What is the Surface Mass Balance of Antarctica? An Intercomparison of Regional Climate Model Estimates

Ruth Mottram[1], Nicolaj Hansen[1,2], Christoph Kittel[3], Melchior van Wessem[4], Cécile Agosta[5], Charles Amory[3], Fredrik Boberg[1], Willem Jan van de Berg[5], Xavier Fettweis[3], Alexandra Gossart[6], Nicole P.M. van Lipzig[6], Erik van Meijgaard[7], Andrew Orr[8], Tony Phillips[8], Stuart Webster[9], Sebastian B. Simonsen[2], and Niels Souverijns[6,10]

[1]DMI, Lyngbyvej 100, Copenhagen, 2100, Denmark
[2]DTU-Space, Kongens Lyngby, Denmark
[3]Laboratory of Climatology, Department of Geography, SPHERES, University of Liège, Liège, Belgium
[4]Institute for Marine and Atmospheric Research Utrecht, Utrecht University, Utrecht, the Netherlands
[5]Laboratoire des Sciences du Climat et de l'Environnement, LSCE-IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, Gif-sur-Yvette, France
[6]Department of Earth and Environmental Sciences, KU Leuven, Belgium
[7]Royal Netherlands Meteorological Institute, De Bilt, the Netherlands
[8]British Antarctic Survey, High Cross, Madingley Road, Cambridge, UK
[9]UK Met Office, FitzRoy Road, Exeter, Devon, EX1 3PB, UK
[10]Unit Remote Sensing and Earth Observation Processes, Flemish Institute for Technological Research (VITO), Mol, Belgium

Correspondence: Ruth Mottram (rum@dmi.dk)

Abstract. We compare the performance of five different regional climate models (RCMs) (COSMO-CLM^2, HIRHAM5, MAR3.10, MetUM and RACMO2.3p2) forced by ERA-Interim reanalysis, in simulating the near surface climate and surface mass balance (SMB) of Antarctica. All models simulate Antarctic climate well when compared with daily observed temperature and pressure with nudged models matching daily observations slightly better than free running models. The ensemble mean annual SMB over the Antarctic Ice Sheet (AIS) including ice shelves is 2329 ± 94 Gigatonnes (Gt) year^{-1} over the common 1987 to 2015 period covered by all models. There is large interannual variability, consistent between models due to variability in the driving ERA-Interim reanalysis. Mean annual SMB is sensitive to the chosen period, over our 30 year climatological mean period (1980 to 2010), the ensemble mean is 2486 Gt year^{-1}. However, individual model estimates vary from 1961 ± 70 to 2519 ± 118 Gt year^{-1}. The largest spatial differences between model SMB estimates are in West Antarctica, the Antarctic Peninsula and around the Transantarctic mountains. We find no significant trend in Antarctic SMB over either period. Antarctic ice sheet (AIS) mass loss is currently equivalent to around half a millimetre year^{-1} of global mean sea level rise (Shepherd et al., 2018b) but our results indicate some uncertainty in the SMB contribution based on RCMs. We compare modelled SMB with a large dataset of observations which, though biased by undersampling, indicates that many of the biases in SMB are common between models. A drifting-snow scheme improves modelled SMB on ice sheet surface slopes with an elevation between 1000 and 2000 m where strong katabatic winds form. Different ice masks have a substantial impact on the integrated total SMB and along with model resolution is factored into our analysis. Targeting undersampled regions with high precipitation for observational campaigns will be key to improving future estimates of SMB in Antarctica.

1 Introduction

The Antarctic Ice Sheet (AIS) is the largest body of freshwater on the planet and an important contributor to global sea level rise. It is also a significant part of the climate system contributing freshwater to the ocean and with high relief that influences atmospheric circulation. Studies by Rignot et al. (2011, 2019) and Shepherd et al. (2018a) showed the AIS to have had a net loss since at least 2002. Current estimates suggest that around 10 % of observed sea level rise since 1993 is from Antarctica, however that rate of contribution is also increasing (IPCC SROCC Chap. 4 Oppenheimer et al. (2019)). Most ice loss in Antarctica occurs as a result of submarine melting, that is melt at the water-ice interface underneath ice shelves, or by the calving of icebergs from ice shelves. Recent ice dynamics studies (DeConto and Pollard, 2003; Edwards et al., 2019; Sutter et al., 2016; Shepherd et al., 2018a) have shown that there is potential for rapid ice sheet loss owing to ice sheet dynamics that are currently poorly understood, especially in West Antarctica. Ice sheet models of the AIS have thus largely concentrated on parameterising sub-shelf and calving processes. However, surface mass balance (SMB), also known as surface mass balance or climate mass balance (Cogley et al., 2011), is also of crucial importance in controlling the stability and evolution of the vast ice sheet. Changes in precipitation and increases in surface melt and runoff will change the mass balance and therefore both ice dynamics and the sea level rise contribution from Antarctica in the future. Moreover there has been disagreement between studies focused on the SMB contribution to the total mass budget of Antarctica and therefore the contribution to sea level rise (Scambos and Shuman, 2016; Zwally et al., 2015), that makes it essential to understand potential biases and uncertainties.

SMB is the difference between accumulation and ablation at the surface of a glacier. In Antarctica, accumulation is derived primarily from solid precipitation, but on local or regional scales wind-driven processes can have a significant effect on accumulation rates. Surface ablation in Antarctica is primarily a result of erosion and sublimation due to the high winds and generally dry atmosphere (Scambos et al., 2012; Das et al., 2013; Agosta et al., 2019), although increasing melt rates are documented in some areas (Stokes et al., 2019). In the future, a "greenlandification" of the ice sheet climate is projected due to anthropologically induced climate change (Trusel et al., 2018). This will lead to more melt with more refreezing in the snowpack as well as increasing runoff.

