided the models into two groups, finer and coarser than 1.25°/1.25°, and performed the same regression analysis. Figure 6 in Supplementary shows the coarser resolution models have a correlation of R = -0.14 with 95% confidence intervals of -0.51 and 0.24 while finer resolution models have a correlation of R = -0.06 with 95% confidence intervals of -0.38 and 0.26. From this, we conclude that there is no significant correlation between model resolution and total score.

6 Conclusions

In this paper, we tested the ability of the suite of models in CMIP5 to capture SMB reconstructed from ice cores and reanalysis products by scoring them using a series of criteria: AIS-integrated mean value, trend, and variability, as well as the spatial variability patterns. This scoring system is designed as a guide for choosing what GCMs to focus on studying for ~~SMB prediction.~~ future SMB projections. Using this scoring system, we found that the top 90th percentile models were GISS E2 H CC, GISS E2 R CC, GISS E2 R, MPI ESM LR, MPI ESM MR, and MPI ESM P of CMIP5 and CESM FV2 and MPI ESM2 LR of CMIP6. A similar study in Agosta et al. (2015) found ACCESS1-3, ACCESS1-0, CESM BGC, CESM CAM5, NorESM1-M, and EC-Earth to most accurately capture AIS sea level pressure, 850 hPa air temperature, precipitable water, and ocean conditions – all of which impact AIS SMB to varying degrees. They focused their investigation into more atmospheric and oceanic dynamics (sea ice extent, sea surface temperature, sea surface pressure, precipitable water, 850 hPa temperature) and were comparing models directly to a reanalysis product. Barthel et al. (2019), another study with a similar goal of analyzing SMB performance among GCMs selected CCSM4, MIROC ESM CHEM, and NorESM1-M as their top three performing models for Antarctica. They ruled out both the GISS and MPI modeling groups due to their initial selection criteria and were also looking more at the impacts thermodynamical processes on SMB.

Our SMB mean value estimates are comparable to Agosta et al. (2019), who found a mean SMB value of roughly 2100 $\pm$ 100 Gt yr^{-1} for the grounded AIS using ERA-Interim products. The SMB trends are also in line with Medley and Thomas (2019) over the 20th century. Unlike previous studies, we use a reconstructed data set based on ice core reanalysis, not RCMs. Also of note is the fact that this data set and the GCMs we use for comparison allow us to investigate much longer time periods (150 years), enhancing the robustness of long-term AIS SMB trends. Using this reconstruction, we are able to refine estimates of SMB mean value and SMB trend by the end of the 21st century using CMIP5 by assigning scores to the models and creating a subset of the most accurate models historically. Also unlike previous studies, we analyze both CMIP5 and the early models of CMIP6 together allowing for direct comparison between the two suites of models. The scores for all CMIP5 models are, on average, ~~lower~~ better than the average score of the currently released CMIP6 models.

All scores are equally weighted to avoid issues with coincidental good or bad performance. Having a spread of criteria against which we score the models limits the possibility that models are recreating one aspect well for the wrong reasons. This scoring method does well in determining simple and consistent criteria to score the accuracy of modeled SMB. In contrast,

it struggles to recognize any difference in the importance of individual criteria as they are all weighted equally and also only reflects a few, simple scoring metrics. The criteria were chosen such that they all carry equal weight which we justify by arguing that not meeting any one of the criteria to within a reasonable degree would significantly impact future SMB estimates.

Using the top ~~four~~ eight best scoring models, ~~we~~ four of which we were able to project out to 2100 under three different RCPs, we refined future SMB predictions to ~~2246~~ 2372 $\pm$ ~~268~~ 282 Gt yr^{-1} for RCP2.6, ~~2358~~ 2452 $\pm$ ~~331~~ 286 Gt yr^{-1} for RCP4.5, and ~~2495~~ 2588 $\pm$ ~~335~~ 291 Gt yr^{-1} for RCP8.5. Over the 21st century this translates to ~~1.82 cm, 3.49~~ 8.6 cm, 9.6 cm, and ~~6.57~~ 11 cm of GMSL rise buffering in RCPs 2.6, 4.5, and 8.5, respectively, for all of CMIP5. ~~The best models, which show lower AIS-integrated SMB over the 21st century compared to all of~~ Our result of these best scoring models projecting AIS SMB at the lower end of the overall CMIP5 ~~, project 1.30 cm, 3.22 cm, and 5.05 cm of GMSL rise buffering in~~ interquartile range in trend is in contrast to Palerme et al. (2017) who found that, especially considering RCPs 2.6 ~~,~~ and 4.5, ~~and 8.5, respectively~~the CMIP5 models that best captured snowfall change rates tended to predict higher snowfall rates into the 21st century. Additionally, model trends were refined to ~~0.9 $\pm$ 0.1~~ 0.47 to 2.45 Gt yr^{-2} for RCP2.6, ~~1.9 $\pm$ 1.0~~ 1.44 to 2.88 Gt yr^{-2} for RCP4.5, and ~~3.8 $\pm$ 1.8~~ 3.06 to 4.63 Gt yr^{-2} for RCP8.5. Comparing the projected change in SMB per degree warming between the emission scenarios gives median sensitivities of 64 $\pm$ 80 Gt °~~C~~K^{-1}, 57 $\pm$ 33 Gt °~~C~~K^{-1}, and 78 $\pm$ 15 Gt °~~C~~K^{-1} for RCPs 2.6, 4.5, and 8.5, respectively, for the best scoring models. ~~Combined, these data tell us that for stronger emission scenarios , the AIS SMB response will be stronger in both magnitude and trend.~~

However, these results are not statistically significantly different from one another across forcing scenarios and indicate that there is no difference in the sensitivity response to changes in temperature between the three forcing scenarios. Given that the best performing models show lower AIS-integrated SMB values and trends compared to the entire CMIP5 spread indicates less sea level rise mitigation from increasing SMB than is implied by looking at all CMIP5 models.

The major limitations of this work stem from the subjective selection of scoring criteria. While each model is scored based on the same criteria, each criterion is chosen specifically to gauge model performance for capturing AIS SMB. As such, these criteria may be ill suited for looking at other variables and, thus, other metrics could yield very different results. Another caveat of this work is that we are only capable of analyzing the CMIP6 models that have been released. As this analysis and the release of CMIP6 are concurrent, this limits the number of models we can reasonably analyze due to time constraints. Additional CMIP6 models may have different results and may skew the comparison between CMIP5 and CMIP6 significantly. Similarly, due to the small number of CMIP6 models released at this point, using statistical analyses becomes moot as the top 90% of models constitutes the single, best scoring model. One final major caveat 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. (2014); 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.

Appendix A

A1

Author contributions. T. G. and J. T. M. L. conceptualized and initiated this work. T. G. performed the analysis, discussed the results with J. T. M. L., and wrote the paper. B. M. provided the reconstructions and guidance on using and interpreting them. All authors reviewed the paper before submission.

Competing interests. The authors declare no competing interests.

Acknowledgements. T. G. and J. T. M. L. acknowledge support from the National Aeronatics and Space Administration (NASA), Grant 80NSSC17K0565 (NASA Sea Level Team 2017–2020).

References

Agosta, C., Fettweis, X., and Datta, R.: Evaluation of the CMIP5 models in the aim of regional modelling of the Antarctic surface mass balance, The Cryosphere, 9, 2311–2321, https://doi.org/10.5194/tc-9-2311-2015, www.the-cryosphere.net/9/2311/2015/, 2015.

Agosta, C., Amory, C., Kittel, C., Orsi, A., Favier, V., Gallée, H., van den Broeke, M. R., Lenaerts, J. T. M., van Wessem, J. M., and Fettweis, X.: Estimation of the Antarctic surface mass balance using MAR (1979–2015) and identification of dominant processes, The Cryosphere Discussions, pp. 1–22, https://doi.org/10.5194/tc-2018-76, 2019.

Barthel, A., Agosta, C., Little, C. M., Hatterman, T., Jourdain, N. C., Goelzer, H., Nowicki, S., Seroussi, H., Straneo, F., and Bracegirdle, T. J.: CMIP5 model selection for ISMIP6 ice sheet model forcing: Greenland and Antarctica, The Cryosphere Discussions, pp. 1–34, https://doi.org/10.5194/tc-2019-191, 2019.

