Evaluation of Greenland near surface air temperature datasets

J. E. Jack Reeves Eyre[1] and Xubin Zeng[1]

[1]Department of Hydrology and Atmospheric Sciences, University of Arizona, Tucson, 85721, USA

Correspondence to: J. E. Jack Reeves Eyre (jeyre@email.arizona.edu)

Abstract. Near-surface air temperature (SAT) over Greenland has important effects on mass balance of the ice sheet, but it is unclear which SAT datasets are reliable in the region. Here extensive in-situ SAT measurements (~ 1400 station-years) are used to assess monthly mean SAT from seven global reanalysis datasets, five gridded SAT analyses, one satellite retrieval and three dynamically downscaled reanalyses. Strengths and weaknesses of these products are identified, and their biases are found to vary by season and glaciological regime. MERRA2 reanalysis overall performs best with mean absolute error less than 2 °C in all months. Ice sheet-average annual mean SAT from different datasets are highly correlated in recent decades, but their 1901–2000 trends differ even in sign. Compared with the MERRA2 climatology combined with gridded SAT analysis anomalies, thirty-one earth system model historical runs from the CMIP5 archive reach ~5 °C for the 1901–2000 average bias and have opposite trends for a number of sub-periods.

1 Introduction

Near-surface air temperature (SAT) over the Greenland ice sheet (GrIS) is important both for its place in wider climate change and for its effects on mass balance of the ice sheet. Due to its remoteness and extreme climate however, continuous widespread climate monitoring over the GrIS has been carried out for only about the last two decades, and even then with rather sparse coverage in some geographic areas and glaciological regimes. Studies of past climate and surface mass balance (SMB) of the GrIS have used a variety of techniques to achieve complete spatial coverage of SAT, including statistical interpolation, atmospheric reanalysis, dynamic downscaling through regional climate modeling, and satellite remote sensing. Projections of future change in Greenland climate and ice sheet evolution have used global earth system models, either directly (e.g., Ridley et al., 2005; Vizcaíno et al., 2013) or through dynamical downscaling (e.g., Fettweis et al., 2013; Rae et al., 2012). Many such studies have involved some form of assessment using weather station data (e.g., Box, 2013; Noël et al., 2015; Rae et al., 2012) and inter-comparison of several SAT data sources (e.g., Box, 2013). Here we build on such work to assess and compare a greater number of widely available products, using a more comprehensive set of in situ observations than has customarily been used in previous work. In doing so we hope to guide future dataset and model development over this region and address a number of outstanding questions.

Our main focus here is on global datasets – reanalyses, gridded SAT analyses and earth system models from the CMIP5 archive – though several regional datasets are also included. Regional climate models (RCMs) have been used widely to downscale reanalysis (e.g., Box, 2013; Box et al., 2009; Burgess et al., 2010; Ettema et al., 2010a; Fettweis et al., 2017; Noël et al., 2015) and global climate model output (e.g, Fettweis et al., 2013b; Rae et al., 2012). While Noël et al. (2016) demonstrated the benefit of high (< 10 km) resolution downscaling for SMB, the benefit for SAT is less clear: because SAT is strongly elevation-dependent, use of a high resolution model may not lead to a significant improvement compared to a lower resolution model with elevation corrections, as shown by Lucas-Picher et al. (2012) for grid sizes 0.25^{o} and 0.05^{o}. By comparing results from a range of resolutions, including RCMs at relatively high resolutions, we aim to investigate the value added by dynamic downscaling.

Inter-comparison of SMB components has been carried out among different RCMs and between RCMs and global reanalyses (Cullather et al., 2016; Rae et al., 2012; Vernon et al., 2013). The results from these studies point to a wide inter-model spread, which are related to differences in model parameterizations (e.g., snow and ice physics), model ice mask and forcing at the domain lateral boundaries. One goal of this work is to investigate how closely RCM forcing affects SAT representation, by comparing differently forced runs of the same RCM (building on the work of Fettweis et al., 2017), and comparing these runs with results taken directly from the forcing dataset.

Satellite remote sensing data has been key in spatially complete reconstruction of GrIS SAT, whether through direct use (e.g., Hall et al., 2013) or through assimilation into reanalyses. One consequence of this, though, is that only a small proportion of studies extend GrIS SAT back before the satellite era. SMB studies that incorporate centennial scale SAT reconstructions include: Hanna et al. (2011), who combined Twentieth Century Reanalysis (Compo et al., 2011) and ERA–40 reanalysis (Uppala et al., 2005); and Box (2013) who adjusted regional climate model output using in situ observations to reconstruct SAT from 1840–2010. The Box (2013) SAT reconstruction was compared to that of Hanna et al. (2011) and found to be cooler over most of the common period, but especially so before about 1930. More recently, Fettweis et al., 2017 investigated the effect on RCM-derived SMB of using different forcing reanalyses and showed that SAT estimates are sensitive to model forcing, with large differences in the first half of the 20th century. By looking at multiple datasets that include the first half of the 20th century (and earlier), we hope to shed light on the climate of the GrIS in this very poorly observed period. In particular, such datasets allow comparison with previous assessments of Greenland SAT climate based on (mainly coastal) station data (e.g., Box, 2002; Chylek et al., 2006; Hanna et al., 2012; Mernild et al., 2014). Long, spatially complete time series also offer the best means of assessing CMIP5 models, without differences introduced by incomplete spatial coverage and short period (~ 30 year) trends and decadal variability.

This paper is structured as follows: in Sect. 2 data sources are described and examples of their past use given; results are broken down into Sect. 3.1, dataset assessment using in situ observations, Sect. 3.2, comparison of long term SAT changes among datasets and Sect. 3.3, further discussion; conclusions are presented in Sect. 4.

2 Data

2.1 Weather station observations

To assess the different SAT products, we use SAT observations made at manned and automatic weather stations (AWSs) from several sources, totalling 17000 station-months or 1400 station-years. These are briefly described here, and further details are shown in Fig. 1. Coastal station records of monthly mean temperature for 11 stations (stretching as far back as 1784) are compiled by the Danish Meteorological Institute (DMI; Cappelen, 2014). Thanks to their long records, SAT from these stations has been studied extensively: Box (2002) found a pattern of warming from ~1900 to ~1940, cooling from ~1940 to ~1990, and warming from ~1990 onwards. In addition, inter-annual variability was found to be closely related to the North Atlantic Oscillation (NAO). Hanna et al. (2012) found similar patterns of warming and cooling using updated SAT data from DMI stations, and concluded that recent temperatures were in excess of SAT from the early 20th century warm period.

In contrast to coastal regions, no long term (e.g., 30 years or more) climate monitoring has occurred on the GrIS. Monthly mean temperatures from mid-20th century expeditions and field camps, concentrated in the 1930s and 1950s, are taken from the appendix of Ohmura (1987). Since the mid-1990s, the number of SAT observations from the ice sheet has greatly increased. We use records from AWSs operated as part of the Greenland Climate Network (GC–Net), predominantly in the accumulation region of the ice sheet (Steffen and Box, 2001), from the K–transect in western Greenland (operated by the Institute for Marine and Atmospheric Research at the University of Utrecht; van de Wal et al., 2005) and from AWSs mostly in the ablation region operated by the Geological Survey of Denmark and Greenland (GEUS) under the PROMICE and GAP programs (Van As et al., 2011). Locations and types of all stations are shown in Fig. 1 and further details are available in the Supplemental Material (Table S1).

The providers of several of these observational datasets employ quality control tests and/or quality inspection as part of their routine data management. In addition, we remove unrealistic values where our inspection of time series reveals them (e.g., with spikes and step changes). Where data were provided as hourly values, we calculate daily averages (the mean of hourly values) for all days with 20 or more hourly values and monthly averages (the mean of daily values) for all months with 24 or more daily values.

2.2 Gridded SAT products

Most of the datasets assessed here fall into two categories: global reanalysis and interpolated global SAT analyses. The spatial and temporal resolution and length of record (Table 1) vary greatly across these products. It should be noted that even though reanalyses are constrained by (in some cases) remote sensing and some local observations to represent observed synoptic–planetary scale weather, the lack of assimilated SAT observations over Greenland means that the SAT data assessed here are largely the result of modelled atmospheric and surface processes.

Several of the latest generation of global reanalyses are used in this study (Table 1). Most of these are reliant on radio-sonde and satellite data, and thus cover only the period when these are available (1979 onwards; 1958 in one case). In addition, we analyze the Twentieth Century reanalysis version 2c (20CRv2c; Compo et al., 2011) and ERA–20C (Poli et al., 2016), which do not assimilate satellite or radio-sonde data, but instead use a subset of observation types that are available over the 20th century (and earlier) and therefore cover much longer periods. GrIS SAT from reanalyses has been used in SMB modeling: Hanna et al. (2005) used ERA-40, while Hanna et al. (2011) combined ERA–40 with 20CR. However, SAT data from a number of other reanalyses remain untested for such applications. It should be noted that, with the exception of ERA-Interim, SAT from land stations is not assimilated into reanalyses and so the SAT observations described in Sect. 2.1 are indeed an independent verification. In ERA-Interim, SAT is assimilated from land stations by the surface analysis scheme, to update surface fields (such as soil moisture) which have an effect on SAT. To the best of our knowledge, for the period analysed here the only Greenland SAT observations that are assimilated by ERA-Interim are from DMI stations, and so the ice sheet stations still provide independent data.

Reanalysis represents a combination of observations and model. In contrast, several research groups have created gridded SAT datasets based almost entirely on statistical analyses of weather station SAT (we refer to these as *gridded SAT analyses*). Such datasets have not been widely used over Greenland (though see, e.g., Fettweis et al., 2008), and their long time series and temporal homogeneity is a potential strength. For example, some reanalyses are known to suffer from spurious trends as observing networks and processing systems change (e.g., Screen and Simmonds, 2010): comparison between reanalyses and gridded SAT analyses, particularly in the early 20th century, can highlight such problems with reanalyses. Some gridded SAT analyses, due to their analysis methods and requirements for data completeness, have large data gaps over Greenland, e.g., HadCRUT4 (Morice et al., 2012) and NOAAGlobalTemp (Smith et al., 2008; Vose et al., 2012). However, here we use four such datasets that have complete (or very nearly so) coverage over Greenland. Three of these (NASA GISTEMP, CRU TS 3.23 and Berkeley Earth; references in Table 1) are widely used global SAT monitoring products, while one (NANSENSAT) covers only the Arctic. Note that GISTEMP (Hansen et al., 2010) is provided as anomalies only (relative to 1951–80 climatology). As the ice sheet weather stations have typically not been operational long enough to calculate a stable climatology, we do not assess GISTEMP using in situ observations; however, we do combine GISTEMP anomalies with MERRA2 climatology to enable assessment of *stationarity* of biases and comparison of long term variability against other datasets.