It is important to distinguish between the continental grounded ice sheet and ice shelves when considering values for SMB integrated over a wider area whether regional or continent wide. Snowfall and melt on ice shelves is not directly relevant to sea level rise contributions as they are already floating but precipitation and ablation on grounded parts of the ice sheet is. As the models used in this study by and large do not distinguish between grounded and floating ice in their ice masks, in this paper when we refer to SMB over an area we include ice shelves unless specifically noted.

Currently, runoff is a relatively minor contribution (Lenaerts et al., 2019) to mass loss in Antarctica. Increasing snowfall, associated with higher saturated vapour pressure is expected to dominate future changes in SMB, compensating for the projected increase in surface runoff (Krinner et al., 2008; Lenaerts et al., 2016) but the balance between these processes is still a matter of debate. This makes it even more important to evaluate the effectiveness of modelled precipitation and sublimation across the continent to be able to estimate SMB at present. Accurate SMB estimates are required to both drive ice sheet dynamical models and to accurately partition sea level rise contributions determined from observations. SMB from regional climate mod-

els (RCMs) is also used to correct altimetry measurements by accounting for firn compaction processes for remote sensing applications.

The most common way to observe SMB is by geodetic mass balance stakes (Lenaerts et al., 2019) but this is challenging due to the size and environmental conditions in Antarctica and the most practical alternative is to use output from (high-resolution) RCMs to make continent-wide estimates. There are now an increasing number of RCMs downscaling Antarctic climate simulations available via the CORDEX (CoOrdinated Regional climate Downscaling EXperiments) database. CORDEX is a project of the World Climate Research Programme that aims to produce representative ensembles of regional climate models for different regions of the world. The purpose is to better understand regional climate change, assess regional impacts and improve adaptation to future climate conditions (http://www.cordex.org/).

In the polar regions, CORDEX simulations can also be used to assess the mass budget of the large polar ice sheets, but have not yet been evaluated together for Antarctica. Souverijns et al. (2019) made a 30 years hindcast with COSMO-CLM^2, and Agosta et al. (2019) estimated the SMB using MAR, while various versions of RACMO2 have been used to estimate the SMB of the AIS (Van Wessem et al., 2014; van Wessem et al., 2018). Both MetUM and HIRHAM5 have been run for the Antarctic domain but evaluation of the SMB results have not yet been published in peer review literature (Hansen, 2019). Here, we use the framework of the Polar CORDEX project to assess climate model performance in Antarctica for the period 1979-2018 derived from an ensemble of six simulations from five different RCMs. The RCMs cover a range of resolutions, physical and dynamical schemes in the atmosphere and types of surface and snow/ice schemes. This allows us to determine the relative importance of individual model components needed to accurately model the climate by comparing the modelled SMB against the sparse observational data-sets available in Antarctica. We also investigate some of the uncertainties within the individual models and between the ensemble members.

In this paper, we seek to quantify present-day Antarctic SMB and understand the sources of variation as a baseline to assess mass budget changes and better understand sea level rise observations and projections both directly in terms of the amount of meltwater added to oceans and indirectly as surface forcing for ice sheet dynamical models (Robel et al., 2019; Nowicki et al., 2016).

2 Methods

We compare six climate simulations made with five different RCMs (COSMO-CLM^2, HIRHAM5, MAR, MetUM, RACMO) in the newest available version of the given RCM. However, to provide backwards continuity, we also briefly compare three older versions that have been widely used in earlier studies, to examine how results have varied (or not) as RCMs have been developed. We assess the climate of Antarctica in the models and derive estimates for SMB. All models were forced on the lateral boundaries with the ERA-Interim climate reanalysis (Dee et al., 2011) but downscaling used different grids, over slightly different domains and at different resolutions with slightly different ice masks used in the different model versions (see A1). Simulations with MAR forced by different reanalyses (Agosta et al., 2019) found that results were rather similar to ERA-Interim but to exclude additional variability potentially introduced by using different boundary forcings, we chose to

use a single common reanalysis only. The MAR, RACMO and COSMO-CLM^2 models were nudged within the domain using upper air relaxation and MetUM was run as a 12 hour reinitialised hindcast. With this technique the model is run in weather forecast mode and restarted with new boundary conditions every 12 hours. The two versions (high and low resolution) of the HIRHAM5 model were allowed to run freely within the domain and forced only on the boundaries.

We first give a brief overview of each of the participating models, summarised in Table 1. The CORDEX protocol (Christensen et al., 2014) prescribes a simulation domain for Antarctica with a minimum common analysis extent and a resolution of 0.44°. Lucas-Picher et al. (2012); Lenaerts et al. (2012b); Franco et al. (2012) and van Wessem et al. (2018), among others, have found that a higher spatial model resolution gives more physically plausible results especially with respect to precipitation processes in areas with steep terrain. Hence, several participating groups have chosen to run their RCMs at higher spatial resolution. To quantify both the absolute and relative integrated and basin scale SMB for the continent, we compare outputs from the different models with each other and the ensemble mean. We also evaluate the models with SMB observations (including ice cores and stakes) and near-surface climate observations (surface pressure, temperature and wind speed) measured across the continent. Unfortunately, as we are constrained to using existing simulations, the models cover slightly differing periods (see table 1 for details). We have therefore defined a common 30 year climatological period of 1980 to 2010 for all models to simplify the integrated mass budget comparison, except for COSMO-CLM^2 where the period covers 1987 to 2010. Figures that show time series of data show the full period relevant for each model.

