JGR Atmospheres

RESEARCH ARTICLE
10.1029/2019JD030897

Assessment of Atmospheric Reanalyses With Independent Observations in the Weddell Sea, the Antarctic

M. O. Jonassen[1,2], I. Välisuo[3], T. Vihma[3], P. Uotila[4], A. P. Makshtas[5], and J. Launiainen[4]

[1]Department of Arctic Geophysics, University Centre in Svalbard, Longyearbyen, Norway, [2]Geophysical Institute, University of Bergen, Bergen, Norway, [3]Finnish Meteorological Institute, Helsinki, Finland, [4]Institute for Atmospheric and Earth System Research, University of Helsinki, Helsinki, Finland, [5]Arctic and Antarctic Research Institute, St. Petersburg, Russia

Key Points:
- ERA Interim and CFSR performed best for meteorological variables and downward radiation
- All four reanalyses assessed are generally too warm in the boundary layer
- The consistency of the ABL warm bias over decades calls for improvement of ABL parameterizations

Supporting Information:
- Supporting Information S1
- Table S1

Correspondence to:
M. O. Jonassen,
marius.jonassen@unis.no

Citation:
Jonassen, M. O., Välisuo, I., Vihma, T., Uotila, P., Makshtas, A. P., & Launiainen, J. (2019). Assessment of atmospheric reanalyses with independent observations in the weddell sea, the antarctic. *Journal of Geophysical Research: Atmospheres*, *124*, 12,468–12,484. https://doi.org/10.1029/2019JD030897

Received 26 APR 2019
Accepted 26 OCT 2019
Accepted article online 11 NOV 2019
Published online 3 DEC 2019

Abstract Surface layer and upper-air in situ observations from two research vessel cruises and an ice station in the Weddell Sea from 1992 and 1996 are used to validate four current atmospheric reanalysis products: ERA-Interim, CFSR, JRA-55, and MERRA-2. Three of the observation data sets were not available for assimilation, providing a rare opportunity to validate the reanalyses in the otherwise datasparse region of the Antarctic against independent data. All four reanalyses produce 2 m temperatures warmer than the observations, and the biases vary from +2.0 K in CFSR to +2.8 K in MERRA-2. All four reanalyses are generally too warm also higher up in the atmospheric boundary layer (ABL), with biases up to +1.4 K (ERA-Interim). Cloud fractions are relatively poorly reproduced by the reanalyses, MERRA-2 and JRA-55 having the strongest positive and negative biases of about +30 % and −17 %, respectively. Skill scores of the error statistics reveal that ERA-Interim compares generally the most favorably against both the surface layer and the upper-air observations. CFSR compares the second best and JRA-55 and MERRA-2 have the least favorable scores. The ABL warm bias is consistent with previous evaluation studies in high latitudes, where more recent observations have been applied. As the amount of observations has varied depending on the decade, season, and region, the consistency of the warm bias suggests a need to improve the modeling systems, including data assimilation as well as ABL and surface parameterizations.

Plain Language Summary Surface layer and upper-air in situ observations from two research vessel cruises and an ice station in the Weddell Sea from 1992 and 1996 are used to validate four atmospheric reanalyses products. Three of the observation data sets were not available in compiling the reanalyses. This provides a rare opportunity to validate the reanalyses in the otherwise datasparse region of the Antarctic against independent data. The reanalyses differ in performance. However, all four reanalyses have large errors in the cloud cover, and they also generally display too high temperatures in the lowermost part of the atmosphere. The latter finding is consistent with previous validation studies in polar regions, in which more recent observations have been applied. As the amount of observations has varied depending on the decade, season, and region, the consistency of the warm bias suggests a need to improve the representation of physical processes in the lowest parts of the atmosphere in the reanalyses investigated.

1. Introduction

Reanalyses combine observations and a numerical prediction model providing four-dimensional gridded and dynamically coherent data with full spatial and temporal coverage that are used for a wide range of applications. Their usefulness is particularly high in the Arctic and Antarctic, where observational data are otherwise sparse and unevenly distributed (Bromwich et al., 2013). Atmospheric reanalyses are also used for reconstructing near-surface temperature (Nicolas & Bromwich, 2014; Steig et al., 2009), evaluating climate models (Perez et al., 2014; Rinke et al., 2006), and providing boundary conditions for land surface models, ice-ocean models, and limited area atmospheric models (Assmann et al., 2013; Dutrieux et al., 2014; Lindsay et al., 2014). The applications of reanalysis data, for example, the provision of input data to models that are highly sensitive to the forcing conditions, make it important to evaluate reanalyses against available independent observational data, which are often rare, particularly in the polar regions. Further, reanalyses are broadly applied in estimating climatological trends, but more attention is needed on the reliability of the trends based on reanalyses (Chung et al., 2013). This calls for evaluation studies based on observations that

are not restricted to recent years but also cover periods when less observations were available for assimilation.