Recognizing that reanalysis SAT over Greenland is dominated by the model formulation and has relatively coarse horizontal resolutions, a number of researchers have sought to improve results over the GrIS by using reanalysis to force higher resolution regional climate models (RCMs) coupled to comparatively sophisticated snow–ice models. Such models are typically run with grid spacing of 10–20 km. This high resolution (compared to global climate models and most reanalyses) is thought to better resolve the large climate gradients that occur around the margins of the ice sheet. Here we include output from version 3.5.2 of the Modèle Atmosphérique Régional (MAR; Fettweis et al., 2013, 2017) run with 20 km grid spacing, then interpolated to the 5 km polar stereographic grid of Bamber et al. (2001). Three different runs of MAR are used here: one forced by ERA–40 (1958–1978) and ERA–Interim (1979–2015) reanalyses; a second forced by 20CRv2c reanalysis; and a third forced by ERA–20C reanalysis. ERA–40 and ERA–Interim reanalyses have been widely used as forcing data (Box et al., 2009; Ettema et al., 2010a, 2010b; Fettweis et al., 2013); 20CRv2c and ERA–20C have seen more limited use (e.g., Fettweis et al., 2017). It should be noted that the field we use from this model is nominally the 3-meter air temperature, whereas most reanalyses output 2-meter air temperature (when specified), and the measurement height at weather stations varies as the snow/ice surface changes. We also include an updated version of the SAT reconstruction of Box (2013) which uses statistical relationships between long-running DMI stations and RACMO2 RCM output (e.g., Noël et al., 2015) to estimate Greenland SAT on a 5 km grid from 1840 to 2014. This dataset can therefore be thought of as a hybrid of an RCM and gridded SAT analysis. We use Box2013 to denote this dataset.

Satellite remote sensing data, in addition to being assimilated by reanalyses, have been used directly to study the GrIS. Several studies have focused on the relationship between SAT and ice sheet surface temperature (IST), and have used data from both microwave (e.g., Shuman et al., 1995, 2001) and infrared sensors (e.g., Comiso et al., 2003; Hall et al., 2008, 2013; Koenig and Hall, 2010). Sounding instruments offer a method to retrieve air temperature more directly, but have received little attention over GrIS. Here we assess SAT from the Atmospheric Infrared Sounder (AIRS; Chahine et al., 2006) on board NASA's AQUA satellite platform. AIRS has been operational since September 2002, providing temperature and humidity retrievals at many vertical levels through the atmosphere. We use the level 3 monthly near surface air temperature from ascending and descending overpasses, taking a weighted average to give a single monthly value at each grid point (further details are given in Table 1). This product is a clear-sky only retrieval: a key part of assessing this product is to understand what effect this has through, for example, seasonally varying cloud amounts and increased wind-driven mixing during winter storms, as discussed in Koenig and Hall (2010).

Earth System Models (ESMs) from the CMIP5 multi-model ensemble archive (Taylor et al., 2011) are included in comparisons of long term areal average SAT. However, comparison of CMIP5 ESMs against in situ observations is not performed because the ESMs are free-running coupled (atmosphere–ocean–land–ice) models, so we do not expect them to have the correct phasing of synoptic weather or inter-annual or even decadal climate. Apparent biases at station locations would therefore combine bias in the long term average and differences in variability over the relatively short station records. The ice sheet areal averages, compared to the longer reanalyses and gridded SAT analyses, should adequately reveal the first order biases in the ESMs' long term average SAT and its trends. Thirty-one different model configurations from 11 modeling centers are used. We use the first ensemble member (r1i1p1) of historical runs from all model configurations that had the necessary data (SAT and glacial ice fraction). Further details of individual models are given in Table S2. In contrast to other datasets above, CMIP5 ESM SAT data are used on their model native grids, rather than interpolated to a common grid (to be discussed below).

3 Results

Our analysis is based on the monthly mean near-surface air temperature. Except for CMIP5 ESMs and the MAR RCM variants, datasets were spatially interpolated from their native grid to a 5 km equal area grid (the EASE grid of the National Snow and Ice Data Center [NSIDC]) using bilinear interpolation. This resolution is used to attempt to resolve the large SAT gradients that occur over the steep topography at the margin of the ice sheet. Interpolating like this presents some potential problems due to model topography: The surface elevation fields used in many of the datasets here are smoother than the actual topography of Greenland, and this leads to elevation biases as seen in Fig. 2. The relatively low resolution 20CRv2c (Fig. 2b) has mostly positive elevation bias around the edge of the ice sheet and negative bias in the interior; however there are also regions of positive bias close to the center of Greenland. The higher resolution MAR (Fig. 2c) does not have the same magnitude of biases in the interior, but still misses much of the small scale detail, as seen by the speckled pattern of biases of alternating sign. All datasets have a negative mean elevation bias on the ice sheet (Table 2), with MAR the smallest and 20CRv2c the largest. Note that elevation errors are not a monotonic function of resolution: despite a smaller grid spacing than MERRA2 and ERA–Interim, CFSR still has a larger bias and mean absolute error.

The elevation biases cause the SAT fields to be smoother than in reality, and interpolation of the smooth SAT fields is unlikely to accurately reflect the true SAT gradients, which are strongly influenced by elevation. To account for this, a correction is applied to the reanalysis and AIRS datasets after interpolation to the EASE grid: for each product, the elevation field is also bilinearly interpolated to the EASE grid, and then compared to the digital elevation model (DEM) of Bamber et al. (2013; provided at 1 km grid spacing, and here bilinearly interpolated to the EASE grid). The elevation bias (product minus DEM) is multiplied by the relevant month's lapse rate from Fausto et al. (2009) and their product added to the interpolated SAT field. The importance of this step can be seen by comparing the results below with comparable figures for un-corrected datasets (Figs. S2 and S3). For some datasets in some seasons, the correction leads to a deterioration, but in most cases there is a clear improvement: in many cases, bias and MAE (averaged over all months) are reduced by 50% or more.

3.1 Monthly mean SAT biases

Comparisons between gridded datasets and in situ observations are made by choosing the nearest EASE grid point (for CRU and Berkeley Earth, which are land-only datasets, the nearest grid point may contain missing data, in which case the nearest non-missing grid point is chosen). Note that an alternative, using bilinear interpolation directly from the native grids to the station locations, gives very similar results. The primary statistics used in the assessment of datasets are mean bias and mean absolute error (MAE). When aggregating results over multiple stations, the average of station-months is taken, rather than averaging over time then over stations. Stations are grouped into coastal (DMI), ice sheet below 1500 m and ice sheet above 1500 m. The elevation of 1500 m is chosen to approximately represent the equilibrium line altitude, as found for the K–transect by van de Wal et al. (2005). The pattern of biases seen below is largely the same for different separation elevations between 1000 m and 2000 m. Aggregating over elevation bands like this can pick out some important aspects of spatial variation in dataset errors, but is likely to miss regional and local patterns of dataset error (see Fig. S1). Note that, when taking the spatial average across the ice sheet, the area above 1500 m dominates: using the DEM and mask of Bamber et al. (2013), Greenland has a total area of 2.16 million km^2, which is 16.5 % ice-free land, 18.6 % ice sheet below 1500 m, and 64.9 % ice sheet above 1500 m.

The seasonal cycle of bias and MAE averaged over all station months from 1979 onwards in Figs. 3 and 4 suggests that many datasets, though not all, show similar seasonal cycles: above 1500 m and at coastal stations, more positive biases in winter and more negative in summer; at ice sheet stations below 1500 m, the opposite cycle. Despite qualitative similarities, a clear picture of dataset performance emerges. On the ice sheet, MERRA2, MAR (all three versions), Box2013 and 20CRv2c are best. The AIRS satellite product is also very good, except in winter months at stations below 1500 m. NANSENSAT is one of the best performers below 1500 m during the summer; however it has large biases and MAE elsewhere and there are concerns with its long term homogeneity (see below). At coastal stations, ERA–Interim performs best (likely related to its assimilation of some of these observations), and JRA–55 and MERRA2 are nearly as good. MAR (all three versions) performs better in summer than in winter. This is thought to be due to the specification of sea ice thickness in the MAR v3.5.2 model: in many regions around the coast of Greenland, sea ice thickness is over-estimated in the model boundary conditions, resulting in a cold bias in adjacent areas (X. Fettweis, personal communication). Note that without elevation corrections, MAR coastal station errors are larger in summer but smaller in winter (Figs. S2 and S3). CRU and Berkeley Earth results are comparable to the best reanalyses at coastal stations, likely because it is SAT observations from the coastal stations that form the majority of the input data for these datasets. Based on a (rather subjective) assessment of the 12-month average bias (absolute value, to avoid cancellation between months) and MAE, the most consistent good performer is MERRA2: the 12-month average biases (absolute values) are approximately equal or less than 1.0 ^{o}C, and 12-month average MAE are less than 1.5 ^{o}C, in all regions. MAR (all three versions) and Box 2013 have comparable performance across the ice sheet, and in some cases better performance in lower elevations during summer (the most important region and season for SMB modeling), but overall are marred by large winter time biases at coastal stations.

The analysis above aggregates all station months from 1979 onwards. To investigate time variations in biases, Fig. 5 compares mean bias before and after 1979 for those datasets which begin before 1979. Note that the datasets beginning in 1979 show only small changes in bias by decade (not shown). GISTEMP is included here with the MERRA2 elevation-corrected climatology: the absolute values of the biases are highly dependent on the climatology, but here can be ignored as, for the purpose of assessing the stationarity in GISTEMP bias (and thereby the credibility of its long term variability and trends), we are interested in the *changes* in bias.

Clear differences are apparent for some seasons and datasets. Statistical significance of these differences (using Student's *t*-test for a difference in means with unequal variances, and defining significance at the 1% level) suggest that a number of the datasets have time-varying biases and so may show spurious long term trends. This is most apparent for the coastal DMI stations, where larger sample sizes give the statistical test greater power. 20CRv2c has negative changes high on the ice sheet and at coastal stations but positive changes in the ablation region. NANSENSAT shows negative changes over time in all regions in both winter and summer. Other than NANSENSAT, the gridded SAT analyses do not seem more prone to time-varying bias than reanalyses.

3.2 Time series

Areal average (weighted by glacial ice fraction) annual mean temperatures for all datasets show close correlation in recent decades: considering only the period 1979 onwards, the correlation (*r)* values are in the range 0.71 to 0.99. In earlier periods, correlations are generally smaller: for the period 1900–1940, we have r = 0.28 to 0.98, and for 1940–1980, r = 0.29 to 0.97. However, CRU, Berkeley Earth and GISTEMP have pairwise correlation coefficients of 0.90 or greater for all these periods since they are based on a similar set of surface stations, and their correlations with Box2013 are greater than 0.84 in all periods. ERA–20C is highly correlated with MAR–ERA–20C (r > 0.97) in all three periods, as the latter is forced by the former. Similarly, 20CRv2c is highly correlated (r > 0.90) with MAR–20CRv2c in all these periods. Interestingly, Box2013 is generally more highly correlated with CRU, Berkeley Earth and GISTEMP than with the MAR datasets.