2.1 Models

The model versions we include in this paper all fulfill the requirements of being the most up-to-date model version as well as being forced on the boundaries with ERA-Interim reanalysis. We also include the earlier RACMO v2.1 and MAR v3.6 as part of the initial SMB comparison as these models have been widely used and are still available for scientific use online; for example, results from RACMO2.1P were used in compiling the IPCC AR5 climate atlas. However, they are no longer considered up to date and have been replaced by RACMO2.3p2 and MAR$_{v3.10}$ respectively therefore we do not consider them in the detailed results analysis in this paper. The models also have snow schemes of differing complexity so the comparison of SMB necessarily includes slightly different terms for different models. For example, the RACMO model has been developed to include the wind blown snow sublimation terms in SMB and both RACMO and MAR$_{v3.10}$ include melt and refreezing of meltwater. As these terms cannot easily be removed without retuning the models, we have opted to include these within the SMB calculation for these two models. We also explicitly include a second simple SMB calculation equation 1, based only on the precipitation and sublimation for a fairer model intercomparison within the results section. The individual model descriptions give further details of each models outputs.

2.1.1 COSMO-CLM^2

COSMO-CLM^2 is a non-hydrostatic RCM developed at the German Weather service together with an extensive scientific community (Rockel et al., 2008). The model is applied over the Antarctic at a spatial resolution of ~25 km and 40 vertical levels in the atmosphere. The model is forced every 6 hours at the boundaries by ERA-Interim. Additionally, this model is

coupled to the Community Land Model (version 4.5; Oleson and Lawrence, 2013), with adjustments in the perennial snow proposed by van Kampenhout et al. (2017) to better represent the SMB of ice sheets (COSMO-CLM^2). Apart from this, several model parameters were adjusted for polar regions, particularly those related to the turbulent kinetic energy scheme and the cloud scheme. A full description of the setup over Antarctica including an evaluation of its performance in simulating the Antarctic climate and SMB is available in Souverijns et al. (2019). In this paper, precipitation minus sublimation is taken as a proxy for the SMB.

2.1.2 HIRHAM5

HIRHAM5 is an RCM developed at the Danish Meteorological Institute and run in this study at both low (0.44 0.44 km) and high (0.11 0.11 km) resolution, with all other model elements being kept identical. The model combines the atmospheric dynamics of the HIRLAM7 numerical weather prediction model (Eerola, 2006), and the physics of the ECHAM5 global climate model (Roeckner et al., 2003). There are 31 vertical levels in the atmosphere and the model is forced at 6 hourly intervals on the lateral boundaries with temperature, pressure, relative humidity and the wind vectors. Sea surface temperatures (SST) and sea ice concentration (SIC) are forced on the lower boundary at daily intervals. The set-up for Antarctica is similar to that of Lucas-Picher et al. (2012) in Greenland, that is with only a very simple surface physics scheme over glacier ice. A subsurface scheme developed for Greenland by Langen et al. (2017) is currently undergoing optimisation for Antarctic SMB processes but was not available for use in these simulations. We used the model outputs of precipitation, evaporation and sublimation to compute a simple SMB.

2.1.3 MetUM

The UK Met Of ce Uni ed Model (MetUM) is a numerical modelling system based on non-hydrostatic dynamics (Walters et al., 2017), which can be run either as a global model or a regional mesoscale model, as presented by e.g Orr et al. (2015). Here, we run version 11.1 of the mesoscale model over the standard Antarctic CORDEX domain at a spatial resolution of 50 km and 70 vertical levels (reaching up to 80 km). The mesoscale model is nested within a global version of the MetUM with a horizontal resolution of N320 (i.e. 640 × 480 longitude-latitude grid implying a nominal 40 km horizontal mesh), which was initialised by ERA-Interim. For this study we ran a series of consecutive twice-daily 24-hour forecasts at 00 and 12 UTC from the beginning of 1980 to the end of 2018. The rst 12 hours of each forecast were discarded as spin-up, with the remaining output concatenated together to form a continuous time-series. Although the mesoscale model includes a multi-layer snow scheme (Walters et al., 2019), in these simulations we used a simpli ed single-layer scheme with for example, no refreezing (Cox et al., 1999). We therefore calculate SMB based on output precipitation and sublimation and evaporation.

2.1.4 $MAR_{v3.10}$

The «Modèle Atmosphérique Régional» (MAR) (Gallée and Schayes, 1994) is a hydrostatic RCM specially designed for polar areas (e.g., Fettweis et al., 2017; Kittel et al., 2018; Agosta et al., 2019). The model has 24 atmospheric vertical levels and an horizontal resolution of 35 km. MAR is coupled to the 1-D multilayer surface scheme SISVAT (Soil Ice Snow Vegetation Atmosphere Transfer; De Ridder and Gallée, 1998), which simulates mass and energy fluxes between the atmosphere and the surface. The snow-ice module, based on the CROCUS model (Brun et al., 1992), represents the evolution of the snowpack for 30 snow layers through subroutines of snow metamorphism, surface albedo, meltwater runoff, percolation, retention and refreezing. MAR is forced with ERA-Interim every 6 hours over 1979 – 2018 at its atmospheric lateral and upper boundaries (pressure, wind, specific humidity and temperature at each vertical level) and over the ocean surface (SST and SIC). Furthermore, an upper-air relaxation is used to constrain the MAR general atmospheric circulation (van de Berg and Medley, 2016). Relative to previous studies over the AIS (Kittel et al., 2018; Agosta et al., 2019), the version used in this study ($MAR_{v3.10}$) only improves the cloud lifetime, the model stability and its computational efficiency enhancing a larger independence of MAR to its timesteps. Furthermore, the definition of the AIS mask has also been improved by taking into account rock outcrops. An extensive description of the adaptation of MAR to the AIS can be found in Agosta et al. (2019).