Beaumet, J., Déqué, M., Krinner, G., Agosta, C., and Alias, A.: Effect of prescribed sea surface conditions on the modern and future Antarctic surface climate simulated by the ARPEGE atmosphere general circulation model, The Cryosphere, 13, 3023–3043, https://doi.org/10.5194/tc-13-3023-2019, https://www.the-cryosphere.net/13/3023/2019/, 2019.

Bromwich, D. H., Nicolas, J. P., and Monaghan, A. J.: An Assessment of precipitation changes over antarctica and the southern ocean since 1989 in contemporary global reanalyses, Journal of Climate, 24, 4189–4209, https://doi.org/10.1175/2011JCLI4074.1, 2011.

Burgener, L., Rupper, S., Koenig, L., Forster, R., Christensen, W. F., Williams, J., Koutnik, M., Miège, C., Steig, E. J., Tingey, D., Keeler, D., and Riley, L.: An observed negative trend in West Antarctic accumulation rates from 1975 to 2010: Evidence from new observed and simulated records, Journal of Geophysical Research Atmospheres, 118, 4205–4216, https://doi.org/10.1002/jgrd.50362, 2013.

Frezzotti, M., Scarchilli, C., Becagli, S., Proposito, M., and Urbini, S.: A synthesis of the Antarctic surface mass balance during the last 800 yr, Cryosphere, 7, 303–319, https://doi.org/10.5194/tc-7-303-2013, 2013.

Fyke, J., Lenaerts, J. T., and Wang, H.: Basin-scale heterogeneity in Antarctic precipitation and its impact on surface mass variability, National Snow and Ice Data Center, 558, 2595–2609, https://doi.org/10.1175/2011JCLI4074.1, http://nsidc.org/cryosphere/quickfacts/icesheets.html, 2017.

Gallée, H., Trouvilliez, A., Agosta, C., Genthon, C., Favier, V., and Naaim-Bouvet, F.: Transport of Snow by the Wind: A Comparison Between Observations in Adélie Land, Antarctica, and Simulations Made with the Regional Climate Model MAR, Boundary-Layer Meteorology, 146, 133–147, https://doi.org/10.1007/s10546-012-9764-z, 2013.

Genthon, C., Krinner, C., and Castebrunet, H.: Antarctic precipitation and climate-change predictions: Horizontal resolution and margin vs plateau issues, Annals of Glaciology, 50, 55–60, https://doi.org/10.3189/172756409787769681, 2009.

Grieger, J., Leckebusch, G. C., and Ulbrich, U.: Net precipitation of Antarctica: Thermodynamical and dynamical parts of the climate change signal, Journal of Climate, 29, 907–924, https://doi.org/10.1175/JCLI-D-14-00787.1, 2016.

Hosking, J. S., Orr, A., Marshall, G. J., Turner, J., and Phillips, T.: The Influence of the Amundsen–Bellingshausen Seas Low on the Climate of West Antarctica and Its Representation in Coupled Climate Model Simulations, Journal of Climate, 26, 6633–6648, https://doi.org/10.1175/JCLI-D-12-00813.1, http://journals.ametsoc.org/doi/10.1175/JCLI-D-12-00813.1, 2013.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and Vertenstein, M.: The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability, Bulletin of the American Meteorological Society, 96, 1333–1349, https://doi.org/10.1175/BAMS-D-13-00255.1, http://journals.ametsoc.org/doi/10.1175/BAMS-D-13-00255.1, 2015.

Kittel, C., Amory, C., Agosta, C., Delhasse, A., Doutreloup, S., Huot, P.-V., Wyard, C., Fichefet, T., and Fettweis, X.: Sensitivity of the current Antarctic surface mass balance to sea surface conditions using MAR, The Cryosphere, 12, 3827–3839, https://doi.org/10.5194/tc-12-3827-2018, https://doi.org/10.5194/tc-12-3827-2018, 2018.

Krinner, G., Largeron, C., Ménégoz, M., Agosta, C., and Brutel-Vuilmet, C.: Oceanic Forcing of Antarctic Climate Change: A Study Using a Stretched-Grid Atmospheric General Circulation Model, Journal of Climate, 27, 5786–5800, https://doi.org/10.1175/JCLI-D-13-00367.1, http://journals.ametsoc.org/doi/10.1175/JCLI-D-13-00367.1, 2014.

Lenaerts, J. T., Van Den Broeke, M. R., Van De Berg, W. J., Van Meijgaard, E., and Kuipers Munneke, P.: A new, high-resolution surface mass balance map of Antarctica (1979-2010) based on regional atmospheric climate modeling, Geophysical Research Letters, 39, 1–5, https://doi.org/10.1029/2011GL050713, 2012a.

Lenaerts, J. T., Vizcaino, M., Fyke, J., van Kampenhout, L., and van den Broeke, M. R.: Present-day and future Antarctic ice sheet climate and surface mass balance in the Community Earth System Model, Climate Dynamics, 47, 1367–1381, https://doi.org/10.1007/s00382-015-2907-4, 2016.

Lenaerts, J. T., Medley, B., van den Broeke, M. R., and Wouters, B.: Observing and Modeling Ice Sheet Surface Mass Balance, Reviews of Geophysics, https://doi.org/10.1029/2018RG000622, 2019.

Lenaerts, J. T. M., van den Broeke, M. R., Déry, S. J., van Meijgaard, E., van de Berg, W. J., Palm, S. P., and Sanz Rodrigo, J.: Modeling drifting snow in Antarctica with a regional climate model: 1. Methods and model evaluation, Journal of Geophysical Research: Atmospheres, 117, n/a–n/a, https://doi.org/10.1029/2011jd016145, 2012b.

Marshall, G. J., Thompson, D. W., and van den Broeke, M. R.: The Signature of Southern Hemisphere Atmospheric Circulation Patterns in Antarctic Precipitation, Geophysical Research Letters, 44, 580–11, https://doi.org/10.1002/2017GL075998, 2017.

Medley, B. and Thomas, E. R.: Increased snowfall over the Antarctic Ice Sheet mitigated twentieth-century sea-level rise, Nature Climate Change, 9, 34–39, https://doi.org/10.1038/s41558-018-0356-x, http://dx.doi.org/10.1038/s41558-018-0356-x, 2019.

Monaghan, A. and Bromwich, D.: Global warming at the poles, Nature Geoscience, 1, 728, https://doi.org/10.1038/ngeo346http://10.0.4.14/ngeo346, 2008.

Monaghan, A. J., Bromwich, D., Fogt, R., Wang, S.-H., Mayewski, P., Dixon, D., Ekaykin, A., Frezzotti, M., Goodwin, I., Isaksson, E., Kaspari, S., Morgan, V., Oerter, H., Van Ommen, T., Van der Veen, C., and Wen, J.: Insignificant Change in Antarctic Snowfall Since the International Geophysical Year Andrew, Science, 313, 827–831, https://doi.org/10.5061/dryad.5t110.Supplementary, 2006.

Palerme, C., Kay, J. E., Genthon, C., L'Ecuyer, T., Wood, N. B., and Claud, C.: How much snow falls on the Antarctic ice sheet?, Cryosphere, 8, 1577–1587, https://doi.org/10.5194/tc-8-1577-2014, 2014.

Palerme, C., Genthon, C., Claud, C., Kay, J. E., Wood, N. B., and L'Ecuyer, T.: Evaluation of current and projected Antarctic precipitation in CMIP5 models, Climate Dynamics, 48, 225–239, https://doi.org/10.1007/s00382-016-3071-1, 2017.

Philippe, M., Tison, J. L., Fjøsne, K., Hubbard, B., Kjær, H. A., Lenaerts, J. T., Drews, R., Sheldon, S. G., De Bondt, K., Claeys, P., and Pattyn, F.: Ice core evidence for a 20th century increase in surface mass balance in coastal Dronning Maud Land, East Antarctica, Cryosphere, 10, 2501–2516, https://doi.org/10.5194/tc-10-2501-2016, 2016.

Previdi, M. and Polvani, L. M.: Anthropogenic impact on Antarctic surface mass balance, currently masked by natural variability, to emerge by mid-century, Environ. Res. Lett, 11, 94 001, https://doi.org/10.1088/1748-9326/11/9/094001, 2016.