Among the datasets covering the entire 20th century, most have similar inter-decadal variations, with a general pattern of early 20th century warming, up to 1930, followed by cooling to around 1990, then strong warming in recent years (Fig. 6). Nonetheless, differences do exist (Table 4). For instance, NANSENSAT shows relatively large early 20th century jumps thought to be caused by changing data sources over this period, indicating this dataset is not suitable for long term monitoring over Greenland. In 20CRv2c, the ~1930 peak is warmer than the most recent years, in contrast to other datasets. Related to the last point, the amount of cooling from 1930 to 1990 varies between datasets, with 20CRv2c showing strongest cooling, and GISTEMP showing least. Anomalies (relative to the 1981–2010 mean; Fig. 6b) reveal some more subtle differences. For example, MAR–ERA and CRU both show less positive anomalies than other datasets since about 2005. Variability in 20CRv2c matches other datasets closely from 1980 onwards, but before this differs significantly (except in the 1940s). Comparison with anomalies at long-running DMI stations (Fig. S4) suggests it is 20CRv2c that is in error here, as might be expected from the consensus of other datasets. MAR–20CRv2c shows agreement for more of the period, but seems to inherit poor representation of variability before about 1920 and from 1950 to 1980.

Of the datasets that extend back before 1900, Box 2013, Berkeley Earth and GISTEMP agree quite closely but show notable differences with 20CRv2c. Box2013, Berkeley Earth and GISTEMP cannot be considered truly independent data sources (as they all rely on similar input data for this period, as suggested by their close correspondence with observations in Fig. S4), and so their consensus is not especially meaningful. However, the fact that their biases are more constant in time (Fig. 5) than those of 20CRv2c suggest that they are more reliable for this period. In common with disparities mentioned above for the first half of the 20th century, users of these SAT datasets should be aware that significant uncertainties exist before 1900, with notable differences in trends and variability (both inter-annual and inter-decadal). We recommend the use of gridded SAT analyses alongside reanalyses and downscaled reanalyses, to assess sensitivity to these differences.

The range of SAT among CMIP5 ESMs is wider than that among the other datasets (Fig. 6), but much of this range comes from a group of four relatively warm models and two relatively cold models: eliminating these gives a range comparable to the gridded analysis and reanalysis datasets. This highlights the fact that choice of verification dataset can have a significant effect on assessments of ESM mean climate. Based on results above, we use GISTEMP with MERRA2 climatology to assess the long term mean temperatures of the CMIP5 ESMs. Using the 1901–2000 mean of ice sheet annual average temperatures, 10 ESMs lie within 1 ^{o}C (namely GFDL's CM3 and ESM2G; GISS's E2–H–CC, E2–R and E2–R–CC; IPSL's CM5A–MR, CM5A–LR_historical and CM5A–LR_esmHistorical; CESM1–CAM5; CMCC–CMS; see Table S1 for further details).

The median of the CMIP5 ESM trends (Table 4) is positive for all periods considered – in marked contrast to the other datasets. However, further investigation shows the picture is not so clear: The number of individual ESMs that have positive trends in each period suggest that, with the possible exception of 1990–2005, the models do not give a clear consensus on signs of trends: this may be because inter-decadal climate variability dominates, and the phasing of this variability differs between models. For the 1990–2005 period, 27 out of 31 ESMs have a positive trend and the median is an order of magnitude larger than for the earlier periods (although still smaller than the 1990–2005 trends from the other datasets). Thus the ESMs seem to agree on accelerated warming since 1990. Significance of the trends is tested using the method described in Santer et al. (2000), which is based on a two-tailed Student's *t*-test modified to account for autocorrelation in the time series. Few of the trends are significant, in any of the periods considered. The ensemble mean has a long term average slightly higher than that from MERRA2+GISTEMP, and trends broadly similar to the median of the individual trends (i.e., with accelerated warming since about 1970); however, it does not feature decadal variability that individual CMIP5 ESMs, reanalyses and gridded SAT analyses show, and thus has limitations in representing historical GrIS SAT.

Due to its importance in SMB calculations, we briefly consider summer mean (June to August) ice sheet average SAT (Fig. 6c). Many features are shared with the annual time series, e.g., periods of warming in the years leading up to 1930 and beginning in the 1990s. In addition, we see that the variability in MAR–20CRv2c and MAR–ERA–20C closely follow that in 20CRv2c and ERA–20C respectively. In contrast with the annual mean time series though, the CMIP5 ensemble mean more closely follows the evolution in the observation–based datasets.

3.3 Further discussion

The majority of in situ SAT observations from the ice sheet have been made since 1995. We have used the relatively small number of observations from the mid-20th century to assess the stationarity of biases, and find that several datasets show significant temporal variations in their bias. At ice sheet stations above 1500m (the region which dominates in areal averages), 20CRv2c shows large (and significant) changes – becoming more negative with time. 20CRv2c biases also become more negative with time at coastal stations (from which there are many more observations), casting further doubt on the suitability of this dataset for long term trend analysis. ERA–20C has more stable biases in the accumulation region and at coastal stations (though not in the ablation region), as do several of the gridded SAT analyses, suggesting that SAT reconstruction based on anomalies is valid over monthly to centennial time scales. This is not a trivial result, as it is not obvious a priori that conditions driving anomalies at coastal stations will result in a similar, smoothly varying response over different surface types.

Trends among the datasets assessed here (excluding CMIP5 ESMs) generally agree with patterns found in previous studies (e.g., Box, 2002). In addition, interannual variability since 1979 matches closely between datasets. However, differences between longer term trends, along with temporal changes in bias (discussed above), suggest that some datasets have limitations in their representation of early to mid-20th century GrIS SAT. In particular, 20CRv2c shows stronger cooling between 1930 and 1990 than most other datasets, and has a 1930s warm period warmer than the 21st century warm period. Such discrepancies between 20CRv2c and anomaly based SAT datasets have been noted at the global scale by Compo et al. (2013), although the differences here are much greater than those for global SAT. Similarity of anomalies among gridded SAT analyses and ERA–20C, along with the greater temporal constancy of their biases, leads us to put greater faith in their representation of long term trends and inter-decadal variability.

While, as noted above, interannual variability in the last 30 years matches closely between datasets, there is variation in the magnitude of ice sheet average trends (Table 4) and spatial variation in trends (Fig. S5) over this period. Box2013 has the largest recent trends, with largest trends in the west. MERRA2 has its largest trends in the south-west, whilst all three MAR versions have their largest trends in the north-east.

One of our central questions in this study is whether global SAT datasets are as good as RCM-downscaled datasets, which are, at least for SMB modeling, the current state of the art. For MAR–ERA and MAR–ERA–20C, results are generally better than for SAT taken directly from the forcing dataset (even with elevation corrections applied). However, at coastal stations, MAR–ERA performs worse than ERA–Interim. For MAR–20CRv2c, the difference is minimal at ice sheet stations and downscaling is detrimental at coastal stations in winter (though without elevation corrections, MAR–20CRv2c has smaller biases and MAE than 20CRv2c; see Fig. S3). Comparing MAR against all global datasets, we find MERRA2 has biases and MAE comparable to or less than MAR (all three forcings) in all seasons and regions. This is likely due to the comprehensive (relative to other reanalyses) snow/ice model in MERRA2 (Cullather et al., 2014) and reinforces the importance of atmosphere–ice sheet coupling in modeling SAT. In summer, and particularly in the ablation region, MAR and Box2013 are among the best datasets, confirming their suitability for SMB modelling. However, for SAT more generally, the benefits of RCM downscaling seem to be limited.

Another question related to the RCM downscaling is: how closely does the forcing dataset constrain climate variability in the downscaled RCM? Correlations (between 20Crv2c and MAR-20CRv2c, and between ERA–20C and MAR–ERA–20C) of ice sheet annual mean SAT before 1979 suggest that the constraint is close: for example, MAR–20CRv2c has correlation coefficients with 20CRv2c greater than 0.9 for both 1900–1940 and 1940–1980, while its correlation with other datasets is lower (0.54–0.62 for 1900–1940; 0.29–0.82 for 1940–1980). The variability of summer SAT is even more closely constrained (Fig. 6c). Downscaling is able to remedy some large biases shown by reanalysis (e.g., for ERA–20C in summer, Fig 6c), and consideration of anomalies (Fig. 6b) suggests that the downscaling improves representation of climate variability by bringing MAR–20CRv2c more into line with other datasets. Nonetheless, differences remain, particularly before 1920 and between 1950 and 1980, and we consider that MAR–20CRv2c still suffers from some shortcomings in 20CRv2c's representation of variability before 1980.

Although the comparison is for a shorter period than for other datasets, we have found that AIRS gives very good results over the ice sheet in summer – with biases and MAE values among the smallest of any dataset in the ablation region for June, July and August. However, its performance is poor in winter over the ablation region and in summer at coastal stations. The wintertime biases in the accumulation region do not agree (although those in the ablation region do) with the findings of Koenig and Hall (2010) at Summit, that satellite-derived clear-sky only temperatures were lower than all-cloud in situ measurements. They attributed this finding to the fact that clear-sky only retrievals miss winter storms – during which strong winds mix warm air from above an inversion down to the surface – which should lead to negative wintertime biases. The fact that AIRS has positive bias in the accumulation region during winter suggests compensating errors from other sources, for example from retrieval of temperature profiles or from times of day of satellite overpass. Attributing the overall bias to different causes is beyond the scope of this study. In summary, the summertime results suggest AIRS may be a useful dataset for studies of recent SMB, but further investigation is needed into the consequences of clear-sky retrievals, particularly the wintertime discrepancy with previous work and the possibility of compensating errors.

Note that there is a discrepancy between various products in calculating monthly mean SAT. As discussed in Wang and Zeng (2013) the daily mean calculated using 24 hourly values per day is different from that calculated using just maximum and minimum SAT. Comparisons for AWSs on the GrIS suggest the difference for monthly mean temperatures is ~ 0.2 ^{o}C, but can exceed 0.5 ^{o}C in some individual months. Other averaging methods (e.g., mean of 3-hourly values; weighted mean of 0800, 1400 and 2100 local time (Box, 2002)) are unlikely to introduce larger errors than the maximum plus minimum method. Overall these relatively small uncertainties are unlikely to affect our conclusions.