2.1.5 RACMO2.3p2

The Regional Atmospheric Climate Model RACMO2.3p2 combines the dynamical processes of the High Resolution Limited Area Model (HIRLAM) (Undén et al., 2002) and the physics package CY33r1 of the European Centre for Medium-range Weather Forecasts (ECMWF) Integrated Forecast System (IFS). RACMO2.3p1 was built by porting the polar physics components that were part of RACMO2.1P into the standard climate model RACMO2.3 developed at the Royal Netherlands Meteorology Institute (KNMI). RACMO2.3p2 is the follow-up of RACMO2.3p1 and has been applied to the polar ice sheets of Greenland and Antarctica by the Institute for Marine and Atmospheric research Utrecht (IMAU). RACMO2.3p2 includes a multilayer snow model that calculates melt, percolation, refreezing and runoff of liquid water (Ettema et al., 2010). RACMO2.3p2 also uses a prognostic scheme for snow grain size used to calculate surface albedo (Kuipers Munneke et al., 2011); and a drifting snow routine that simulates the interaction of drifting snow with the surface and the lower atmosphere (Lenaerts et al., 2012a). For this study, the model operates at a horizontal resolution of 27 km, with 40 vertical atmospheric levels. Surface topography is based on Cook et al. (2012) and Bamber and Gomez-Dans (2009). At the lateral and the upper atmospheric boundaries the model is forced by ERA-Interim reanalysis data every 6 hours, and at the ocean boundaries by prescribed ocean temperatures and sea ice cover. The model atmosphere is initialised Jan, 1st, 1979 with the ERA-Interim reanalysis data, and the snow/firn layer with data generated by the IMAU Firn Densification Model (IMAU-FDM) (Ligtenberg et al., 2011). The precursor version, RACMO2.3p1 includes an older ice mask and surface topography, no upper air nudging, a more severe drifting snow formulation eroding more snow and changes in the formulations of surface melting and precipitation. Further details can be found in van Wessem et al. (2018) that intercompares versions p1 and p2 more fully.

2.1.6 RACMO2.1P

RACMO2.1P is an earlier version of RACMO2 using the ECMWF-IFS physics package CY23r4 that does not include ice cloud super-saturation and utilizes earlier parameterizations for short wave radiation and boundary-layer turbulence as described in Van Wessem et al. (2014). This version of RACMO2.1 includes the polar multi-layer snow routines, as well as the schemes for drifting snow and albedo as described for RACMO2.3p2 above. In essence, its polar physics components are identical to those in RACMO2.3p1. Simulations with RACMO2.1P have been performed on a modelling domain matching the CORDEX ANT-44 domain in the interior plus a 16-point extension on each domain side for boundary relaxation of ERA-Interim elds. There is also no nudging within the domain in this version.

Model	Period	Resolution [km] (degrees)	Nudging	SMB scheme	Topography	Atmos. Levels
COSMO-CLM^2	1987-2016	25 (0.22)	Yes	Yes	GLOBE	40
HIRHAM5	1979-2017	50 (0.44); 12.5 (0.11)	No	No	GTOPO	31
MetUM	1979-2018	50 (0.44)	Reinit.	No	GLOBE	70
$MAR_{v3.6}$	1979-2018	35	Yes	Yes	Bedmap2	23
$MAR_{v3.10}$	1979-2018	35	Yes	Yes	Bedmap2	24
$RACMO2.1P_{v1}$	1979-2012	50 (0.44)	No	Yes	RAMPv2	40
RACMO2.3p2	1979-2018	27 (0.25)	Yes	Yes	Cook, Bamber	40

Table 1. Summary of differences and similarities between the RCMs. Horizontal resolution is given in degrees and (kilometres), while the number of atmospheric levels refers to the vertical resolution. Nudging refers to the level of forcing within the domain, refer to the individual model descriptions for more details.

2.2 Model Set-up and Outputs

2.2.1 Surface Mass Balance Calculations in RCMs

Two of the models (RACMO and MAR) have subsurface schemes optimised over snow and ice for Antarctica (see references under the model descriptions). The models include parameterisations to account for retention and refreezing of meltwater and also in the case of RACMO2.3p2 wind-driven processes such as erosion at the surface and sublimation of blowing snow. Thus, the de nition of the calculation of the SMB changes depending on the complexity of the model. Three models (HIRHAM5, METUM, COSMO-CLM^2,) have only simple surface snow physics over ice surfaces in these experiments. The basic SMB we calculate for them in this study is:

$$SMB = precipitation \quad evaporation \quad sublimation \tag{1}$$

For MAR with an optimised subsurface schemes the SMB is calculated from Eq.2:

$$SMB = precipitation \quad evaporation \quad sublimation \quad runoff \tag{2}$$

This differs slightly in RACMO2.3p2/RACMO2.1P as sublimation and erosion of drifting snow (SU_{ds} and ER_{ds} respectively) are also included as a mass loss term as in Equation 3:

$$SMB = precipitation \quad evaporation \quad sublimation \quad runoff \quad SU_{ds} \quad ER_{ds} \tag{3}$$

Both models account for refreezing and retention and thus use runoff rather than melt. Due to the low temperatures in Antarctica, most meltwater refreezes and runoff is negligible in the current climate (van Wessem et al., 2018; Agosta et al., 2019) so for the remaining models without the multi-layer subsurface schemes SMB is calculated without the runoff component.