Shepherd, A., Ivins, E. R., A, G., Barletta, V. R., Bentley, M. J., Bettadpur, S., Briggs, K. H., Bromwich, D. H., Forsberg, R., Galin, N., Horwath, M., Jacobs, S., Joughin, I., King, M. A., Lenaerts, J. T. M., Li, J., Ligtenberg, S. R. M., Luckman, A., Luthcke, S. B., McMillan, M., Meister, R., Milne, G., Mouginot, J., Muir, A., Nicolas, J. P., Paden, J., Payne, A. J., Pritchard, H., Rignot, E., Rott, H.,

Sørensen, L. S., Scambos, T. A., Scheuchl, B., Schrama, E. J. O., Smith, B., Sundal, A. V., van Angelen, J. H., van de Berg, W. J., van den Broeke, M. R., Vaughan, D. G., Velicogna, I., Wahr, J., Whitehouse, P. L., Wingham, D. J., Yi, D., Young, D., and Zwally, H. J.: A Reconciled Estimate of Ice-Sheet Mass Balance, Science, 338, 1183 LP – 1189, https://doi.org/10.1126/science.1228102, http://science.sciencemag.org/content/338/6111/1183.abstract, 2012.

Thomas, E. R., Hosking, J. S., Tuckwell, R. R., Warren, R. A., and Ludlow, E. C.: Twentieth century increase in snowfall in coastal West Antarctica, Geophysical Research Letters, 42, 9387–9393, https://doi.org/10.1002/2015GL065750, 2015.

Thomas, E. R., Melchior Van Wessem, J., Roberts, J., Isaksson, E., Schlosser, E., Fudge, T. J., Vallelonga, P., Medley, B., Lenaerts, J., Bertler, N., Van Den Broeke, M. R., Dixon, D. A., Frezzotti, M., Stenni, B., Curran, M., and Ekaykin, A. A.: Regional Antarctic snow accumulation over the past 1000 years, Climate of the Past, 13, 1491–1513, https://doi.org/10.5194/cp-13-1491-2017, 2017.

Turner, J., Phillips, T., Hosking, J. S., Marshall, G. J., and Orr, A.: The amundsen sea low, International Journal of Climatology, 33, 1818–1829, https://doi.org/10.1002/joc.3558, 2013.

van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G. C., Kram, T., Krey, V., Lamarque, J. F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S. J., and Rose, S. K.: The representative concentration pathways: An overview, Climatic Change, 109, 5–31, https://doi.org/10.1007/s10584-011-0148-z, 2011.

van Wessem, J. M., van de Berg, W. J., Noël, B. P. Y., van Meijgaard, E., Birnbaum, G., Jakobs, C. L., Krüger, K., Lenaerts, J. T. M., Lhermitte, S., Ligtenberg, S. R. M., Medley, B., Reijmer, C. H., van Tricht, K., Trusel, L. D., van Ulft, L. H., Wouters, B., Wuite, J., and van den Broeke, M. R.: Modelling the climate and surface mass balance of polar ice sheets using RACMO2, part 2: Antarctica (1979–2016), The Cryosphere Discussions, pp. 1–35, https://doi.org/10.5194/tc-2017-202, 2017.

Dear editor,

Thank you for reading over our paper and providing your initial feedback before sending this manuscript out to the editors. We appreciate the work required to find reviewers and manage the review process. We have worked on the revisions and to address what we feel were the most significant suggestions, we plan to do the following:

1. include the CESM-LENS project into our analysis to investigate the spread among ensemble members of a single model
2. include the HighResMIP experiment from CMIP to investigate the impact that varying resolutions have on the final scoring result
3. perform the suggested leave-one-out cross validation process to help identify which criterion/criteria is/are most important for future projections
4. analyze the impact of using different masking regimes on, particularly, the AIS-integrated SMB mean value.

Item 4. here is in direct response to the editor's initial comments wherein they inquired about the effect of using model-specific masks as opposed to a common mask interpolated to the various grids. We feel that this is a very good question and we want to make sure that it is addressed in our future revisions. All other revisions are explained in detail below, with the original text in bold, the reviewer/editor comment in italics, and our response in green. We hope that you approve of our plan and that we can move forward with applying these revisions.

Thank you for your consideration,

Tessa Gorte and co-authors.

Editor

We interpolated an AIS mask (Shepherd et al., 2012) using the nearest sample grid point and applied it to all data sets."
Please elaborate. Is it possible you include in your ice mask model grid cells that were not land ice in the original model? If so, how could this affect your results?

We selected three CMIP5 models: GFDL ESM2M, HadGEM2 CC, and GISS R to represent low, medium, and high overall scoring models. We applied to each model both the interpolated mask from Zwally et al. 2012 and from the model output itself. We then calculated the resulting AIS-integrated SMB time series (fig. 1). For each model, the Zwally interpolated mask results in a lower AIS-integrated mean value – a result that benefits models that overestimate SMB – due, in large part, to the exclusion of ice shelves in this mask. For our work, we aim to derive the sea level contribution of future SMB in CMIP models. Because SMB on ice shelves does not impact sea level, we find that the logical choice to fulfill our objective is to use the Zwally interpolated mask.

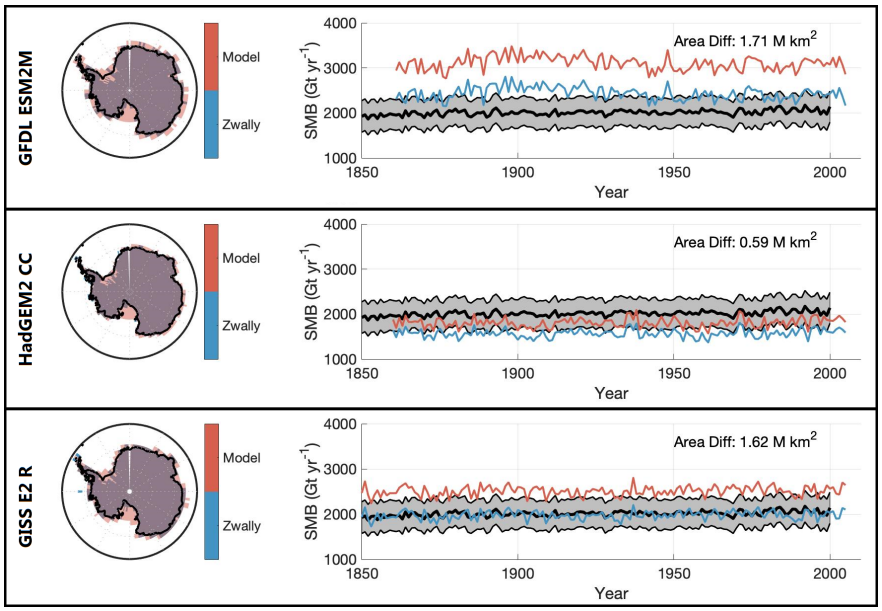

Figure 1: Masking comparison of the model mask in red/pink and the Zwally interpolated mask in blue with overlapping regions appear purple in the left column. Time series comparison of AIS-integrated SMB with the model mask applied in red and the Zwally interpolated mask applied in blue in the right column. The black line represents the reconstruction AIS-integrated SMB with the uncertainty in grey. The area differential between the two model masks is displayed in the upper right hand corner in the right column. **Top** GFDL ESM2M is a poor overall scoring model. **Middle** HadGEM2 CC is a medium overall scoring model. **Bottom** GISS R is a good overall scoring model.

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. 2 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?*

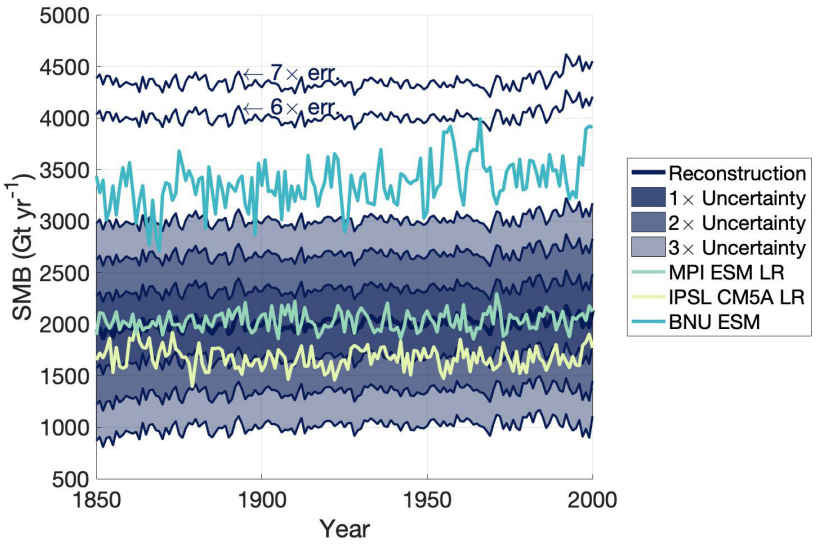

Figure 2: Time series of the reconstructed AIS-integrated SMB time series (purple) with 1×, 21×, and 31× the uncertainty in dark purple, medium purple, and light purple, respectively. Three model AIS-integrated SMB time series, MPI ESM LR (green), IPSL CM5A LR (yellow), and BNU ESM (cyan) have been plotted 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.