Our evaluation of 5 km grid box values using point measurements may also be affected by the sampling errors due to the SAT variation within a grid box (e.g., in grid boxes containing a large range of elevations and different surface types). Quantifying such an error could in principle be done using several stations within the same grid box; we do not have any 5 km grid boxes containing more than one station, however. Instead we look to the variation of elevation, assuming that this is the dominant source of SAT variation at small spatial scales and implicitly neglecting effects of varying surface type and other factors. Elevation variation at any particular location is quantified by taking the standard deviation of elevation values at the nearest and 24 surrounding grid boxes from the 1 km version of the Bamber et al. (2013) DEM. This is then multiplied by a (slightly conservative) lapse rate of 9.0 ^{o}C km^{-1}, to give a likely range of SAT variation over this elevation range. This formulation gives smaller sampling error over relatively flat terrain: ~ 0.1 ^{o}C above 1500 m on the ice sheet. In more variable terrain, around the margins of the ice sheet and in coastal land regions, sampling errors are larger – usually in the range 0.3

to 1.0 °C. Overall these uncertainties are relatively small in magnitude compared to the large biases and MAEs between various datasets and in situ observations, and hence our conclusions are largely unaffected.

In our assessment of biases and their changes through time we have assumed that all observations are un-biased. Observation biases are likely to exist (e.g., the positive bias of un-aspirated thermometer shields in low wind, high solar radiation conditions; Genthon et al., 2011) and are likely to vary in space and time due to differences in station siting, instrumentation and observing practices (e.g., number per day and timing of manual thermometer readings). By breaking down the bias assessment into two altitude bands (below and above 1500 m), our analysis aims to reduce the impact of station siting changes (e.g., a large increase in the proportion of ablation zone observations as the PROMICE network has been set up). Our analysis also, to some extent, isolates different instrument types, as the PROMICE network and K–transect stations are mostly below 1500 m, while GC–Net stations are mostly above 1500 m. Side-by-side comparisons of different instrument types, across different climatological regimes on the ice sheet, is needed for a future study to better understand the spatial and temporal patterns of bias shown here. This could include the replication of historic observing practices and instruments, to better understand, and make the most of, the limited number of mid-20th century ice sheet SAT observations.

4 Conclusions

We have assessed a number of global SAT datasets using in situ observations over Greenland, and found large differences in their performance. Reanalyses generally perform better than gridded SAT analyses – particularly at high elevations on the ice sheet. Simple elevation-based corrections applied to reanalyses lead, in most cases, to improved performance: changes in mean monthly MAE (weighted as in Table 3) vary from a 3% increase to a 42% decrease. Considering all regions and seasons, the smallest biases are seen in (elevation-corrected) MERRA2 reanalysis. Biases vary by season and by region of the ice sheet: in the ablation region (demarcated here by the 1500 m elevation contour) during summer, most reanalyses have a ~1 °C positive bias (though 20CRv2c and ERA–20C have negative biases) while CRU and Berkeley Earth gridded SAT analyses have larger positive biases. These biases have implications for SMB reconstruction, as this region and season contribute a large proportion of meltwater creation.

Among global datasets that cover the entire 20th century, 20CRv2c generally has the smallest biases and MAEs when comparing against observations made since 1979. However, combining GISTEMP anomalies with the MERRA2 climatology gives slightly better results and, given concerns about spurious long term trends in 20CRv2c (in particular, a warm bias before 1950), we recommend this type of approach (i.e., combining GISTEMP with MERRA2) to represent monthly SAT over the early and mid-20th century. Similarity of anomalies between gridded SAT analyses (except NANSENSAT) suggests that observed biases result from their climatology fields, but their anomalies are suitable alternatives to GISTEMP.

Alongside multi-decadal global SAT datasets, we have analyzed SAT from recent (2002 to present) AIRS satellite retrievals and from RCM-downscaled reanalysis. AIRS has among the smallest biases and MAE in summer months over the ice sheet, but larger errors in winter and when comparing to coastal stations. RCMs are found to reduce biases in comparison to their respective forcing datasets and provide among the best representations of SAT on the ice sheet. However, MERRA2 reanalysis performs comparably on the ice sheet, and better in comparison to coastal stations. The long term variability of RCM SAT closely follows that from the forcing dataset; the shortcomings that we highlight for 20CRv2c thus also persist, to some degree, in the version of MAR forced by 20CRv2c. MAR–ERA–20C has long term variability closer to gridded SAT analyses and long-running DMI stations, but differences remain. The Box2013 dataset, by using spatial information from a similar RCM, has similar patterns of bias to the MAR datasets. However, Box2013 inherits its long term variation from the same SAT observations as used in global SAT analyses, rather than from (as in MAR) reanalysis forcing; thus its anomalies closely follow those from CRU, GISTEMP and especially Berkeley Earth.

We have assessed CMIP5 ESMs by comparing their ice sheet average SAT with that from other datasets. A key finding is that such an assessment depends crucially on the choice of verification dataset. Using GISTEMP combined with MERRA2 climatology (due to its overall good performance in comparison with in situ observations), we find that a large number of the CMIP5 ESMs have similar ice sheet long term annual average SATs (10 within 1 ^{o}C, 19 within 2 ^{o}C). The 1901–2000 trends from most individual models and the ensemble mean are positive. For a number of sub-periods examined, some individual ESMs have negative trends, though the ensemble mean does not, highlighting the fact that the ensemble mean does not exhibit realistic decadal variability. The 1990–2005 trends are positive and larger than for earlier periods (though mostly not statistically significant) for the majority of CMIP5 ESMs analyzed here, suggesting that forced changes dominate over internal variability in this period.

Our analysis highlights several avenues for future work. Comparison of different instrument types and measurement practices would allow a quantitative assessment of the effects of instrument bias on the results shown here. Such work is also crucial to investigations of GrIS diurnal temperature variation, for example in model assessment and SMB studies using positive degree day methods (Fausto et al., 2011; Rogozhina and Rau, 2014). Results for AIRS retrievals suggest it may provide useful SAT information over the GrIS in summer, but further work is needed on the effects of only sampling clear-sky SAT. Investigation is required to establish the cause of disparities in trends and variability between 20CRv2c and ERA–20C – which are ostensibly formulated in similar ways. Possible causes include different representation of atmospheric circulation and different sea ice and sea surface temperature datasets. While RCM downscaling is currently an important tool in assessing past and future GrIS mass balance changes, our results provide new evidence that results from RCMs are highly dependent on the forcing. For SAT, RCM downscaling can reduce biases and give realistic spatial patterns compared to the forcing dataset, but does not seem to greatly alter the long term evolution of the areal average. It remains to be seen whether the same is true for SMB. The greatest SAT differences between the versions of MAR used here occur before 1980, but there are differences since 2000 too, highlighting that uncertainties in GrIS SMB exist even in the better-observed recent past.

5 Code and data availability

Most of the data used in this work are freely and publicly available. Full dataset references are given in the Supplemental Material. Derived data fields (e.g., elevation-corrected SAT) and code used to analyze data and plot figures are available from the corresponding author on request.

Acknowledgements

This research was supported by NASA (NNX14AM02G), DOE (DE-SC0016533), and the Agnese Nelms Haury Program in Environment and Social Justice. We thank Jason Box, Xavier Fettweis and an anonymous reviewer for their constructive comments and suggestions. Chris Castro and Guo-Yue Niu are thanked for useful discussions during the preparation of this manuscript. We also thank the various groups and centers for making their datasets and model results available. We thank C.J.P.P Smeets and the Institute for Marine and Atmospheric Research at the Utrecht University for providing the K–transect data. DMI AWS data were downloaded from http://www.dmi.dk/laer-om/generelt/dmi-publikationer/2013/, GC–Net AWS data from http://cires1.colorado.edu/steffen/gcnet/, and PROMICE data from http://www.promice.dk/. MERRA, MERRA2 and AIRS data were downloaded from the NASA Goddard Earth Sciences Data and Information Services Center (GES DISC). ERA–Interim and ERA–20C were downloaded from the EMCWF website (http://apps.ecmwf.int/datasets/). JRA–55, CFSR and CFSv2 data were obtained from the Research Data Archive at the National Center for Atmospheric Research (NCAR) Computational and Information Systems Laboratory. 20CRv2c data were downloaded from http://www.esrl.noaa.gov/psd/. CRU data were downloaded from the British Atmospheric Data Centre (BADC), Berkeley Earth from http://berkeleyearth.org/data/, GISTEMP from http://data.giss.nasa.gov/gistemp/, and NANSEN SAT from http://www.niersc.spb.ru. MAR data were downloaded from ftp://ftp.climato.be/fettweis/MARv3.5.2/Greenland/. CMIP5 data were obtained from the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison (http://cmip-pcmdi.llnl.gov/cmip5/data_portal.html).

References

Bamber, J. L., Layberry, R. L. and Gogineni, S. P.: A new ice thickness and bed data set for the Greenland ice sheet: 1. Measurement, data reduction, and errors, J. Geophys. Res. Atmospheres, 106(D24), 33773–33780, doi:10.1029/2001JD900054, 2001.

Bamber, J. L., Griggs, J. A., Hurkmans, R., Dowdeswell, J. A., Gogineni, S. P., Howat, I., Mouginot, J., Paden, J., Palmer, S., Rignot, E. and Steinhage, D.: A new bed elevation dataset for Greenland, The Cryosphere, 7(2), 499–510, doi:10.5194/tc-7-499-2013, 2013a.

Bamber, J. L., Griggs, J. A., Hurkmans, R. T. W. L., Dowdeswell, J. A., Gogineni, S. P., Howat, I., Mouginot, J., Paden, J., Palmer, S., Rignot, E. and Steinhage, D.: A new bed elevation dataset for Greenland, The Cryosphere, 7(2), 499–510, doi:10.5194/tc-7-499-2013, 2013b.

Box, J. E.: Survey of Greenland instrumental temperature records: 1873–2001, Int. J. Climatol., 22(15), 1829–1847, doi:10.1002/joc.852, 2002.

Box, J. E.: Greenland Ice Sheet Mass Balance Reconstruction. Part II: Surface Mass Balance (1840–2010), J. Clim., 26(18), 6974–6989, doi:10.1175/JCLI-D-12-00518.1, 2013.

Box, J. E., Yang, L., Bromwich, D. H. and Bai, L.-S.: Greenland Ice Sheet Surface Air Temperature Variability: 1840–2007, J. Clim., 22(14), 4029–4049, doi:10.1175/2009JCLI2816.1, 2009.

Burgess, E. W., Forster, R. R., Box, J. E., Mosley-Thompson, E., Bromwich, D. H., Bales, R. C. and Smith, L. C.: A spatially calibrated model of annual accumulation rate on the Greenland Ice Sheet (1958–2007), J. Geophys. Res. Earth Surf., 115(F2), F02004, doi:10.1029/2009JF001293, 2010.

Cappelen, J.: Greenland - DMI Historical Climate Data Collection 1784-2013, Technical report, Danish Meteorological Institute. [online] Available from: http://www.dmi.dk/fileadmin/user_upload/Rapporter/TR/2014/tr14-04.pdf (Accessed 3 October 2016), 2014.