Early reanalyses are known to have larger errors in the Arctic and Antarctic with respect to, for example, wind speed and direction, air humidity, air temperature, cloud cover, and radiative fluxes (Bromwich et al., 2007; Screen & Simmonds, 2011; Sturaro, 2003; Vancoppenolle et al., 2011; Walsh & Chapman, 1998). Due to these shortcomings, extensive work has been carried out in producing new reanalyses, incorporating among others more sophisticated assimilation methods, better representation of sea-ice and land-surface processes, and better horizontal and vertical grid resolution. Examples of these current reanalyses are ERA-Interim (Dee et al., 2011) and ERA-5 (Copernicus Climate Change Service Climate Data Store [CDS]) from the European Centre for Medium-Range Weather Forecasts (ECMWF), the Climate Forecast System Reanalysis (CFSR; Saha et al., 2006) from the National Centers for Environmental Prediction (NCEP), the Japanese 55-year reanalysis (JRA-55; Kobayashi et al., 2015) from the Japan Meteorological Agency (JMA), and MERRA Version 2 (MERRA-2; Gelaro et al., 2017) from the National Aeronautics and Space Administration (NASA). Even though substantial progress has been made in these products with respect to their predecessors (Nygård et al., 2016; Tastula et al., 2013), also these have their deficiencies. For example, spurious warming trends have been identified in many parts of East Antarctica (Y. Wang et al., 2016) and near-surface cold biases have been found along the Antarctic coast (Bracegirdle & Marshall, 2012; P. D. Jones & Lister, 2014; R. W. Jones et al., 2016b).

In this paper we evaluate the ERA-Interim, CFSR, JRA-55, and MERRA-2 reanalyses against a range of in situ observational data from the Weddell Sea. These data include vertical profiles and surface layer time series of temperature, humidity, and wind speed as well as radiative and turbulent surface fluxes and cloud fraction. To extend the evaluation more toward process level, we compare the relationships between different variables in observations and reanalyses. Our main data source is the Ice Station Weddell (ISW), which was a U.S.-Russian campaign conducted from February to June in 1992, providing all of the above-mentioned types of data (Gordon & Ice Station Weddell Group of Principal Investigators and Chief Scientists, 1993). Much of our knowledge of the atmospheric boundary layer over the Antarctic sea ice zone comes from this campaign (Andreas, 1995, 2002; Andreas et al., 2000, 2004, 2005; Andreas & Claffey, 1995; Tastula et al., 2012, 2013). The ISW is the hitherto longest lasting Antarctic drift station of its kind, and only two other ice stations with a duration of at least weeks have been deployed on the Antarctic sea ice: Ice station Polarstern (ISPOL), which lasted for 5 weeks in December January 2004 and 2005 (Bareiss & Görgen, 2008; Vihma et al., 2009), and the Sea Ice Mass Balance in the Antarctic (SIMBA), which took place over 2 weeks in September–October in 2007 (Vancoppenolle et al., 2011). The other data sources that we use in this study are radiosounding data sets originating from two research vessel cruises in the Weddell Sea: the first by RV *Akademik Fedorov* (hereafter *Fedorov*), in February 1992, and the second by RV *Aranda* (Vihma et al., 1997), in January to mid-February 1996. All data sets except the one from *Fedorov* were withheld from assimilation and are thus independent data for evaluation of reanalyses. This study is made particularly actual by the ongoing Year of Polar Prediction (YOPP), which has a main focus on improving environmental prediction capabilities in polar regions (Goessling et al., 2016; Jung et al., 2016).