At line 109, we changed "For example, if a model time series was fully captured within 2× the reconstruction uncertainty, the model would receive a score of 2" to "the minimum possible score, then, is one, for a model who represents SMB that fits entirely within 1× the reconstruction uncertainty." We have also included Fig. 2 with the associated caption at line 110.

"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?)*

At line 127, we have changed "EOF analysis, as applied to these annual data, involves finding what spatial SMB patterns explain the highest variance in the AIS-integrated SMB time" to "EOF analysis maps the spatial pattern of a variable associated with the highest temporal variance of another variable. As applied to these annual data, EOF analysis maps the spatial pattern of sea level pressure associated to the highest variability in SMB integrated over the AIS."

"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?*

At line 130, we added the text "The top three modes of variability 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?*

From line 131 to line 133, we replaced "We then sorted the top modes of variability for each model based on smallest difference thus giving the models the 'benefit of the doubt.'" with "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:

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

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.*

To address this point, we performed an analysis wherein we isolated the outliers for each criterion score and removed them from consideration in the rescaling and retroactively assigned them a score of 10. This generated a score range of 1 to 10 for all non-outlier models and a score of 10 for all outlier models for each of the five criteria. We then took the average across the five criteria to generate the overall score for each model. Fig. 3 shows the overall scores for CMIP5 using the original scoring metric on the left and the outlier-removal-based scoring on the right. The final score of a few models did change and the mean overall score increased slightly, but the top 90th percentile and bottom 10th percentile of models remained in the same order. We have changed our scoring metric to reflect this newer approach. At line 142, we changed "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." to "After compiling scores for all five of the aforementioned scoring criteria, we removed from consideration any outliers and normalized each set of scores to be on a scale from one to ten to ensure that each criterion was equally weighted. After this normalization, the outliers for any given criterion were retroactively assigned a score of ten for that criterion."

"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)*

We appreciate the reviewer's comment here. We have performed some analysis aimed at addressing the robustness of choosing a small subset of models. At line 149, we included the text "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

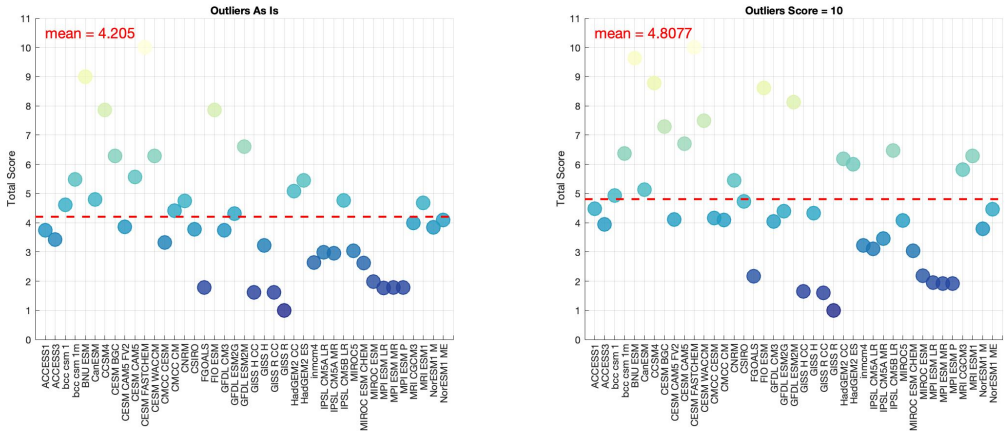

Figure 3: Caption

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. To shore up some of the language regarding our trend uncertainty calculation, from line 114 to line 118, we changed "The reconstructed trend uncertainties were calculated by performing a Monte Carlo simulation assuming a normal distribution of SMB values centered around the reconstructed SMB value with a standard deviation of the reconstruction uncertainty for each year. From those distributions, we generated 10,000 simulated SMB time series based on the reconstruction and calculated the trends for each. The standard deviation in trend, then, is the reconstructed trend uncertainty" to "We performed a Monte Carlo simulation wherein we assumed a normal distribution where the standard deviation of the distribution is equal to the reconstruction uncertainty of possible SMB values for each year. We then 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." Additionally, at line 114, we included the text "The first two of these three time slices confirm the robustness of the trends with longer periods for trend analysis. The last time slice, 1950-2000, allows us to view SMB in the context of significant anthropogenic warming. However, the large interannual 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. As a result, we have used the original reanalysis product for the temporal variability criterion and added to the text "For this criterion, we used the original MERRA-2 reanalysis precipitation minus evaporation data (1980-2019). Likely due to sampling only 53 ice core sites, the reconstruction produced a relatively low variability record. The reconstructed variability at any location can only be as large as the maximum variability in the ice cores. Thus, undersampling regions of stronger interannual variability will dampen the variability signal in the reconstruction. Analyses of the AIS-integrated SMB mean value and trend show that the reconstruction is generally in line with the literature (**?**)."

"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.*

At line 270, we have added a separate "Discussion" section and many parts of the EOF analysis discussion that were in the results section have been moved here.

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

At line 205, we have changed "The Amundsen Sea Low, which represents the pole of circulation variability in Antarctica ..." to "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 Antarctica 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. At line 206, we have added "High values on the EOF map indicate regions that explain large amounts of the variability in AIS SMB." From line 211 to line 213, we have change "The second mode of variability represents high variability in West Antarctica and the Antarctic Peninsula. This could be caused by the topography in these regions which can induce large amounts of snowfall" to "Looking at mode 2, 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 have multiplied the appropriate model EOFs by -1 to make the comparison easier (fig. 4).

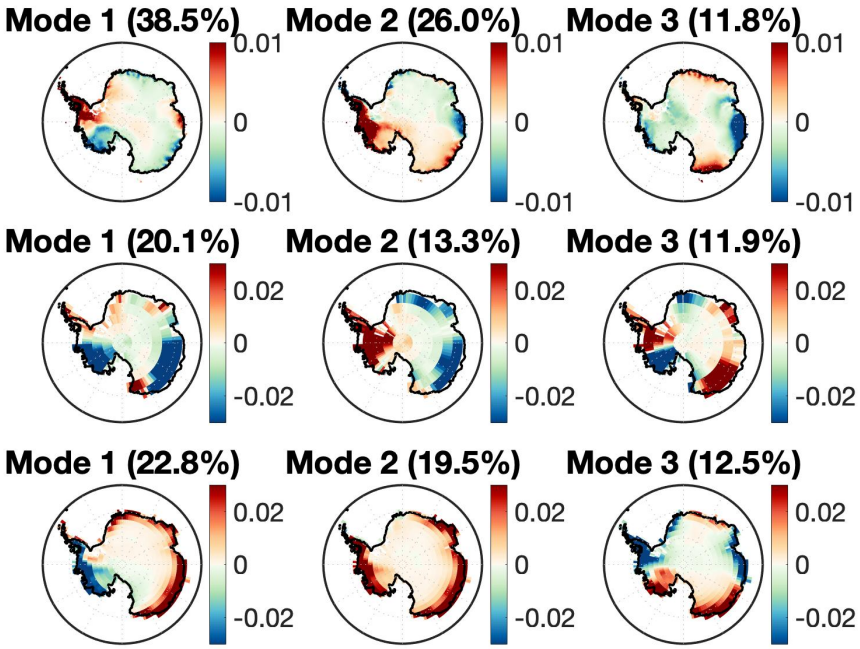

Figure 4: EOF analysis plots of the top 3 modes of variability for **A** the reconstruction, **B** a relatively high scoring model (CMCC CM), and **C** a low scoring model (CESM1 WACCM).

Generally, though, 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.*

At line 265 and 295, we have changed all instances of $^{\circ}C$ to $^{\circ}K$. We have also changed fig. 9 to reflect SMB changes per $^{\circ}$K. Changing the temporal variability criterion did, in fact, result in a change in the top four scoring models. At line 197, we included the text "(For comparison, the reconstructed normalized and detrended SMB standard deviation is about 2.9 Gt yr^{-1}.)".