2.2.2 Nudging and upper atmosphere relaxation

As von Storch et al. (2000) pointed out, nudging, whether spectral or with simpler techniques, keeps a regional model closer to the driving large-scale fields (GCM or reanalysis) and is thus a valuable technique where a close match to observations or to a driving GCM is required. Within Polar CORDEX, upper-air relaxation and other forms of nudging have been included as a standard where observational campaigns in large domains require close matches between modelled and observed weather. For example, Arctic cyclone systems and the presence of clouds in particular, appear to be better resolved in models that include nudging (Akperov et al., 2018) and (Sedlar et al., 2011). Similarly, nudging of RCMs run over Antarctica ties their synoptic evolution to these of the driving reanalysis, improving the representation of the interannual variability in SMB to similar levels as in the reanalysis as shown in van de Berg and Medley (2016).

In the experiments presented here, COSMO-CLM^2, MAR$_{v3.10}$, and RACMO2.3p2 are nudged by adjusting temperature and wind fields to the global fields with a minimum relaxation time scale of 6 hours. Strongest relaxation is applied at the top of the atmosphere and relaxation decreases gradually for lower levels. Below typically 4 km (ocean) to 6.2 km (4 km land topography) no relaxation is applied. In the case of MAR$_{v3.10}$, the relaxation of the temperature is weaker than the relaxation of the wind between the highest cloud level and the lowest nudging level. This prevents inconsistency between the temperature inherited from the reanalyses and the humidity and clouds conditioned by the MAR microphysics scheme. Moisture fields are not adjusted by nudging as this would introduce artificial uphill moisture transport. HIRHAM5 and MetUM are not nudged but MetUM is run in a 12-hourly reinitialisation hindcast that keeps the model evolution close to the driving reanalysis.

2.2.3 Grids and land-sea-ice masks

All models have been run for a domain covering the entire Antarctic continent but not all of the domains are the same. HIRHAM5 0.44 and MetUM use the standard CORDEX domain and grid. However, COSMO-CLM^2 extends this slightly to cover more ocean around Dronning Maud Land while the HIRHAM5 0.11 simulations and MAR$_{v3.10}$ were run over slightly smaller domains than the CORDEX domain to reduce computational time, though only after running experiments to determine that e.g. precipitation was not affected. RACMO2.3p2 and RACMO2.1 are run for a domain slightly larger than CORDEX but are trimmed back to remove the relaxation zone such that final results are presented on the CORDEX domain. As the model resolutions are different and each model had its own land-sea mask, the area of Antarctica is not the same in all models, which

complicates the SMB results when integrated over the continent. To correct for this areal difference, all the data have been bilinearly regridded to the HIRHAM5 0.11grid, with the unglaciated land of MAR3.10 included, with a threshold for the ice mask of 50%. This was used to generate a common ice mask for the models in order to calculate the integrated SMB over the ice sheet, ice shelves and in the individual basinc. In the appendix, gure A1 shows all masks compared to the common mask. Most models had very few grid points different from the common mask but these are also areas with high precipitation rates and this therefore would give measurable differences in annual SMB. We do not report these differences here but it is important to bear in mind the ice masks used when comparing our results with those from other studies.

Modelled SMB is integrated over drainage basins de ned as in Shepherd et al. (2018b). The horizontal resolution of the models is not altered and the drainage basin masks are de ned by selecting all model grid points that fall within the drainage basin outlines. In addition to the drainage basins, that are by de nition grounded ice, outlines of the ice shelves that the basins drain into are also used. This allows us to partition SMB over grounded ice (GRIS) and ice shelves (IS) as well as over the ice sheet as a whole including ice shelves (TotIS).

2.3 Observations

2.3.1 AWS observations

We use weather observations to assess how well RCMs reproduce the meteorological conditions over the AIS. Although a detailed evaluation of the near-surface model climates of each of the models is not the purpose of this study, this comparison helps to explain model biases in simulating SMB and especially the coherence between the modelled SMB and the near-surface climate. The original dataset is a compilation of surface pressure, near-surface temperature and wind speed from 30 AWS over the ice sheet used in the MET-READER database (Turner et al., 2004), but also collected by the BAS, IMAU (van Wessem et al., 2014), and the IGE/IPEV (Amory, 2019). The original data were available at several sampling time steps (sub-hourly, hourly, 3-hourly) and were averaged to obtain daily values. Only daily averages computed from more than 75% of the original data are considered as representative of the entire measurement (UTC) day and are used for comparison. Sever stations displayed suspicious measurements (sudden discontinuity in pressure and temperature, temperature values capped the lower bound of the measurement range during the whole winter season, etc) and these were removed from the dataset Stations occasionally exhibited wind speeds of 0 m/s for day-long periods, probably as a result of sensor riming. For these cases the daily averages were considered as no data (See Kittel et al. in preparation for details on the full list of AWS and data selection protocol). Although we use a homogenised and quality-controlled dataset for the comparison, observations may still be biased in ways that are hard to quantify due to e.g. burial of stations by snow, battery failures, tilt due to strong winds and other instrument failures that remained undetected, re ecting the dif culties involved in collecting data in the harsh and remote Antarctic environment.