2. Data Sets and Methodology

2.1. Observations

We use observational data from three different sources in this study. The first is the ice drift station ISW, from which we use data on cloud fraction, snow surface skin temperature (the sea ice surface was covered by snow), radiative and turbulent surface fluxes, and near-surface air temperature, air humidity, and wind speed. Near-surface temperature and humidity were observed at heights of 0.1 and 5 m above the surface level (ASL), while wind speed was observed at 5 m ASL. The turbulent sensible and latent heat fluxes were obtained using a sonic anemometer/thermometer and a hygrometer, both mounted at 4.65 m ASL. All these data are available as hourly averages from the ISW data set, and we extracted these data for the time period between 25 February 18:00 and 29 May 1992 18:00. More details about these observations and the postprocessing can be found in Andreas et al. (2004) and Andreas et al. (2005).

In addition, 40 airsonde soundings and 128 tethersondes soundings are available from the ISW (Claffey et al., 1994), the first of which we found to provide reliable data only of temperature and the latter of which

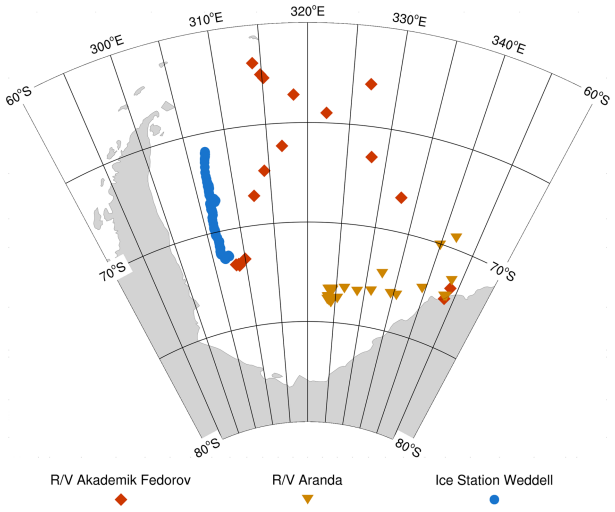

Figure 1. Tracks of *R/V Akademik Fedorov* (red rectangles and red, solid line) and *Ice Station Weddell* (blue circles and blue, solid line) from 1992 and of *R/V Aranda* (yellow triangles and yellow, solid line) from 1996.

provided reliable temperature, humidity, and wind data. For the tethersonde data, the maximum height reached by each sounding varied substantially (from 92 to 1,350 m ASL), and the average height was 613 m ASL. In this study we only use soundings that reached at least 600 m ASL. Some time periods during ISW had more intensive sounding activity than others. In order not to weigh these periods excessively in the error statistics calculations, we selected profiles with a minimum temporal spacing of 12 hr for comparison against the reanalyses. In total we utilize 40 and 56 individual soundings from the ISW airsonde and tethersonde data sets, respectively.

Two other upper-air data sets that we use in this study consist of radiosonde soundings launched from cruises in the Weddell Sea with *Fedorov* and *Aranda*. A total of 77 soundings are available from *Aranda* and 40 from *Fedorov*. Of these, we excluded 17 from further analysis from *Aranda* and 11 from *Fedorov* due to poor data quality. Furthermore, like with the ISW soundings, we selected soundings with a minimum time difference of 12 hr, leading to 24 and 34 available soundings from *Aranda* and *Fedorov*, respectively. The airsonde, tethersonde, and radiosonde soundings all have a fairly similar vertical resolution averaging to about 20 m. Figure 1 presents a map of the locations where the data sets used in this study originate from.

Among all these observational data sets, the *Fedorov* data set is the only one sent to the Global Telecommunication System (GTS) and thereby the only data set made available for assimilation for the reanalysis products considered herein. It is not clear, however, to what degree these observations were actually used for assimilation in the different products and or how much they were weighted in any assimilation. As part of our data postprocessing, we removed clearly erroneous values from the observations before we used them for model evaluation.

2.2. Reanalyses

A wide range of observations have been assimilated in the reanalyses validated in this study, either using three-dimensional variational data assimilation (3D-Var; CFSR and MERRA-2) or 4D-Var (ERA-Interim and JRA-55). The reanalyses are available at horizontal resolutions in the range of 0.5–0.75° in a regular latitude-longitude grid (ERA-Interim, CFSR, and MERRA-2) and 1.25° (JRA-55). We note that JRA-55 is run at a higher native resolution (about 0.5°), but the pressure level data that we use in this study are only available at this coarser resolution. The pressure levels used in this study are located between 1,000 and 250 hPa in all four reanalyses, with a vertical resolution of 25 up to 750 hPa and 50 hPa farther aloft. MERRA-2 has an additional level at 725 hPa. The considered reanalysis variables are from the analysis fields. An exception is the near-surface temperature, humidity, and wind in CFSR, which are only available as forecast fields. Another exception is the radiative and turbulent surface heat fluxes, which we obtained and or calculated as 6-hourly averages accumulated over the model forecasts.