Chahine, M. T., Pagano, T. S., Aumann, H. H., Atlas, R., Barnet, C., Blaisdell, J., Chen, L., Divakarla, M., Fetzer, E. J., Goldberg, M., Gautier, C., Granger, S., Hannon, S., Irion, F. W., Kakar, R., Kalnay, E., Lambrigtsen, B. H., Lee, S.-Y., Le Marshall, J., McMillan, W. W., McMillin, L., Olsen, E. T., Revercomb, H., Rosenkranz, P., Smith, W. L., Staelin, D., Strow, L. L., Susskind, J., Tobin, D., Wolf, W. and Zhou, L.: AIRS: Improving Weather Forecasting and Providing New Data on Greenhouse Gases, Bull. Am. Meteorol. Soc., 87(7), 911–926, doi:10.1175/BAMS-87-7-911, 2006.

Chylek, P., Dubey, M. K. and Lesins, G.: Greenland warming of 1920–1930 and 1995–2005, Geophys. Res. Lett., 33(11), L11707, doi:10.1029/2006GL026510, 2006.

Comiso, J. C., Yang, J., Honjo, S. and Krishfield, R. A.: Detection of change in the Arctic using satellite and in situ data, J. Geophys. Res. Oceans, 108(C12), 3384, doi:10.1029/2002JC001347, 2003.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D. and Worley, S. J.: The Twentieth Century Reanalysis Project, Q. J. R. Meteorol. Soc., 137(654), 1–28, doi:10.1002/qj.776, 2011.

Compo, G. P., Sardeshmukh, P. D., Whitaker, J. S., Brohan, P., Jones, P. D. and McColl, C.: Independent confirmation of global land warming without the use of station temperatures, Geophys. Res. Lett., 40(12), 3170–3174, doi:10.1002/grl.50425, 2013.

Cullather, R. I., Nowicki, S. M. J., Zhao, B. and Suarez, M. J.: Evaluation of the Surface Representation of the Greenland Ice Sheet in a General Circulation Model, J. Clim., 27(13), 4835–4856, doi:10.1175/JCLI-D-13-00635.1, 2014.

Cullather, R. I., Nowicki, S. M. J., Zhao, B. and Koenig, L. S.: A Characterization of Greenland Ice Sheet Surface Melt and Runoff in Contemporary Reanalyses and a Regional Climate Model, Cryospheric Sci., 10, doi:10.3389/feart.2016.00010, 2016.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. R. Meteorol. Soc., 137(656), 553–597, doi:10.1002/qj.828, 2011.

Ettema, J., van den Broeke, M. R., van Meijgaard, E., van de Berg, W. J., Box, J. E. and Steffen, K.: Climate of the Greenland ice sheet using a high-resolution climate model – Part 1: Evaluation, The Cryosphere, 4(4), 511–527, doi:10.5194/tc-4-511-2010, 2010a.

Ettema, J., van den Broeke, M. R., van Meijgaard, E. and van de Berg, W. J.: Climate of the Greenland ice sheet using a high-resolution climate model – Part 2: Near-surface climate and energy balance, The Cryosphere, 4(4), 529–544, doi:10.5194/tc-4-529-2010, 2010b.

Fausto, R. S., Ahlstrøm, A. P., Van As, D., Bøggild, C. E. and Johnsen, S. J.: A new present-day temperature parameterization for Greenland, J. Glaciol., 55(189), 95–105, doi:10.3189/002214309788608985, 2009.

Fausto, R. S., Ahlstrøm, A. P., Van As, D. and Steffen, K.: Present-day temperature standard deviation parameterization for Greenland, J. Glaciol., 57(206), 1181–1183, doi:10.3189/002214311798843377, 2011.

Fettweis, X., Hanna, E., Gallee, H., Huybrechts, P. and Erpicum, M.: Estimation of the Greenland ice sheet surface mass balance for the 20th and 21st centuries, The Cryosphere, 2(2), 117–129, 2008.

Fettweis, X., Franco, B., Tedesco, M., van Angelen, J. H., Lenaerts, J. T. M., van den Broeke, M. R. and Gallée, H.: Estimating the Greenland ice sheet surface mass balance contribution to future sea level rise using the regional atmospheric climate model MAR, The Cryosphere, 7(2), 469–489, doi:10.5194/tc-7-469-2013, 2013.

Fettweis, X., Box, J. E., Agosta, C., Amory, C., Kittel, C., Lang, C., van As, D., Machguth, H. and Gallée, H.: Reconstructions of the 1900–2015 Greenland ice sheet surface mass balance using the regional climate MAR model, The Cryosphere, 11(2), 1015–1033, doi:10.5194/tc-11-1015-2017, 2017.

Genthon, C., Six, D., Favier, V., Lazzara, M. and Keller, L.: Atmospheric Temperature Measurement Biases on the Antarctic Plateau, J. Atmospheric Ocean. Technol., 28(12), 1598–1605, doi:10.1175/JTECH-D-11-00095.1, 2011.

Hall, D. K., Box, J. E., Casey, K. A., Hook, S. J., Shuman, C. A. and Steffen, K.: Comparison of satellite-derived and in-situ observations of ice and snow surface temperatures over Greenland, Remote Sens. Environ., 112(10), 3739–3749, doi:10.1016/j.rse.2008.05.007, 2008.

Hall, D. K., Comiso, J. C., DiGirolamo, N. E., Shuman, C. A., Box, J. E. and Koenig, L. S.: Variability in the surface temperature and melt extent of the Greenland ice sheet from MODIS, Geophys. Res. Lett., 40(10), 2114–2120, doi:10.1002/grl.50240, 2013.

Hanna, E., Huybrechts, P., Janssens, I., Cappelen, J., Steffen, K. and Stephens, A.: Runoff and mass balance of the Greenland ice sheet: 1958–2003, J. Geophys. Res. Atmospheres, 110(D13), D13108, doi:10.1029/2004JD005641, 2005.

Hanna, E., Huybrechts, P., Cappelen, J., Steffen, K., Bales, R. C., Burgess, E., McConnell, J. R., Peder Steffensen, J., Van den Broeke, M., Wake, L., Bigg, G., Griffiths, M. and Savas, D.: Greenland Ice Sheet surface mass balance 1870 to 2010 based on Twentieth Century Reanalysis, and links with global climate forcing, J. Geophys. Res. Atmospheres, 116(D24), D24121, doi:10.1029/2011JD016387, 2011.

Hanna, E., Mernild, S. H., Cappelen, J. and Steffen, K.: Recent warming in Greenland in a long-term instrumental (1881–2012) climatic context: I. Evaluation of surface air temperature records, Environ. Res. Lett., 7(4), 45404, doi:10.1088/1748-9326/7/4/045404, 2012.

Hansen, J., Ruedy, R., Sato, M. and Lo, K.: Global Surface Temperature Change, Rev. Geophys., 48(4), RG4004, doi:10.1029/2010RG000345, 2010.

Harris, I., Jones, P. d., Osborn, T. j. and Lister, D. h.: Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset, Int. J. Climatol., 34(3), 623–642, doi:10.1002/joc.3711, 2014.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K. and Takahashi, K.: The JRA-55 Reanalysis: General Specifications and Basic Characteristics, J. Meteorol. Soc. Jpn. Ser II, 93(1), 5–48, doi:10.2151/jmsj.2015-001, 2015.

Koenig, L. S. and Hall, D. K.: Comparison of satellite, thermochron and air temperatures at Summit, Greenland, during the winter of 2008/09, J. Glaciol., 56(198), 735–741, doi:10.3189/002214310793146269, 2010.

Kuzmina, S. I., Johannessen, O. M., Bengtsson, L., Aniskina, O. G. and Bobylev, L. P.: High northern latitude surface air temperature: comparison of existing data and creation of a new gridded data set 1900–2000, Tellus A, 60(2), doi:10.3402/tellusa.v60i2.15260, 2008.

Lucas-Picher, P., Wulff-Nielsen, M., Christensen, J. H., Aðalgeirsdóttir, G., Mottram, R. and Simonsen, S. B.: Very high resolution regional climate model simulations over Greenland: Identifying added value, J. Geophys. Res. Atmospheres, 117(D2), D02108, doi:10.1029/2011JD016267, 2012.

Mernild, S. H., Hanna, E., Yde, J. C., Cappelen, J. and Malmros, J. K.: Coastal Greenland air temperature extremes and trends 1890–2010: annual and monthly analysis, Int. J. Climatol., 34(5), 1472–1487, doi:10.1002/joc.3777, 2014.

Molod, A., Takacs, L., Suarez, M. and Bacmeister, J.: Development of the GEOS-5 atmospheric general circulation model: evolution from MERRA to MERRA2, Geosci Model Dev, 8(5), 1339–1356, doi:10.5194/gmd-8-1339-2015, 2015.

Morice, C. P., Kennedy, J. J., Rayner, N. A. and Jones, P. D.: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, J. Geophys. Res. Atmospheres, 117(D8), D08101, doi:10.1029/2011JD017187, 2012.

Noël, B., van de Berg, W. J., van Meijgaard, E., Kuipers Munneke, P., van de Wal, R. S. W. and van den Broeke, M. R.: Evaluation of the updated regional climate model RACMO2.3: summer snowfall impact on the Greenland Ice Sheet, The Cryosphere, 9(5), 1831–1844, doi:10.5194/tc-9-1831-2015, 2015.

Noël, B., van de Berg, W. J., Machguth, H., Lhermitte, S., Howat, I., Fettweis, X. and van den Broeke, M. R.: A daily, 1 km resolution data set of downscaled Greenland ice sheet surface mass balance (1958–2015), The Cryosphere, 10(5), 2361–2377, doi:10.5194/tc-10-2361-2016, 2016.

Ohmura, A.: New temperature distribution maps for Greenland, Z. Gletscherkunde Glaziolgeologie, 23(1), 1–45, 1987.

Poli, P., Hersbach, H., Dee, D. P., Berrisford, P., Simmons, A. J., Vitart, F., Laloyaux, P., Tan, D. G. H., Peubey, C., Thépaut, J.-N., Trémolet, Y., Hólm, E. V., Bonavita, M., Isaksen, L. and Fisher, M.: ERA-20C: An Atmospheric Reanalysis of the Twentieth Century, J. Clim., 29(11), 4083–4097, doi:10.1175/JCLI-D-15-0556.1, 2016.

Rae, J. G. L., Adalgeirsdottir, G., Edwards, T. L., Fettweis, X., Gregory, J. M., Hewitt, H. T., Lowe, J. A., Lucas-Picher, P., Mottram, R. H., Payne, A. J., Ridley, J. K., Shannon, S. R., van de Berg, W. J., van de Wal, R. S. W. and van den Broeke, M. R.: Greenland ice sheet surface mass balance: evaluating simulations and making projections with regional climate models, The Cryosphere, 6, 1275–1294, doi:10.5194/tc-6-1275-2012, 2012.