As the different models have different ice masks and topographies, we only retain stations on the common mask where the difference in elevation is lower than 500 m for each model, this gives a total of 184 AWS (See gure A2 in the appendix for locations of AWS used in this study). We compute the modelled surface pressure, near-surface temperature and wind speed, a

well as the model elevation, using a four-nearest inverse-distance-weighted method. Finally, since the measurement height is not known for every station, we use the vertical level closest to the surface (10 m or 2 m) of the models for all comparisons with the observations.

2.3.2 Comparison with 10 m snow temperature observations

Deep snow temperatures in Antarctica are indicative of the annual long-term mean surface air temperature. Here, we use 64 observations of 10 m snow temperature, collected from a broad range of climatic regions of Antarctica, representing a spatially complete picture of climatological surface temperature (Van Wessem et al., 2014) to compare with model output.

2.3.3 Observed SMB

Observations of SMB are sparse over the wide Antarctic continent, and have been obtained from diverse measurement techniques such as stake measurements, ice cores and radar stratigraphy. For the purpose of our model evaluation, we use the SAMBA dataset from Favier et al. (2013), that has been updated with observations from Wang et al. (2016) and yearly values of shallow ice cores from Thomas et al. (2017) giving a total dataset of 7136 observations for various time periods and for a wide range of locations scattered across the AIS. We did not use the radar measurements published by Medley et al. (2014) in this study as the spatial variability is very high and dif cult to smooth appropriately for all model grids.

To evaluate the models, we selected observations of SMB on the common ice mask and for which the measurement period falls between 1950 and 2018. These conditions reduced the total number of observations used in the comparison to 3671. We used observations between 1950 and 1987, or 2015 and 2018 that are not fully included in the common modelling period of 1987 to 2015, for evaluation only if they covered more than 5 years. These 1849 SMB observations are compared to modelled values averaged over the common modelling period in order to compute a climatological mean while we averaged modelled SMB values over the exactly same period for the observations between 1987 to 2015 (1822 observations).

Since the models have different resolutions and grids, we do not directly compare the modelled SMB values to the observations. As in Kittel et al. (2018) and Agosta et al. (2019), we compute modelled and observed SMB values in 2 steps. Firstly, the original resolution modelled SMB values were interpolated, as for AWS observations, to the observation location using a four-nearest inverse-distance-weighted method. Secondly, all the interpolated SMB values contained in the same grid cell from the common ice mask were averaged as well as the observations for nally creating 923 comparison pairs. This leads to a fair comparison for each model that takes into account the bene t of using a higher resolution for a speci c model and removing the very high spatial variability of the observations that cannot be reproduced by the models.

Like the meteorological data, SMB observations are subject to measurement biases notably due to post-depositional redistribution of snow and the related formation of sastrugi that can considerably complicate the interpretation of measurements at the very local scale (Andersen et al., 2006). SMB observations should therefore be considered as a best estimate of accumulation rather than an absolute value. As SMB observations are not evenly distributed over the ice sheet, the comparison statistics are arti cially in uenced by over- and/or under-sampled regions.

3 Results

We first focus on how the RCMs characterise the surface climate over the ice sheet before turning to assessing the SMB and take note of the differences in precipitation distribution.

3.1 Temperature, Surface Pressure and Wind Speed from Models and Observations

Weather observations in Antarctica extend further back in time and there is generally better spatial and temporal coverage than for direct SMB measurements. In figure 1, we show Taylor diagrams for pressure, temperature and wind velocities. Taylor diagrams offer an efficient way to assess model skill by comparing the Pearson correlation coefficient, the centred root mean square error (CRMSE) and the standard deviation of the modelled output with the observed values. CRMSE is equivalent to the Root Mean Square Error but systematic biases are removed by subtracting the mean observation and mean modelled value from each value as shown in equation 4

$$CRMSE = \sqrt{\frac{\sum_{i=0}^{n}(m_i - o_i)^2}{n} - (\bar{m} - \bar{o})^2} \quad (4)$$

Where n is the number of observations, m_i is the modelled value, o_i is the observed value and $\bar{m}$ and $\bar{o}$ are the average of the modelled and observed values respectively.

A perfect model should be in the same place as the observations (shown by the black star in figure 1 with a correlation of 1, the same standard deviation, and zero CRMSE). The further away a model is from the observations, the more poorly it matches the observed weather. Mean biases and the observational mean are also indicated. In this case, modelled values closest to the dashed line have a more correct representation of the standard deviation and the closer to the black reference star the closer the model correlates to the observations values. We list the bias below the diagrams.

Figure 1. Taylor diagrams showing model performance compared to daily observations of surface pressure (a), near-surface temperature (b), and observed wind speeds (c).The horizontal and vertical axes represent the standard deviation, the dashed line in bold shows the standar deviation of the observations. The Taylor plot also shows the correlation which is measured by the angle with the x-axis. Finally, the CRMSE is represented by the curved lines in light grey. The units of standard deviation, CRMSE, mean bias and mean of the observations are the same (hPa for surface pressure, K for near-surface temperature, and m/s for wind speed).