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"*

From line 17 to line 18, we have changed the text from "Integrated over the grounded Antarctic ice sheet (AIS), the blowing snow and runoff terms are negligibly small (**?**)" to "We neglect blowing snow and runoff and estimate SMB as precipitation minus sublimation (**?**)."

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

At line 22, we have changed "Over longer time scales..." to "Over longer (~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"*

From line 29 to line 30, we have changed "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))" to "Synoptic-scale variability induces a strong regional variability of the SMB (Fyke et al. (2017); Marshall et al. (2017))".

"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. From line 31 to line 33, we have rewritten this sentence to reflect this more accurately: "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 undetectable due to the significant natural SMB variability (**?**). 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?*

From line 38 to 39, we have changed the text from "...but many of those models tend to overestimate annual precipitation values due to poor representation of coastal topography" to "...but many of those models tend to overestimate annual precipitation values 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 (**?**)."

"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?*

At line 41, we have changed the reference here to Palerme et al. (2017) to 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: at line 46, we have fixed this typo as follows: "Barthel et al. (2019) investigated the Ice Sheet Model Intercomparison Project for CMIP6 to determine..." Addressing the second point, from line 48 to line 49: we haved rephrased this sentence to "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 has been moved to line 58 at the beginning of the SMB Reconstructions 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.*

From line 50 to 52, we have changed "These reconstructed data sets now allow for a new avenue to investigate the ability of GCMs to capture SMB into the more distant past (**?**)" to "These reconstructed data sets now allow for

a new avenue to investigate the ability of GCMs to capture SMB into the more distant past (**?**) – an avenue that we leverage for climate model evaluation of AIS SMB to compare the suite of CMIP5 and CMIP6 climate models to this new SMB reconstruction" and removed lines 54-55.

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

At line 66, we have added "... 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"*

Large parts of this section were removed for succinctness.

"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 have removed the sentence "Global climate models tend to show higher skill at representing interannual variability compared to regional climate models (**?**)" at line 79. We have also adjusted the following sentence to "In this work, we use global climate models due to their ability to project decades to centuries into the future. As such, the objective of this work is to guide the selection of GCMs for ice sheet modelers to investigate the global impacts of changing ice sheets" to stress our other main reasons for using global climate models for comparison.

"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?*

At line 81, we have changed this sentence to "... 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 have changed "We downloaded CMIP5 and CMIP6 precipitation and evaporation/sublimation data with monthly resolution in units of kg m^{-2} s^{-1}. After calculating SMB as precipitation - evaporation/sublimation, we converted these to annual time scales and integrated them across the AIS using the Ice Sheet Mass Balance Inter-comparison Exercise Team's (IMBIE Team) AIS grounded ice sheet masks and units of Gt yr^{-1} by multiplying each grid cell by its area, converting s^{-1} to yr^{-1}, and converting kg to Gt (1 Gt = 10^{12} kg)**?**" to "We downloaded CMIP5 and CMIP6 precipitation and evaporation/sublimation data with monthly resolution and, after calculating SMB as precipitation - evaporation/sublimation, converted an annual time scale and integrated across the AIS using the Ice Sheet Mass Balance Inter-comparison Exercise Team's (IMBIE Team) AIS grounded ice sheet mask."

P4 L99-100

remove parentheses

From lines 99 to 100, we have removed 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.*

At lines 10, 66, 107, 108, 233, 241, 242, 248, 249, 251, and 297, we have changed "magnitude" to "mean value." At line 107, we also included the text "... (mean value referring to the value obtained by integrating SMB over the entire AIS)..." to define 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?*

At line 114, we have changed the last time period to "1950-2000" to correct this typo.

"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"*

From line 119 to 120, we have changed "To score the time series variability, we detrended and normalized each time series to separate the SMB trend from its absolute magnitude using:" to "For temporal variability, 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 detrend and normalize each time series, then, to separate the SMB variability from its mean value, we performed the following analysis:"

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

At line 126, we added "on annual data from 1850-2005" after "... we performed an empirical orthogonal function (EOF) analysis..."

"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?*

At line 128, we expanded 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)*

At line 157, we have added the paragraph "The final overall scores are an average of all the scores from all five criteria. After performing the analysis outlined in the Methods section the top 90th percentile overall scoring models were determined to be GISS E2 H, GISS E2 R, MPI ESM LR, and MPI ESM MR. These four models have been added in retroactively to figures 2-3 for comparison of their performance in each scoring criterion relative to the rest of the CMIP model suites."

"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*

From line 146 to line 147, we have removed the sentence "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."

" 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*

From lines 157 to 159, we have changed "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). The highest absolute SMB trends are around the coastal regions of East Antarctica and the Antarctic Peninsula (Fig. 1B)" to "Along with higher SMB values, the coastal regions of East Antarctica and the Antarctic Peninsula also show the highest absolute SMB trends (Fig. 1B)."

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.

At line 270, we have added a "Discussion" section in which we have added the following text "The CESM Large Ensemble (CESM-LENS) is an experiment wherein the Community Earth System Model Version 1 (CESM) is run 40 times with random temperature perturbations at the level of round-off error applied in 1920 (**?**). Because of its large number of ensemble members, the CESM-LENS experiment is useful for quantifying the role of internal variability. Only 35 of the original 40 ensemble members contain the necessary information for assessing AIS SMB. Figure 5 shows the final scores of the five CESM simulations that are included in the CMIP5 suite of models as well as the final scores of the CESM-LENS experiment. The final scores for the CESM-LENS model runs are calculated the same way for all model criteria except for AIS-integrated trend. Because these runs only differ after 1920, we only use the third time slice (1950-2000) to assess the quality of trend reproduction.

The final scores of the five CMIP5 CESM model runs range from 3.99 to 9.74 while the final scores of the 35 CESM-LENS runs range from 1.32 to 5.96. Given that the scores range by 5.74 and 4.65 for the CMIP5 CESM runs and the CESM-LENS runs, respectively, it is reasonable to conclude that internal variability plays as significant a role in determining final score as model parameterizations.

A major caveat of this finding, though, is that the CESM-LENS runs and the reconstruction only overlap from

1920-2000. This will likely most significantly impact the assessment of the trend and EOF analyses. With longer model runs, the CESM-LENS ensemble members would likely deviate further from one another and exacerbate the spread in their final scoring.

With that, internal variability plays a significant role in our AIS SMB assessment. Some models within the CMIP5 and CMIP6 frameworks, like CESM, have many ensemble members for certain parameterizations. However, not all models – and even not all model parameterizations – have multiple ensemble members. As such, performing a direct comparison of the models using the ensemble mean would not necessarily yield an accurate result as models with more ensemble members would have their final score shifted significantly while the same is not true for models with a single ensemble member. For considering using GCMs for AIS SMB analysis, then, we strongly suggest taking into account the fact that internal variabiltity could be playing a strong role in some models final score and that the number of ensemble members available should be considered along with the final score.

We have also added Fig. 5 to the supplementary material for reference.

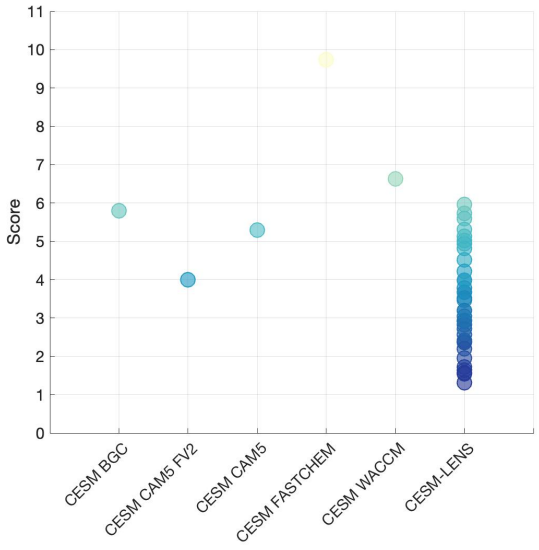

Figure 5: Final scores of the five CESM models from CMIP5 compared to the CEMS-LENS simulations.