2.3. Methodology

For comparison of the near-surface data from ISW, we extracted 6-hourly data from the hourly observations in order to match the time resolution of the reanalysis output. In the case of the temperature, humidity, wind speed, and cloud fraction, we did this by extracting every sixth data point from the respective time series. For the radiative and turbulent surface heat fluxes, however, we averaged the hourly observations into 6-hourly periods. If more than two values were missing within a 6-hourly period, the whole period was flagged as missing. In the resulting observational 6-hourly data sets, about half of the data are missing in the case of the turbulent surface heat fluxes. For the surface radiative data, however, there are only 2% missing, and for the cloud fraction data about 5% is missing. For the temperature, humidity, and wind data about 15% are missing. We ignored all these data gaps in our error statistics calculations.

The observed cloud fraction was reported on a scale from 0 to 10 with intervals of 1, which we converted to % with intervals of 10% for comparison with the reanalyses, whose cloud fractions we also rounded off to the

nearest 10%. We calculated 2 m values for the observed temperature and humidity and 10 m values for the wind speed using an iterative algorithm provided by Launiainen and Vihma (1990). Surface pressure, which is essential for calculating the 2 m values, is not available in the ISW data set, and we therefore estimated it using the average of surface pressure in all four reanalysis products.

For comparison of the sounding data, we first linearly interpolated both the reanalysis and observation profile data to a common, vertical grid with 100 m intervals between 200 and 4,000 m ASL. Thereafter, we linearly interpolated the reanalyses horizontally and in time to each location and timestamp of the observations. Naturally, in some cases the lowest pressure level (1,000 hPa), and even the second lowest (975 hPa), in each reanalysis profile considered are located below the surface. To minimize the influence of this problem on the error statistics, we only focus on data down to 200 m ASL.

The error statistics that we apply in this study are bias, root mean square error (RMSE), and the correlation coefficient (r). In addition, for the surface layer and cloud fraction data, we also consider slopes of linear regression lines and ratios of standard deviations (standard deviation of the reanalyses divided by that observed). We estimated the statistical significance at the 95% level for the bias and r using the Student t test. The values of r are largely significant throughout the results, and we will express in the text if the values are not significant.

3. Results

3.1. Comparison of Surface Layer Data

3.1.1. Error Statistics

Figure 2 presents error statistics of the surface layer data of temperature, specific humidity, relative humidity with respect to water and ice, and wind speed from the ISW period. Scatterplots between the observed and modeled variables are shown in Figure 3.

The mean observed 2 m temperature is 253.2 K, and all the reanalyses are biased warm compared to this (Figure 2). MERRA-2 has the highest bias (+2.8 K) and CFSR the lowest (+2.0 K). Furthermore, we see that the largest warm biases are generally found for the lowest observed temperatures in all four reanalyses (Figure 3). Correspondingly, the slope of all four linear regression lines is below 1. CFSR has the least scatter among the products and thus has the most favorable RMSE (4.0 K, same as ERA-Interim) and r (0.87), while JRA-55 has the most scatter and the least favorable RMSE and r (5.2 K and 0.78).

The mean observed skin temperature is 251.1 K and all four reanalyses feature warm biases, which are larger than for the 2 m temperature. JRA-55 has the lowest bias (+3.0 K), whereas MERRA-2 has the highest (+5.7 K). As for the 2 m temperature warm biases, the strongest warm biases in skin temperature are generally found for the lowest observed temperature (Figure 3), and the linear regression line slopes are all below 1. CFSR has the lowest RMSE (5.3 K) and highest r (0.84), while MERRA-2 has the highest RMSE (7.3 K) and JRA-55 the lowest r (0.72).