Ridley, J. K., Huybrechts, P., Gregory, J. M. and Lowe, J. A.: Elimination of the Greenland ice sheet in a high CO2 climate, J. Clim., 18(17), 3409–3427, 2005.

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G.-K., Bloom, S., Chen, J., Collins, D., Conaty, A., da Silva, A., Gu, W., Joiner, J., Koster, R. D., Lucchesi, R., Molod, A., Owens, T., Pawson, S., Pegion, P., Redder, C. R., Reichle, R., Robertson, F. R., Ruddick, A. G., Sienkiewicz, M. and Woollen, J.: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications, J. Clim., 24(14), 3624–3648, doi:10.1175/JCLI-D-11-00015.1, 2011.

Rogozhina, I. and Rau, D.: Vital role of daily temperature variability in surface mass balance parameterizations of the Greenland Ice Sheet, The Cryosphere, 8(2), 575–585, doi:10.5194/tc-8-575-2014, 2014.

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., Rosenfeld, A., Wurtele, J., Curry, J., Wickham, C. and Mosher, S.: Berkeley Earth Temperature Averaging Process, Geoinformatics Geostat. Overv., 1(2), doi:10.4172/2327-4581.1000103, 2013.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.-T., Chuang, H.-Y., Juang, H.-M. H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Van Delst, P., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., Van Den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J.-K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.-Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R. W., Rutledge, G. and Goldberg, M.: The NCEP Climate Forecast System Reanalysis, Bull. Am. Meteorol. Soc., 91(8), 1015–1057, doi:10.1175/2010BAMS3001.1, 2010.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M. and Becker, E.: The NCEP Climate Forecast System Version 2, J. Clim., 27(6), 2185–2208, doi:10.1175/JCLI-D-12-00823.1, 2014.

Santer, B. D., Wigley, T. M. L., Boyle, J. S., Gaffen, D. J., Hnilo, J. J., Nychka, D., Parker, D. E. and Taylor, K. E.: Statistical significance of trends and trend differences in layer-average atmospheric temperature time series, J. Geophys. Res. Atmospheres, 105(D6), 7337–7356, doi:10.1029/1999JD901105, 2000.

Screen, J. A. and Simmonds, I.: Erroneous Arctic Temperature Trends in the ERA-40 Reanalysis: A Closer Look, J. Clim., 24(10), 2620–2627, doi:10.1175/2010JCLI4054.1, 2010.

Shuman, C. A., Alley, R. B., Anandakrishnan, S. and Stearns, C. R.: An empirical technique for estimating near-surface air temperature trends in central Greenland from SSM/I brightness temperatures, Remote Sens. Environ., 51(2), 245–252, doi:10.1016/0034-4257(94)00086-3, 1995.

Shuman, C. A., Steffen, K., Box, J. E. and Stearns, C. R.: A Dozen Years of Temperature Observations at the Summit: Central Greenland Automatic Weather Stations 1987–99, J. Appl. Meteorol., 40(4), 741–752, doi:10.1175/1520-0450(2001)040<0741:ADYOTO>2.0.CO;2, 2001.

Smith, T. M., Reynolds, R. W., Peterson, T. C. and Lawrimore, J.: Improvements to NOAA's Historical Merged Land–Ocean Surface Temperature Analysis (1880–2006), J. Clim., 21(10), 2283–2296, doi:10.1175/2007JCLI2100.1, 2008.

Steffen, K. and Box, J.: Surface climatology of the Greenland Ice Sheet: Greenland Climate Network 1995–1999, J. Geophys. Res. Atmospheres, 106(D24), 33951–33964, doi:10.1029/2001JD900161, 2001.

Taylor, K. E., Stouffer, R. J. and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bull. Am. Meteorol. Soc., 93(4), 485–498, doi:10.1175/BAMS-D-11-00094.1, 2011.

Uppala, S. M., KÅllberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L. V. D., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. a. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P. and Woollen, J.: The ERA-40 re-analysis, Q. J. R. Meteorol. Soc., 131(612), 2961–3012, doi:10.1256/qj.04.176, 2005.

Van As, D., Fausto, R. S., Ahlstrøm, A. P., Andersen, S. B., Andersen, M. L., Citterio, M., Edelvang, K., Gravesen, P., Machguth, H., Nick, F. M., Nielsen, S. and Weidick, A.: Programme for Monitoring of the Greenland Ice Sheet (PROMICE): first temperature and ablation records, Geol. Surv. Den. Greenl. Bull., 23, 73–76, 2011.

Vernon, C. L., Bamber, J. L., Box, J. E., van den Broeke, M. R., Fettweis, X., Hanna, E. and Huybrechts, P.: Surface mass balance model intercomparison for the Greenland ice sheet, The Cryosphere, 7(2), 599–614, doi:10.5194/tc-7-599-2013, 2013.

Vizcaíno, M., Lipscomb, W. H., Sacks, W. J., van Angelen, J. H., Wouters, B. and van den Broeke, M. R.: Greenland Surface Mass Balance as Simulated by the Community Earth System Model. Part I: Model Evaluation and 1850–2005 Results, J. Clim., 26(20), 7793–7812, doi:10.1175/JCLI-D-12-00615.1, 2013.

Vose, R. S., Arndt, D., Banzon, V. F., Easterling, D. R., Gleason, B., Huang, B., Kearns, E., Lawrimore, J. H., Menne, M. J., Peterson, T. C., Reynolds, R. W., Smith, T. M., Williams, C. N. and Wuertz, D. B.: NOAA's Merged Land–Ocean Surface Temperature Analysis, Bull. Am. Meteorol. Soc., 93(11), 1677–1685, doi:10.1175/BAMS-D-11-00241.1, 2012.

van de Wal, R. S. W., Greuell, W., van den Broeke, M. R., Reijmer, C. H. and Oerlemans, J.: Surface mass-balance observations and automatic weather station data along a transect near Kangerlussuaq, West Greenland, Ann. Glaciol., 42(1), 311–316, doi:10.3189/172756405781812529, 2005.

Wang, A. and Zeng, X.: Development of Global Hourly 0.5° Land Surface Air Temperature Datasets, J. Clim., 26(19), 7676–7691, doi:10.1175/JCLI-D-12-00682.1, 2013.

Table 1: Temperature products assessed in this work. Latitude longitude spacing refers to the grids downloaded for this work (not necessarily the native model grid). Maximum output frequency refers to the maximum available – monthly averages are used in the analysis.

Type	Dataset	Center	Latitude longitude spacing [a]	Maximum output frequency	Period	Reference
Reanalysis	MERRA	NASA/ GMAO	0.5^{o} x 0.667^{o}	Hourly	1979–2015	Rienecker et al., 2011
	MERRA2	NASA/ GMAO	0.5^{o} x 0.625^{o}	Hourly	1980–2015	Molod et al., 2015
	CFSR and CFSv2 [b]	NCEP	0.5^{o} x 0.5^{o}	Hourly	1979–2015	Saha et al., 2010; Saha et al., 2014
	20th Century Reanalysis V2c	NOAA/ CIRES	~1.9^{o} x 1.875^{o}	3–hourly	1851–2014	Compo et al., 2011
	ERA–Interim	ECMWF	0.75^{o} x 0.75^{o}	3–hourly	1979–2015	Dee et al., 2011
	ERA–20C	ECMWF	1^{o} x 1^{o}	3–hourly	1900–2010	Poli et al., 2016
	JRA–55	JMA	~0.56^{o} x ~0.56^{o}	3–hourly	1958–2014	Kobayashi et al., 2015
Gridded temperature analysis	GISTEMP	NASA/ GISS	2^{o} x 2^{o}	Monthly	1880–2015	Hansen et al., 2010
	CRU TS 3.23	CRU	0.5^{o} x 0.5^{o}	Monthly	1901–2014	Harris et al., 2014
	Berkeley Earth Surface temperature	Berkeley Earth	1^{o} x 1^{o}	Monthly	1750–2016	Rohde et al., 2013
	NANSENSAT	Nansen Centers	2.5^{o} x 2.5^{o}	Monthly	1900–2008	Kuzmina et al., 2008
	Box2013	GEUS	5 km x 5 km [c]	Monthly	1840–2014	Box, 2013
Satellite	AIRS	NASA	1^{o} x 1^{o}	Monthly	2002–2015	Chahine et al., 2006
Regional down-scaling	MAR–ERA	University of Liège	5 km x 5 km [c]	Monthly	1958–2015	Fettweis et al., 2013, 2017
	MAR–20CRv2c				1900–2014	
	MAR–ERA–20C				1900–2010	

[a] as downloaded for this study; [b] CFSR, covering 1979–2010, and CFSv2, covering 2011–2015, are appended and referred to together as CFSR in the text; [c] Box2013 and MAR are on the polar stereographic grid of Bamber et al. (2001).

Table 2: Error statistics of model elevation fields (interpolated to EASE grid, except for MAR) relative to the digital elevation model (DEM) of Bamber et al. (2013). Bias and deciles are calculated as (model minus DEM). Averages are taken over all ice sheet grid points, classified using the mask of Bamber et al. (2013).