While we do appreciate the reviewer's concern regarding model resolution, we feel that assessing the full High-ResMIP is beyond the scope of this work. This model intercomparison project is comprised of not-fully-coupled climate models and the historical data only dates back to 1950. As such, we feel that direct comparison with the entire CMIP5 and CMIP6 suites would not accurately reflect a robust analysis. We have added a "Discussion" section at line 270 with the text "The CMIP5 and CMIP6 models vary in resolution from rougly 0.75°×0.75° to 3°×3°. Figure 6 shows a scatter plot of resolution versus total score. Resolution, here, is the latitudinal resolution multiplied by the longitudinal resolution such that a model with latitude/longitude resolutions 0.9375°/1.25° would have a resolution of 1.1719°. A linear regression yields a correlation of R = -0.40 with 95% confidence intervals of -0.62 and -0.17. From this, there is a statistically significant negative correlation between resolution and total model score. This result is further exemplified when looking at total scores from the same model run at different resolutions. CESM CAM5, IPSL CM5A MR, MPI ESM MR, CESM2, CESM2 WACCM, and MPI ESM2 HR all perform worse than their coarser resolution counterparts – CESM CAM5 FV2, IPSL CM5A LR, MPI ESM LR, CESM2 FV2, CESM2 WACCM FV2, and MPI ESM2 LR. Because so many models close to 1°/1° resolution and there is large spread in these models' final scores, we also divided the models into two groups, finer and coarser than 1.25°/1.25°, and performed the same regression analysis. Figure 7 shows the coarser resolution models have a correlation of R = -0.14 with 95% confidence intervals of -0.51 and 0.24 while finer resolution models have a correlation of R = -0.06 with 95% confidence intervals of -0.38 and 0.26. From this, we conclude that there is no significant correlation between model resolution and total score."

We have also added Figures 6 and 7 to the supplementary material for reference.

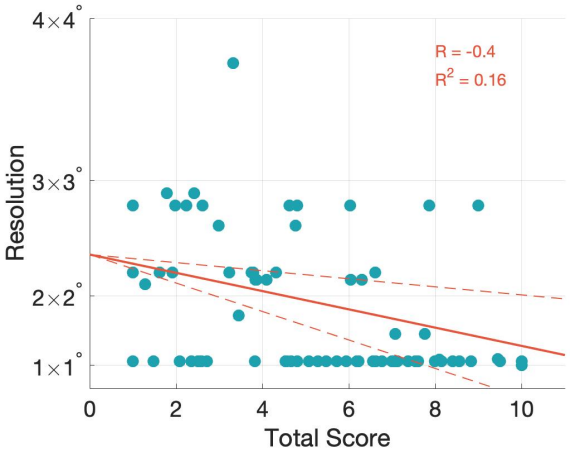

Figure 6: A scatter plot of total score versus model resolution in lat×lon. The correlation is 0.45 and the variance is 0.21.

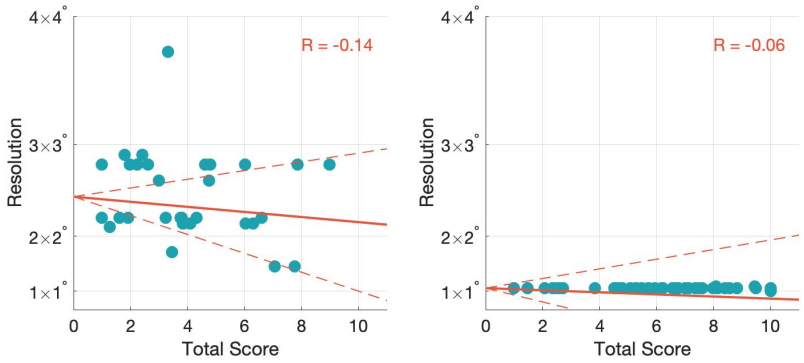

Figure 7: A scatter plot of total score versus model resolution in lat×lon. The correlation is 0.45 and the variance is 0.21.

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.

At line 54, we changed "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" to "For this work, we investigate AIS SMB in GCMs. GCMs are, compared to RCMs, incredibly low resolution which, thus, making it difficult for them to reproduce the detailed SMB response. RCMs have been shown to be more accurate

in capturing AIS SMB **?**, however, due to their high resolution, RCMs are also relatively computationally expensive to run for long periods (~100s of years). Because one of the goals of this paper is to investigate the future of SMB over Antarctica, we analyze GCMs for their ability to simulate these long-term climate effects. As RCMs are by definition regional, they need boundary forcings, which adds an additional layer of complexity and a source of uncertainty to running RCMs into the long-term future. An additional reason we choose to analyze GCMs is simply to figure out which GCMs perform best at capturing these SMB phenomena. There has been extensive work investigating SMB in RCMs (**?**; **?**; **?**), but relatively little looking at GCMs. To investigate the global coupled response to future SMB changes, one needs GCMs. As such, this work is meant to inform modelers who are concerned with global ramifications of changing AIS SMB."

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. As such, we have also added to the text at the end of the introduction "We use criteria that look exclusively at SMB and not its underlying causes. Several studies such as **?** and **?** have investigated the impacts of causes like thermodynamical phenomena and sea ice extent on SMB, but, here, we are looking solely at a model's ability to reproduce SMB without necessarily capturing the exact root causes." We also appreciate the suggestion of the leave-one-out though we feel that, since the criteria are all equally weighted, that no single criterion would have more or less of an impact on the overall scoring. To that point, we have added to the supplemental material Fig. 8 which denotes the overall scores of each CMIP5 model using a weighted scoring system (as we have) versus using an absolute scoring system. For many models, the score did not change significantly. For the models where the score did change, scores decreasing by switching to the absolute scoring system were almost completely offset by score increasing by switching. As a result, the average overall score is roughly the same between the two scoring systems.

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,

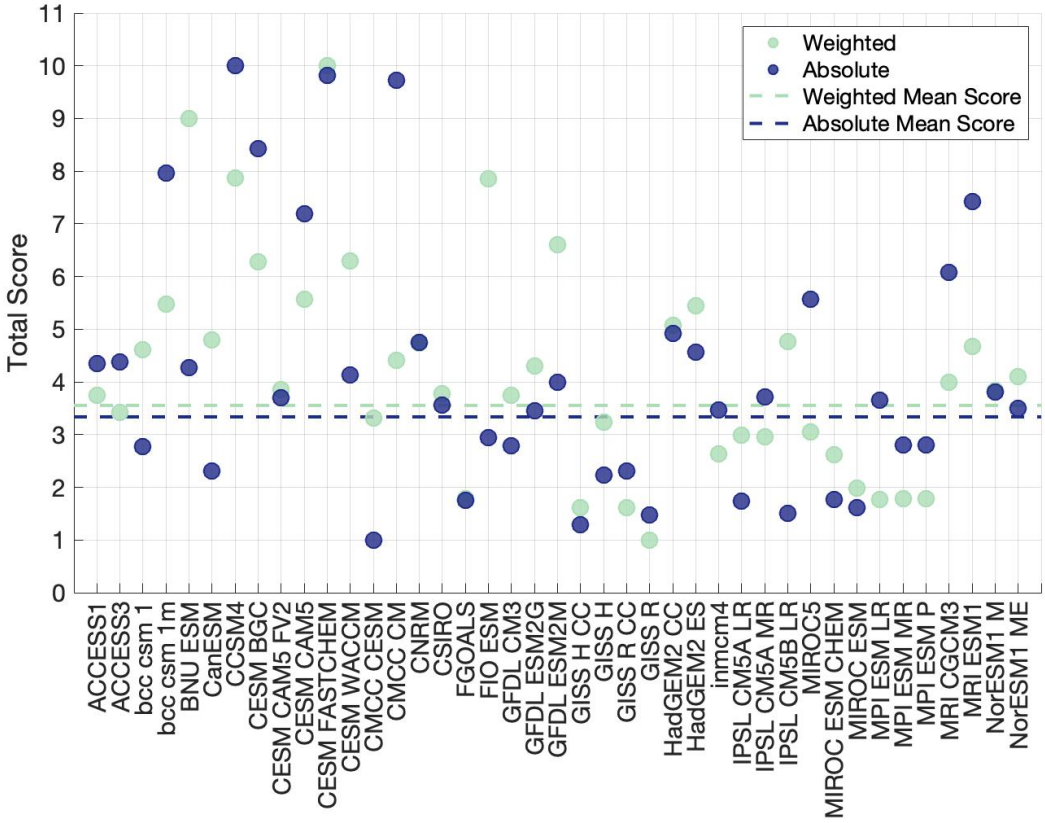

Figure 8: CMIP5 model scores using a weighted scoring system (light green) versus an absolute scoring system (dark blue).

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.