Figure 1 analysis shows that, depending on the variable, all the models perform reasonably well though with some variation. With respect to surface pressure, the majority of models are similarly skillful with the exception of HIRHAM5 0.11 which has the lowest correlation and highest bias, although the model is still close to the pattern of the standard deviation. The other models have quite a high degree of nudging, including upper atmosphere pressure elds within the domain, so it is not so surprising to see the good performance here as the nudging forces the models to be closer to the observed pressure. Withou nudging, the large domain size in Antarctica means that synoptic scale systems have more degrees of freedom to evolve away from the observed quantities. This is likely to be a particular problem for higher resolution models where there are more grid points between the boundary and a given station, compared to a lower resolution model with fewer grid points. Our results show that the high resolution (0.11 version of HIRHAM5, that has many more grid cells than the low resolution (0.44 version, has a higher divergence due to internal variability. MetUM is not nudged by surface relaxation but is run in daily reinitialisation mode and while this probably also helps to keep surface pressure close to observed it is also likely that the large number of atmospheric levels in MetUM also improves modelled surface pressures. The near-surface temperatures in gure 1 show that,

although overall the models perform well (Pearson correlation of 0.85 and higher), on average all the models are too cold and only $MAR_{v3.10}$ and RACMO2.3p2 have a bias of less than 1 degree (respectively -0.16 and -0.51 K), with MetUM having the highest bias (-3.44 K). As with the surface pressure analysis, the HIRHAM5 high resolution simulations have a relatively lower correlation coef cient (0.85 compared to above 0.9 for the other simulations)and this may well be again the consequence of the un-nudged simulations. However, biases in cloud cover and long-wave radiation reaching the surface are likely the main explanation for divergence from observations and should be investigated for all RCMs run for Antarctica as shown by van Wessem et al. (2014). In their study, signi cant improvements in the RACMO2.3p2 model were obtained by adjustments to the cloud microphysics. Furthermore, the lack of detailed subsurface snow pack schemes including processes such as refreezin (and subsequent latent heat release) and densi cation also likely has an impact on the near surface and subsurface temperatu bias in HIRHAM5 and MetUM (see also gure 2).

Figure 1 shows that all of the models perform less well for wind speeds than for temperature or pressure observations. The wind speed plot shows all models have higher CRMSE, higher standard deviation and lower correlation values when compared with observations. Even so, the RACMO2.3p2, MetUM and $MAR_{v3.10}$ still show a correlation above 0.9 with observations, suggesting the nudging schemes in these models are effective in helping to reproduce observed wind speeds. There are als likely to be large uncertainties in the observations, especially at unattended stations where burial by snow, changes in orientation and sensor breakdown are more likely. However, the effects of different resolution and differences in turbulent schemes between the models may also be important. In particular, the extremely stable boundary layer over most of Antarctica is hard to represent in models particularly at lower resolutions (Zentek and Heinemann, 2019). The models appear to fall in two groups on the Taylor Diagram: $MAR_{v3.10}$, MetUM and RACMO2.3p2 on the one hand, and the two HIRHAM5 runs and COSMO-CLM^2 on the other hand. In the case of COSMO-CLM^2 wind speeds are output at 20 m and then interpolated to 10 m using Monin-Obukhov theory (Souverijns et al., 2019), which may not be suf cient to properly represent near-surface winds and associated interactions. The HIRHAM5 results may again be biased due to the lack of nudging within the domain. However, it is worth pointing out that HIRHAM5 correctly represents the mean spatial variability (both runs are the closest to the dashed line indicating the standard deviation) and, in the case of the high resolution run, has a very low bias in the the mean observed wind speed.

3.2 Comparison with 10 m snow temperature observations

Figure 2 shows the modelled surface temperature of the RCMs as a function of 64 measurements of temperature at 10 m depth as also used by Van Wessem et al. (2014). The majority of the AIS has negligible snow melt and in these regions the 10 m snow temperature is representative of the annual long-term average surface temperature. This comparison, therefore, i a robust assessment of the climatological surface signal calculated by the models, also because the observations are even scattered across the continent and represent most climatic regions. All models capture the wide range of surface temperature from 218 K to 260 K. HIRHAM5 0.44consistently underestimates temperature for most locations, a bias that closely resembles RACMO2.1 in Van Wessem et al. (2014) and which the authors concluded was predominantly related to biases in the downwelling long-wave radiation. The other models overestimate temperature in the higher elevation, colder locations,

while underestimating temperature at lower elevations in the coastal regions. For the colder regions below 240 K, these biases are most likely related to discrepancies in cloud cover likely snowfall, affecting downwelling longwave radiation and surface albedo. Some of the Antarctic models have been tuned to improve the dry and cold biases in the interior that were persistent in earlier model versions (see RACMO2.1P; Van Wessem et al. (2014); van Wessem et al. (2018)), but now overestimate temperature slightly instead. While subsequent model updates have led to signi cant improvements in simulated SMB, this has come at the expense of surface temperature due to excessive increases in downwelling radiative uxes that accompany increases in snowfall.

For the lower elevation, mostly coastal regions most models have a cold bias. This bias is likely related to the effects of surface meltwater percolating into the rn and refreezing within, raising deeper snow temperature, implying modelled surface temperature is not a good metric for observed 10 m snow temperatures in the percolation zone. A more accurate comparison would therefore be to directly compare 10 m snow temperatures from the models with the observations. However, not all models calculate snow temperatures, and given the scope of this manuscript, we only intercompare the surface temperature. Here, Figure 2 illustrates a consistent intermodel scatter, with mainly the models that do not include a sophisticated snow model outside of this range. This points to a signi cant potential source of improvements for modelled SMB in the future.

Figure 2. Modelled surface temperature as a function of observed 10 m-snow temperature (Van Wessem et al., 2014). Observations not fully located on the model ice-mask are excluded.