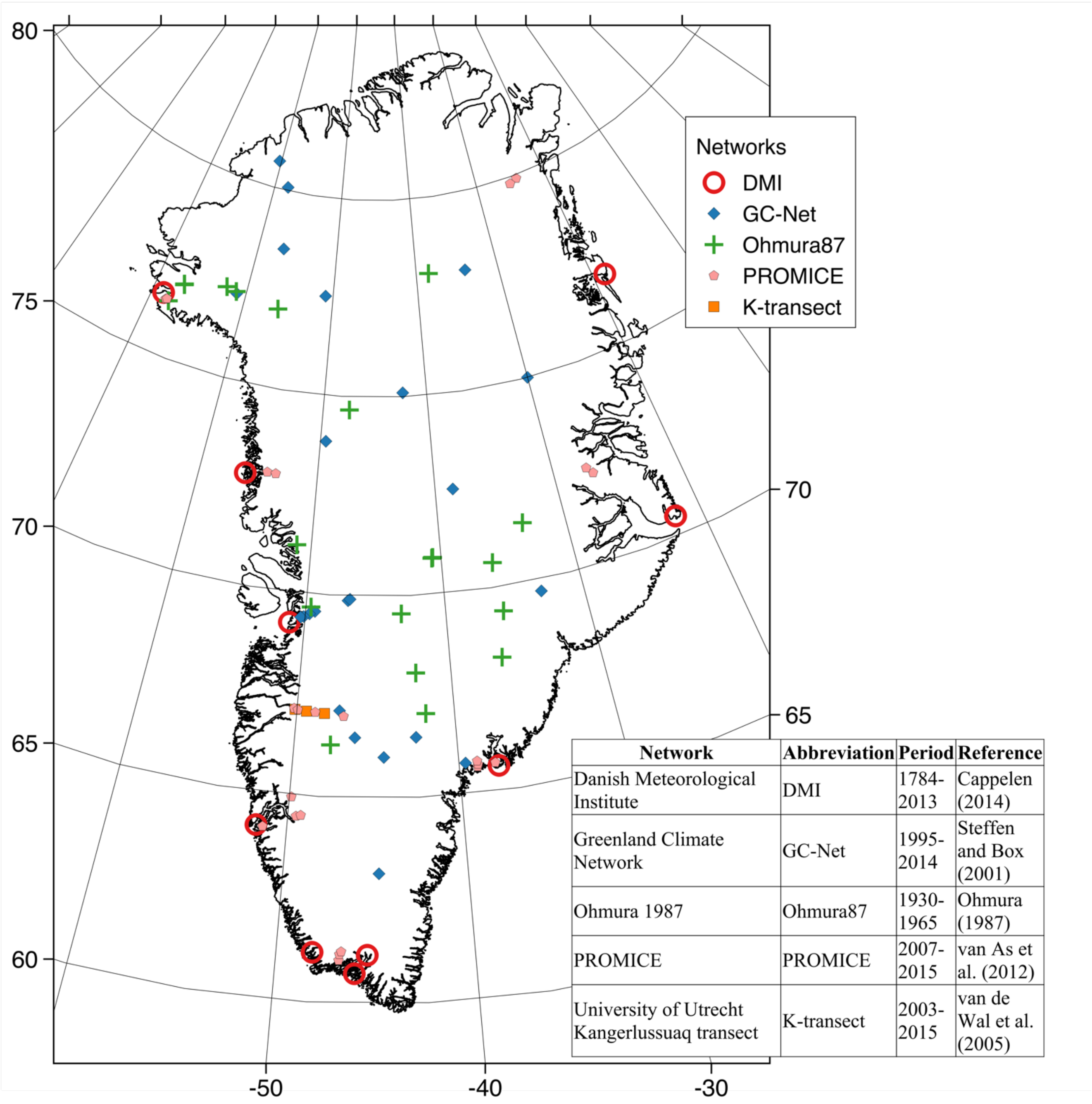
Dataset	Bias (m)	RMSE (m)	MAE (m)	Lower decile (m)	Upper decile (m)
MERRA	-126.3	290.1	199.7	-466.3	141.7
MERRA2	-48.3	172.3	88.4	-194.6	16.9
ERA–Interim	-67.5	215.0	119.5	-281.3	33.3
ERA–20C	-103.1	274.6	173.0	-380.9	51.2
CFSR	-114.9	262.8	192.0	-422.0	143.1
20CRv2c	-244.3	447.1	337.6	-733.8	151.8
JRA–55	-131.9	272.4	199.2	-439.9	123.5
AIRS	-132.3	274.2	200.3	-440.1	120.2
MAR	-13.4	94.1	37.2	-56.0	18.0

Table 3. Mean bias and mean absolute error (MAE) for all datasets ranked from smallest (top) to largest (bottom) MAE. These numbers represent an average of results from Figs. 3 and 4, with unweighted average over months and an area-weighted average over glaciological regimes (64.9% ice sheet above 1500m, 18.6% ice sheet below 1500 m and 16.5% coastal DMI).

Dataset	Bias (^{o} C)	MAE (^{o} C)
MERRA2	0.81	1.27
AIRS	0.95	1.67
MAR–ERA	1.38	1.72
MERRA	1.26	1.94
MAR–ERA–20C	1.69	2.04
MAR–20CRv2c	1.59	2.05
20CRv2c	1.04	2.07
Box2013	1.47	2.08
ERA–interim	1.76	2.16
CFSR	1.87	2.28
JRA–55	1.95	2.34
ERA–20C	2.15	2.52
Berkeley Earth	2.67	3.21
NANSENSAT	3.42	3.74
CRU TS3.23	3.96	4.34

Table 4: Trends (^{o}C per decade) of ice sheet areal average annual mean SAT for given periods. For the CMIP5 ESMs, the trend of the ensemble mean and the median of the individual trends are shown, along with the number of positive, negative and significant trends. Bold type indicates significance at the 0.05 level.

Dataset	Linear trends (^{o}C decade^{-1})					
	1901–1930	1931–1960	1931–1990	1990–2005	1990–2014	1901–2000
MERRA				**1.495**	**0.823**	
MERRA2				**1.739**	**0.696**	
ERA–Interim				**1.979**	**0.907**	
ERA–20C	0.513	0.023	**-0.183**	**1.637**		0.020
CFSR				**1.526**	**0.589**	
20CRv2c	**0.393**	-0.468	**-0.295**	**2.042**	**1.045**	**-0.156**
JRA–55				**1.799**	**0.949**	
MAR–ERA				**1.358**	**0.559**	
MAR–20CRv2c	0.161	-0.120	**-0.156**	**1.649**	**0.806**	-0.061
MAR–ERA–20C	**0.506**	-0.019	**-0.212**	**1.619**		-0.025
Box (2013)	0.774	-0.402	**-0.312**	**2.196**	**1.343**	0.035
CRU TS3.23	0.411	-0.118	**-0.169**	**1.249**	0.505	0.023
Berkeley Earth	0.628	-0.334	**-0.227**	**1.703**	**0.965**	0.054
GISTEMP	0.361	-0.197	**-0.139**	**1.447**	**0.865**	0.100
NANSENSAT	-0.065	-1.003	**-0.497**	**1.066**		**-0.239**
CMIP5: ensemble mean	0.119	0.016	**0.088**	**0.485**		**0.098**
CMIP5: median	0.081	0.007	0.086	0.407		0.094
CMIP5: number positive (significant)	23 (3)	16 (0)	23 (7)	27 (4)		30 (20)
CMIP5: number negative (significant)	7 (0)	14 (1)	7 (0)	4 (0)		1 (1)



Network	Abbreviation	Period	Reference
Danish Meteorological Institute	DMI	1784-2013	Cappelen (2014)
Greenland Climate Network	GC-Net	1995-2014	Steffen and Box (2001)
Ohmura 1987	Ohmura87	1930-1965	Ohmura (1987)
PROMICE	PROMICE	2007-2015	van As et al. (2012)
University of Utrecht Kangerlussuaq transect	K-transect	2003-2015	van de Wal et al. (2005)

Figure 1: Map of study area and weather stations used in this work. Symbol types represent the different monitoring networks summarized in the inserted table.

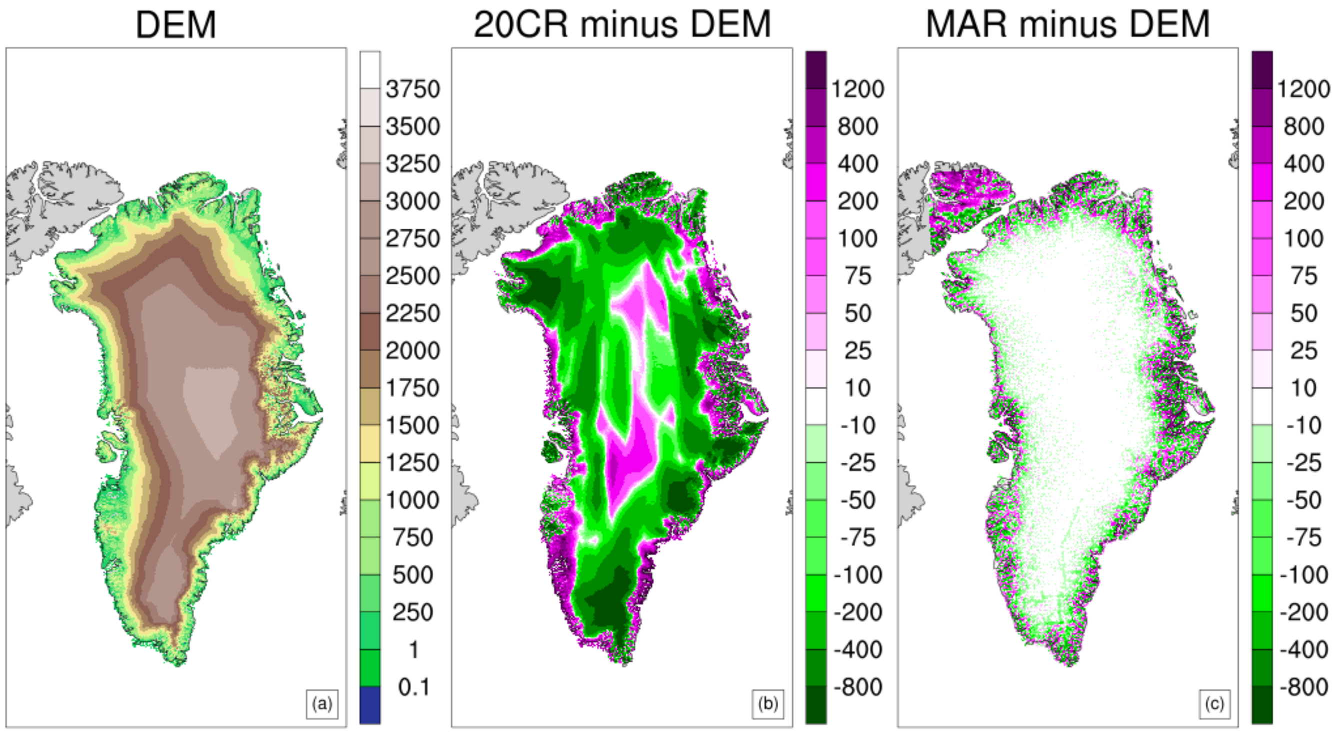


Figure 2: (a) Digital elevation model (DEM) of Bamber et al. (2013) interpolated to EASE 5 km grid; (b) bias of 20CRv2c surface elevation field interpolated to EASE grid, relative to Bamber et al. (2013); (c) bias of MAR surface elevation field relative to Bamber et al. (2013). Units are meters.