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 have added Fig. 2 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, at line 270, we have added a "Discussion" section with the text "The CESM Large Ensemble (CESM-LENS) is an experiment wherein the Community Earth System Model Version 1 (CESM) is run 40 times with random temperature perturbations at the level of round-off error applied in 1920 (**?**). Because of its large number of ensemble members, the CESM-LENS experiment is useful for quantifying the role of internal variability. Only 35 of the original 40 ensemble members contain the necessary information for assessing AIS SMB. Figure 4 in Supplementary shows the final scores of the five CESM simulations that are included in the CMIP5 suite of models as well as the final scores of the CESM-LENS experiment. The final scores for the CESM-LENS model runs are calculated the same way for all model criteria except for AIS-integrated trend. Because these runs only differ after 1920, we only use the third time slice (1950-2000) to assess the quality of trend reproduction.

The final scores of the five CMIP5 CESM model runs range from 3.99 to 9.74 while the final scores of the 35 CESM-LENS runs range from 1.32 to 5.96. Given that the scores range by 5.74 and 4.65 for the CMIP5 CESM runs and the CESM-LENS runs, respectively, it is reasonable to conclude that internal variability plays as significant a role in determining final score as do model parameterizations.

A major caveat of this finding, though, is that the CESM-LENS runs and the reconstruction only overlap from 1920-2000. This will likely most significantly impact the assessment of the trend and EOF analyses.

With that, internal variability plays a significant role in our AIS SMB assessment. Some models within the CMIP5 and CMIP6 frameworks, such as CESM1-CAM5, have many ensemble members. However, not all models – and even not all model versions – have multiple ensemble members. As such, performing a direct comparison of the models using the ensemble mean would not necessarily yield an accurate result as models with more ensemble members would have their final score shifted significantly while the same is not true for models with a single ensemble member. For considering using GCMs for AIS SMB analysis, then, we strongly suggest taking into account the fact that internal variability could be playing a strong role in some models final score, and that the number of ensemble members available should be considered along with the final score.

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.

Starting at line 275, we have added "Using this scoring system, we found that the top 90th percentile models were GISS H CC, GISS R CC, GISS R, MPI ESM LR, MPI ESM MR, and MPI ESM P of CMIP5 and CESM FV2 and MPI ESM2 LR of CMIP6. A similar study in **?** found ACCESS1-3, ACCESS1-0, CESM BGC, CESM CAM5, NorESM1-M, and EC-Earth to most accurately capture SMB from the reanalysis product, ERA-Interim. They focused their investigation into more atmospheric and oceanic dynamics (sea ice extent, sea surface temperature, sea surface pressure, precipitable water, 850 hPa temperature) and were comparing models directly to a reanalysis product. **?**, another study with a similar goal of analyzing SMB performance among GCMs selected CCSM4, MIROC ESM CHEM, and NorESM1-M as their top three performing models for Antarctica. They ruled out both the GISS and MPI modeling groups due to their initial selection criteria and were also looking more at the impacts thermodynamical processes on 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. At line 307, we have added "One final major caveat 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 (**?**; **?**), 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 have removed this statement and replaced it with further justification for the use of GCMs at line 79: "For this work, we investigate AIS SMB in GCMs. GCMs are, compared to RCMs, incredibly low resolution which, thus, making it difficult for them to reproduce the detailed SMB response. RCMs have been shown to be more accurate in capturing AIS SMB **?**, however, due to their high resolution, RCMs are also relatively computationally expensive to run for long periods (~100s of years). Because one of the goals of this paper is to investigate the future of SMB over Antarctica, we analyze 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 into the future for century-length timescales. An additional we analyze GCMs is simply to figure out which GCMs perform best at capturing these SMB phenomena. There has been extensive work investigating SMB in RCMs (**?**; **?**; **?**), but relatively little looking at GCMs. To investigate the global coupled response to future SMB changes, one needs GCMs. As such, this work is meant to inform modelers who are concerned with global ramifications of changing AIS SMB. 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. We use criteria that look exclusively at SMB and not its underlying causes. Several studies such as **?** and **?** have investigated the impacts of causes like thermodynamical phenomena and sea ice extent on SMB, but, here, we are looking solely at a model's ability to reproduce SMB without necessarily capturing the exact root causes." 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 are now using the original reanalysis data and make a strong note about the interannual variability issues in the reconstruction. At line , we changed "To score the time series variability, we detrended and normalized each time series to separate the SMB trend from its absolute magnitude using:" to "For temporal variability, 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 detrend and normalize each time series, then, to separate the SMB variability from its mean value, we performed the following analysis: ...". 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.

Reviewer #3

We thanks the reviewer for providing a lot of thoughtful insight and asking a lot of very good questions. To ad-

dress the predominant remarks, here, we will make the necessary adjustments to the wording to make the paper more accurate and comprehensible. We have also added Fig. 2 and eq. (1) to help with this process. Additionally, we would like to thank the reviewer for their helpful comments on figure adjustments. We appreciate how important for overall comprehension good figures are and we will strive to make ours as palatable as possible.

Section 2.1 SMB reconstructions:
This section summarizes the methods used by Medley & Thomas in creating their ice-core derived SMB reconstructions. I found that I was confused by how these were created, as if all the details might be correct but without the "big picture" context. Once I read the abstract for Medley & Thomas, however, I understood. This section can be re-written (and shortened) to better summarize the reconstructions. If the reader wants all the details of the SMB reconstruction, he/she can refer to Medley & Thomas for that.

From line 58 to 78, we have changed "In this paper, we use the AIS SMB reconstruction generated by Medley & Thomas (2019) ... and refer to it as "reconstruction" to "In this paper, we use the AIS SMB reconstruction generated by **?**. The authors synthesize SMB time series from an extensive ice-core database with reanalysis-derived spatial coherence patterns to generate a continent-wide AIS SMB data set. While **?** compared three reanalysis products, they also show that MERRA-2 performed better than the other two reconstructed products in matching observations. As such, we will use the MERRA-2 based data set as a proxy for all three reconstructions and refer to it as "reconstruction."

Section 4 Results Lines 167-168: "The interquartile ranges for CMIP5 and CMIP6 are 1727 to 2282 Gt yr1 and 1728 to 2229 Gt yr1, respectively, with means of 1940 Gt yr1 and 2115 Gt yr1, respectively."
What is the take away? For example, something like "CMIP5 models tend to have a slightly smaller mean AIS SMB with a larger range than the CMIP6 models (Table XXX)". The figure shows this, and a table could present the quantitative results for any readers that want them. Similarly for the other results throughout this section.

Section 4 results: AIS SMB sensitivities to changes in temperature: Lines 294-297: "Comparing the projected change in SMB per degree warming between the emission scenarios gives median sensitivities of 64±80 Gt C1, 57±33 Gt C1, and 78±15 Gt C for RCPs 2.6, 4.5, and 8.5, respectively, for the best scoring models. Combined, these data tell us that for stronger emission scenarios, the AIS SMB response will be stronger in both magnitude and trend."
The results do not back up this claim. The mean sensitivity for RCP4.5 is lower than that for the RCP2.6! Furthermore, there is no indication here if the differences in the means are statistically significant or not. If model sensitivities of AIS SMB-Temp change with different scenarios – this is a very interesting result (and needs to be backed up better if it is your result – with some explanation to the apparent contradiction of the RCP4.5 having the lowest sensitivity – or maybe there's a typo?). If so, some discussion about what mechanism might explain this. For example, AIS SMB is driven by precipitation and evaporation/sublimation. Are there processes in changing climate that might drive changes in precipitation in addition to changes in temperature? Changes in synoptic weather patterns? Or? Do sensitivities of AIS-SMB to changes in CO2 remain same in all scenarios or do these change? (or do changes in CO2 combine temperature and precipitation sensitivities into "one" proxy for these?)