3.3 Comparison with Observed SMB

Evaluating SMB is hindered by poor observations across the cryosphere, particularly in Antarctica where remoteness and extreme weather conditions add to the challenge of observing SMB. Our analysis uses a large dataset of observations but there are large areas signi cantly undersampled (See for example, gure A2). We therefore separate the comparison of modelled and observed SMB into elevation bins in gure 3 in order to make the results clearer. Note that gure 3 is plotted on logarithmic axes because the distribution of both the observed and simulated SMB is not Gaussian. As linear regression is strongly in uenced by the extreme values, which skew errors in both modelled and observed SMB for the largest values but is only weakly in uenced by the errors on the smallest absolute values, a logarithmic plot better displays how well models reproduce SMB in both high and low SMB regions. It is also important to note that for the scatter plots by elevation class, if an observation or one of the models had a negative value, the observation and modelled values were removed from the analysis using logarithmic values for the scatter plots by elevation class, if an observation or one of the models had a negative value, the observation and

modelled values were removed from the analysis using logarithmic values (hereafter, rlog is the correlation computed on the logarithm of SMB values), but are retained in the analysis using the original populations. We show detailed statistics for the SMB comparison in Table 2. In order to show the large scatter in the observations and the models clearly, we also plot all modelled SMB values against observed SMB values in Figure 4. We show individual model comparisons in the appendix to save space here (Figures A3 , A4, A5 A6, A7,A8).

Apart from COSMO-CLM^2 and HIRHAM5 0.11, the RCMs show similar Root Mean Square Error (RMSE) and r^2 values when compared over the full dataset but broken down by elevation class or locally by regions as in gure 4 shows a more complex story. In general, all models, with the possible exception of MAR$_{v3.10}$, underestimate SMB at the ice shelf observation locations as well as in the low elevation coastal regions of Antarctica (see also statistics in Table 2 a. and b. and Figure 3). The highest mean bias, lowest RMSE and lowest r values in particular are given in the COSMO-CLM^2 and HIRHAM5 0.11 models at the lowest elevations. However, while all the other RCMs underestimate SMB over especially the Ross ice shelf, MAR$_{v3.10}$ overestimates it, probably related to a poorer representation of the surface climate by the model over this ice shelf. There are indications in gure 4 that both HIRHAM simulations overestimates SMB on the Ronne Ice Shelf, but we lack observations to be able to test this properly.

The blowing snow module included in RACMO2.3p2 may explain the lower bias and RMSE in this model at elevations between sea level and 1200 m and especially 1200 m and 2200 m (we show all statistics in details in Table 2 b and c), compared to the other models. A previous comparison shows higher sublimation in RACMO2.3p2 than in MAR$_{v3.10}$ (Agosta et al., 2019) notably at the elevations where katabatic winds are strong due to the slope of the ice sheet and where the atmosphere is not too cold enabling large amounts of sublimation from blowing snow particles. COSMO-CLM^2 and HIRHAM5 0.44 have the highest RMSE while HIRHAM5 0.11, MAR$_{v3.10}$ and MetUM have similar statistics at this elevation. For the highest elevations (above 2200 m), all the model RMSE scores are relatively low and similar to each other except HIRHAM5 0.44 (and to a lesser extent MAR$_{v3.10}$) between 2800 m and 3400 m (Table 2 e). However, the less extensively optimised models (HIRHAM5 at both resolutions and MetUM) are both too dry over the high plateau of the AIS.

If we look at all the elevation ranges, no model is systematically in the top 3 for every range but, RACMO2.3p2 has the best comparison with all the observations, closely followed by MetUM with MAR$_{v3.10}$ and HIRHAM5 0.44 performing almost equally. It is worth emphasising though that as Fig. 4 shows, the observations in this elevation class are also very noisy and the poor relative performance of the models may result as much from unrepresentative and sparse repeat observations as it does from missing or poorly resolved processes in models. Analysis of these results not only indicates areas where models need to be improved but also areas where more observations to test models are desirable, notably between 1200 and 2200 where the mean biases of the models used in this study display large discrepancies (Table 2 c). It is also likely that there are compensating errors within each model that hide the true performance. For example, the mean bias between the two different HIRHAM runs has opposite signs in the 1200 - 2800 m range, likely re ecting the difference in model resolution. Orographic precipitation is very sensitive to slope effects and the presence of steep topography is very different between the two resolutions, affecting where precipitation falls across the continent. The wide scatter in modelled SMB in the 2200-2800 m elevation range is therefore also likely to re ect in part the resolution of the different models and how well they capture orography and the consequent

precipitation. Studies by for example, Hermann et al. (2018) and Schmidt et al. (2017) show that hydrostatic models like HIRHAM5 and RACMO2.3 typically overestimate precipitation on the upslope and have a dry bias downwind of initial steep topography, this pattern seems to some extent to be repeated in ANtarctica in gure 3 and 4. Comparing the observations used in this analysis with the RCM ensemble modelled SMB in gure 6 also highlights that the largest differences between models and compared with the ensemble mean are mostly in regions with very few or no observations. These are also regions where precipitation is typically high, making it dif cult to assess the ability of models to truly simulate the SMB of Antarctica. Our analysis therefore also helps to identify areas where increased observations will be most useful to help assess and improve model processes.

Figure 3. Comparison between modelled SMB and observed SMB in a gridded dataset. Trend lines and points are plotted for each model in a different colour, note different x and y axes for different elevation bins. The gures are plotted on logarithmic axes because the datasets are not gaussian distributed and this better represents the relative error in both high and low SMB regions than linear axes.

Mean bias and RMSE for each model by elevation bin is summed up in the supplementary materials in Figure A9. However as Figure 4 also shows, this is not a straightforward comparison either due to the large areas with only few observations.