5

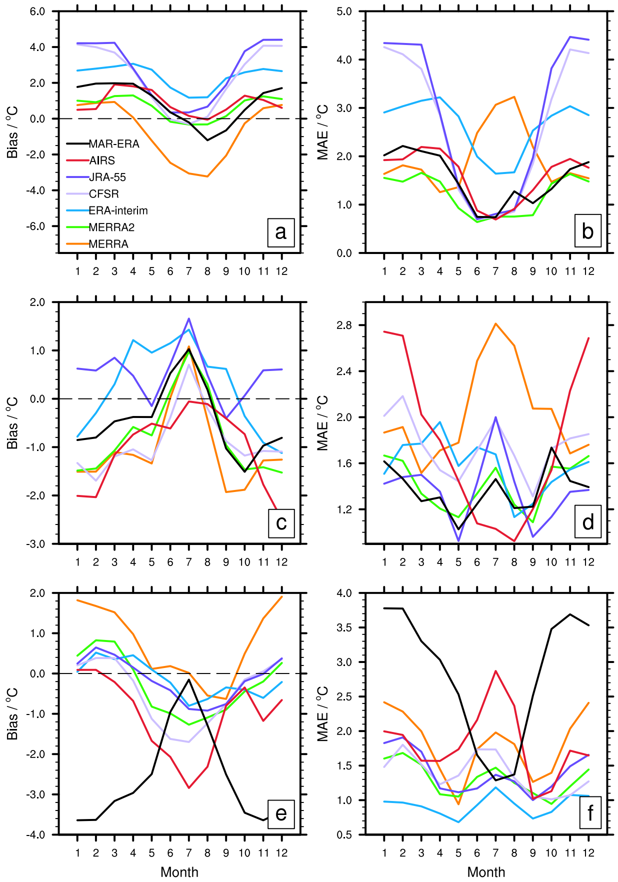


Figure 3: Mean over station-months of bias (a, c, e) and absolute error (b, d, f) relative to monthly mean SAT at: ice sheet stations above 1500 m (a and b); ice sheet stations below 1500 m (c and d); and coastal (DMI) stations (e and f). Ice sheet stations are from GC–Net, PROMICE and K–transect. All available station months from 1979 onwards are used. All datasets included in this figure are elevation corrected: several shorter reanalyses, AIRS, and MAR–ERA. Note that the vertical scales vary with panels. 5

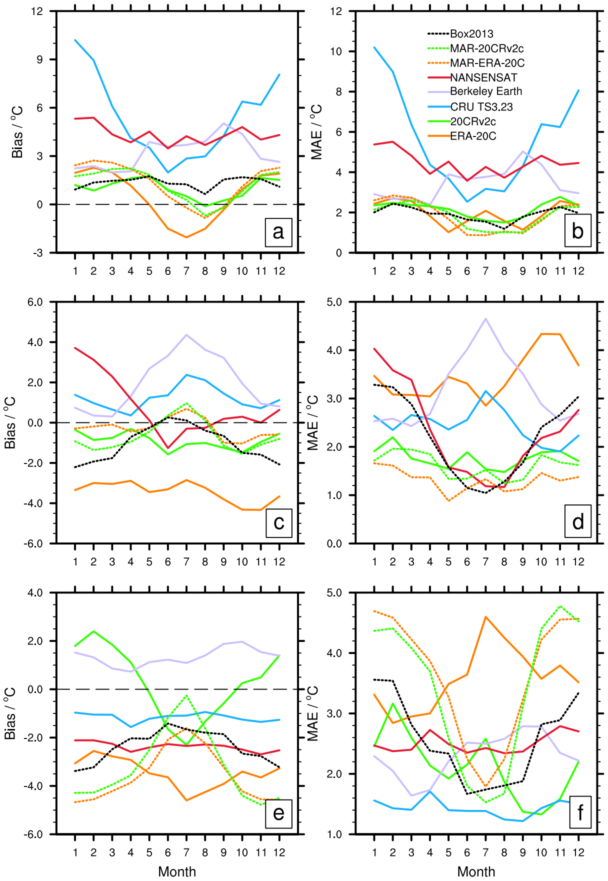


Figure 4: As in Fig. 3, but for elevation-corrected long reanalyses, MAR–ERA–20C and MAR–20CRv2C, Box2013 data and three gridded SAT analyses (not elevation corrected).

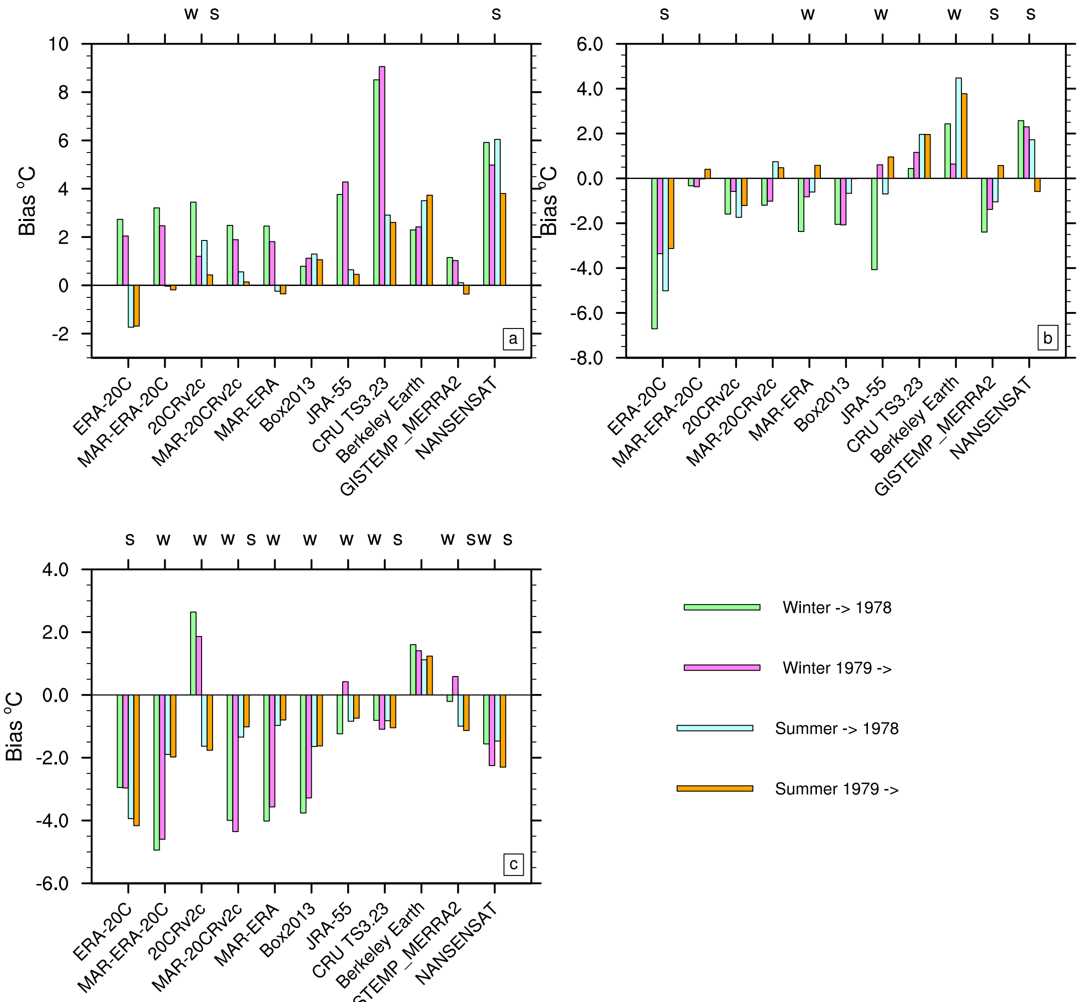


Figure 5: Monthly mean SAT bias for winter (DJF) and summer months (JJA) before and after 1979, for all datasets that extend back before 1979 (elevation-corrected where applicable) at: ice sheet stations above 1500 m (a); ice sheet stations below 1500 m (b); and coastal (DMI) stations (c). Note that these are monthly SAT biases averaged over all months in a season, not biases of seasonal mean SAT. Changes significant at the 99% level (using Student's *t*-test with unequal variances) are denoted by w (for winter) and s (for summer) on the top axis.

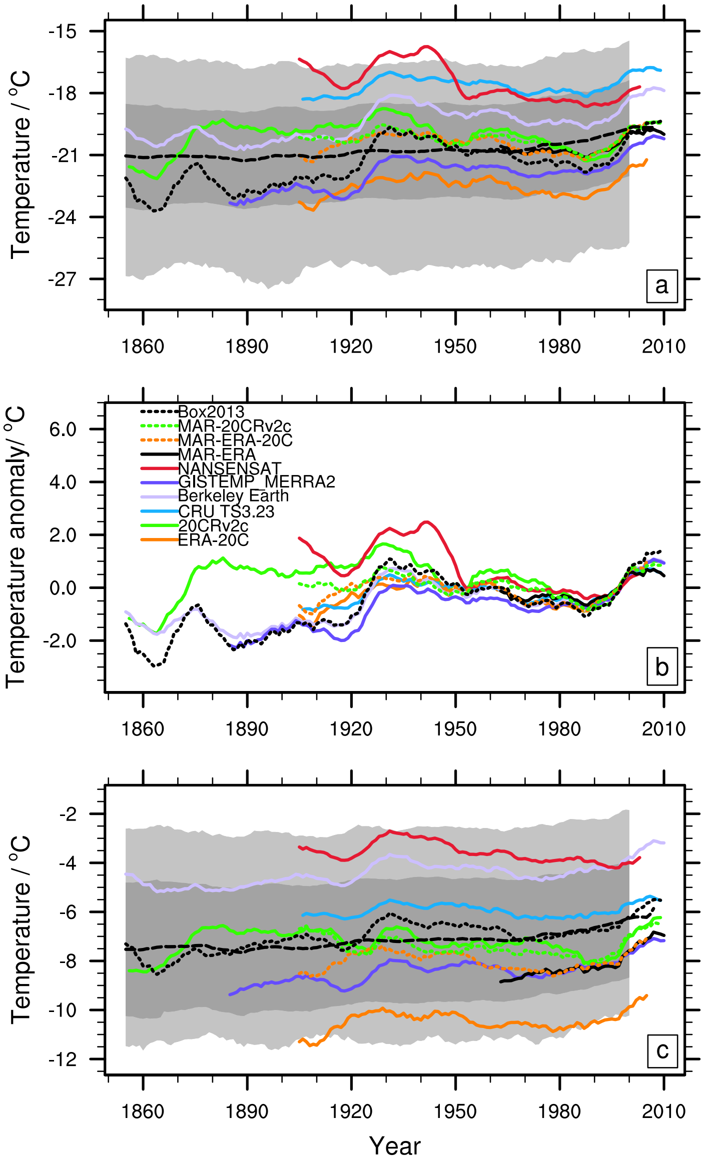


Figure 6: (a) Time series of ice sheet areal average smoothed annual mean SAT for long reanalyses, gridded temperature analyses and both MAR variants (elevation corrected where applicable; colored lines) and CMIP5 climate models (not elevation corrected; ensemble mean in dashed black line; +/- 1 standard deviation in dark grey shading; maxima and minima in light grey shading). (b) Anomalies of ice sheet areal average smoothed annual mean SAT from long reanalyses, gridded temperature analyses and both MAR variants, relative to 1981-2010 mean. (c) As in (a) but for June-August mean SAT. In all panels, time series are smoothed using a centered, uniform-weighted 11-year window, to highlight decadal variability and aid legibility.