We would like to apologize here. The reviewer is correct in that the results are not statistically different between the three forcing scenarios. This text is based off an early result that had since been updated in the figure and we failed to update the text as a result. To reflect the updated figure, from line 294 to 297, we have changed "Comparing the projected change in SMB per degree warming between the emission scenarios gives median sensitivities of 64 ± 80 Gt $^{\circ}C^{-1}$, 57 ± 33 Gt $^{\circ}C^{-1}$, and 78 ± 15 Gt $^{\circ}C^{-1}$ for RCPs 2.6, 4.5, and 8.5, respectively, for the best scoring models. Combined, these data tell us that for stronger emission scenarios, the AIS SMB response will be stronger in both magnitude and trend" to "Comparing the projected change in SMB per degree warming between the emission scenarios gives median sensitivities of 64 ± 80 Gt $^{\circ}K^{-1}$, 57 ± 33 Gt $^{\circ}K^{-1}$, and 78 ± 15 Gt $^{\circ}K^{-1}$ for RCPs 2.6, 4.5, and 8.5, respectively, for the best scoring models. These results are not statistically significantly different from one another across forcing scenarios and indicate that there is no difference in the sensitivity response to changes in temperature

between the three forcing scenarios." Additionally, from line 261 to line 265, we have changed "Box plots of modeled SMB sensitivity to changes in temperature (i.e. how much SMB will change per degree warming) show that SMB responds differently in different warming scenarios (Fig. **??**). The CMIP5 models project that each warming scenario with higher CO_2 concentrations will see greater SMB sensitivity to increases in temperature than those with lower CO_2 concentrations. While the ranges differ from scenario to scenario, the projected sensitivity medians for RCPs 2.6, 4.5, and 8.5 are 101.7 Gt $°C^{-1}$, 111.2 Gt $°C^{-1}$, and 128.2 Gt $°C^{-1}$, respectively" to "Box plots of modeled SMB sensitivity to changes in temperature (i.e. how much SMB will change per degree warming) are shown in Fig. **??**. The projected sensitivity medians for RCPs 2.6, 4.5, and 8.5 are 101.7 Gt $°C^{-1}$, 111.2 Gt $°C^{-1}$, and 128.2 Gt $°C^{-1}$, respectively. These results are not statistically significantly different from one another indicating no significant response by SMB to increased warming scenarios."

Line 71 "calculated spatial sampling uncertainty is based"
should be "calculated spatial sampling uncertainty based"

Large parts of this section were removed for succinctness.

lines 84-87
How many CMIP6 models? Later it is claimed that there were so few CMIP6 models available that statistics are not robust for that set...yet the numbers here (53 models, 28 independent and of these 30/19 are CMIP5 which leaves at least 20 CMIP6?).

We have updated the number of CMIP6 models to 40 which more accurately reflects the current scope of the project. Numbers and references to this data set have been changed throughout the manuscript.

Line 114
4 Repeat 1850-2000...think you mean 1950-2000 in second instance

At line 114, we have corrected this typo to "1950-2000."

Language is a bit cumbersome and over the top in 3.1 (AIS-integrated SMB criteria)

We have rewritten large portions of section 3.1 to improve comprehensibility. We have added to the text "To score the time series mean value, 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. The minimum possible score, then, is one, for a model that represents SMB within 1× the reconstruction uncertainty. Fig. 2 illustrates that a model that fits entirely within 1× the reconstruction uncertainty (dark purple) – MPI ESM LR – would receive a score of 1. A model that fits within 2× the reconstruction uncertainty (medium purple) – IPSL CM5A LR – would receive a score of 2. A poorer scoring model, BNU ESM, would receive a score of 6."

Got lost again in 3.2 Maybe a couple equations and a map (example) would help. I have the sense it's pretty straightforward but description overcomplicates

To clarify the wording in this section, from line 131 to line 133, we replaced "We then sorted the top modes of variability for each model based on smallest difference thus giving the models the 'benefit of the doubt.'" with "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:

$$\text{difference number} = \sum_{lat} \sum_{lon} \left| \text{reconstruction}_{lat,lon} - \text{model}_{lat,lon} \right| \tag{2}$$

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 recon-

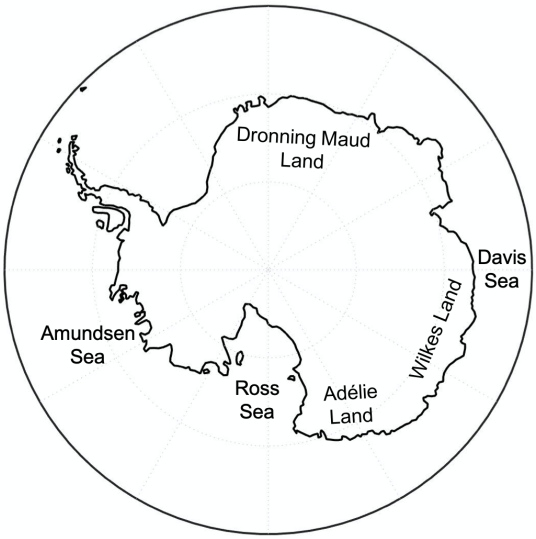

Figure 9: Map of the AIS.

struction 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."

Figure 2
Can't see dots in Figure 2B (they overlap too much?)

All markers denoting the top scoring models have been changed to a single color throughout the figures. While this means that one cannot differentiate these models from one another, we feel is allows for easier overall readability of the figures. Additionally, in the appropriate figure captions, we have added wording that addresses the fact that some markers may be close to, if not entirely, overlapping.

Line 190
Just because there are fewer models does not necessarily imply that the spread in trends will be less! For example, one could pick CMIP5 models and only use a subsampling and still get same spread if the models selected have large range in trends.

As we have added numerous additional CMIP6 models, this section has been rewritten to reflect the updated analysis.

Line 200
Not only melt and discharge distributed unequally, but also accumulation (precipitation)!

At line 200, we have changed this sentence to "... spatial variations in SMB are also important in AIS SMB representation in models as precipitation, melt, and discharge are not distributed equally."

Lines 213-216
If using place names, have a map showing where these are

We have added to the supplementary Fig. 9.

Lines 235-236
already defined RCP earlier, no need to do so again here...

At lines 235 to 236, we have removed "Future CMIP5 projections are created in the context of warming scenarios called Representative Concentration Pathways (RCPs). The RCPs we used to investigate SMB projections are RCP2.6, RCP4.5, and RCP8.5 which have progressively higher CO_2 concentration projections and, thus, higher projected global warming."

Conclusions
The recent and similar work of Barthel et al (2019) is mentioned (lines 45-49). Bartel et al was addressing a related albeit slightly different question (than this submission), namely "which climate models would best be used to force a stand-alone ice sheet model?" and compared climate model output to atmospheric reanalysis products. Did their suggestions (best models for stand alone Antarctic Ice Sheet forcing) differ than yours (best models for AIS SMB in the coupled system) or were they similar? Why do you think that is? (perhaps in conclusions – and only need a couple of sentences). Essentially tie in the results of this submission to other current related results.

We have added to our conclusion section "A similar study in **?** found ACCESS1-3, ACCESS1-0, CESM BGC, CESM CAM5, NorESM1-M, and EC-Earth to most accurately capture AIS sea level pressure, 850 hPa air temperature, precipitable water, and ocean conditions – all of which impact AIS SMB to varying degrees. They focused their investigation into more atmospheric and oceanic dynamics (sea ice extent, sea surface temperature, sea surface pressure, precipitable water, 850 hPa temperature) and were comparing models directly to a reanalysis product. **?**, another study with a similar goal of analyzing SMB performance among GCMs selected CCSM4, MIROC ESM CHEM, and NorESM1-M as their top three performing models for Antarctica. They ruled out both the GISS and MPI modeling groups due to their initial selection criteria and were also looking more at the impacts thermodynamical processes on SMB." We feel this adds context from two recent, prominent studies with similar characteristics in approach and overarching objective.

Figure 4
condense A-D onto one figure

We have combined all of the top scoring models into one panel so there are currently 3 panels: one for the CMIP5 variability distributions, one for the CMIP6 variability distributions, and one for the distribution of standard deviations.

Figure 5
Reconstruction EOFs are low enough that on scale plotted hard to see patterns. Recommend a different scale for reconstruction (and point out in figure caption). Also to help clarify, only need one legend for reconstruction (if you re-scale) and one for 6 panels of model (do not need 9 identical legend bars - extraneous). This will simplify.

We have reduced the scale on the reconstruction EOF by a factor of three. We have also made note of the change in the figure caption.

Figure 6
Yellow x's very difficult to see. Make more visible.

We have changed the yellow x's to red outlines to make the top scoring models more visible.

Figure 7
Hard to see differences from different scenarios (and until 2006 they are identical). Find a way to combine these three panels into one – this will give same information and also new, comparative information

We agree that it is a bit difficult to see accurately the differences between the scenarios in this figure. We have tried multiple ways to convey this information succinctly in a single frame to alleviate this issue but have repeatedly found that our attempts to do so only reduce the readability of the figure. For instance, combining the figure as is into one frame makes it such that the larger model spread (for all models) are difficult – if not impossible – to differentiate due simply to the fact that there is significant spread amongst all the models in every forcing scenario. We can, however, try to add a frame that just looks at the four best models in each of the forcing scenarios to be able to make direct comparison among a smaller subset of models.