Considering the 2 m specific humidity, the mean observed value is 0.8 g/kg. All four reanalyses are significantly moister than this, corresponding to the warm biases found in the 2 m temperature. The highest moist bias is found in MERRA-2 at +0.2 g/kg, and all the other products have biases of +0.1 g/kg (Figure 2). A large portion of the humidity values are clustered below 1 g/kg in all four products, and the slopes of the regression lines are all close to one (Figures 2 and 3). Thus, the model moisture biases are not strongly affected by the observed humidity, though CFSR, JRA-55, and MERRA-2 have a slight tendency toward more positive humidity biases for higher observed humidity values. In terms of RMSE, ERA-Interim and CFSR have the lowest values of 0.3 g/kg, and considering r, CFSR has the highest value of 0.91, both ERA-Interim and CFSR having relatively small scatter. JRA-55 has the worst values for both RMSE and r (0.5 g/kg and 0.80), and these are reflected in relatively large scatter.

The mean observed 10 m wind speed is 4.6 m/s. All four reanalyses, except CFSR, are biased high compared to this, and the highest bias of +0.6 m/s is found in MERRA-2 (Figure 2). All these biases are statistically significant. CFSR, on the other hand, has no bias (0.0 m/s). The positive biases are dominated by a large portion of relatively low observed wind speeds, and for higher observed wind speeds, all four products on average underestimate the wind speed (Figure 3). This is reflected by the slopes of the linear regression lines being below 1 for all four reanalyses. ERA-Interim and CFSR have the best RMSE (1.5 m/s), and ERA-Interim

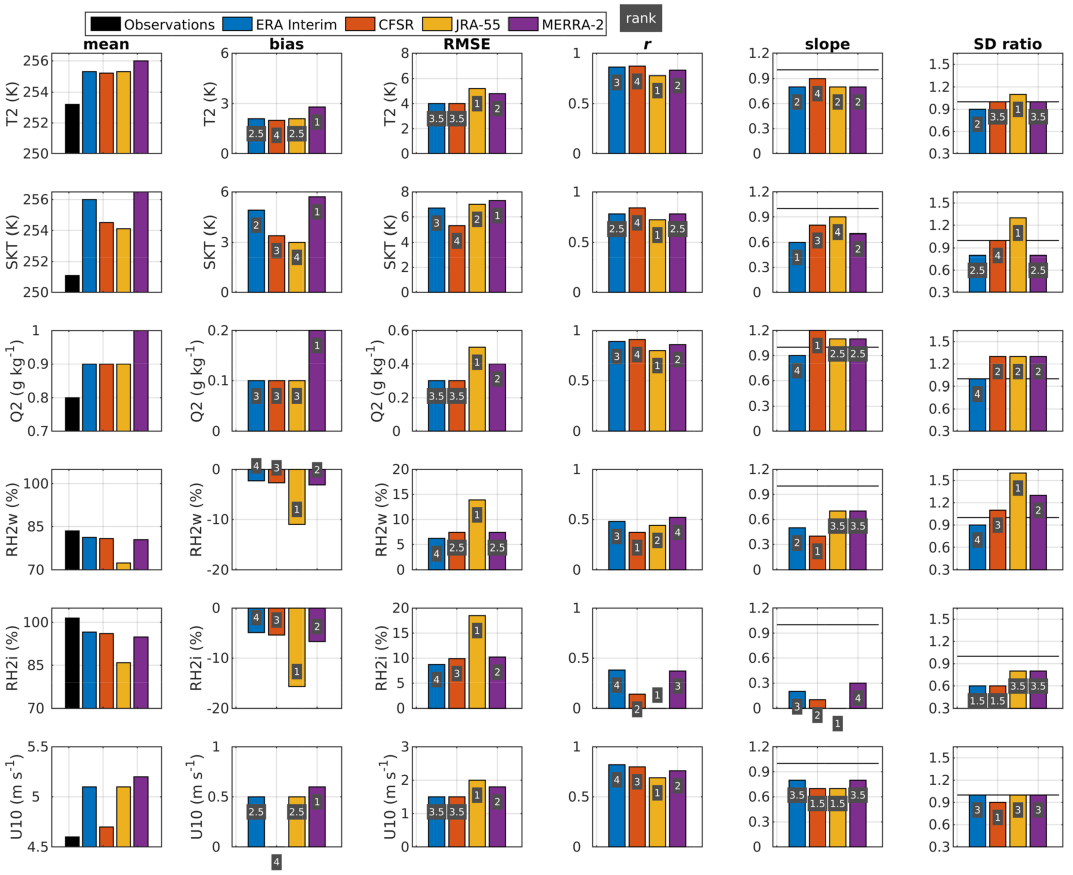

Figure 2. Mean values, mean bias, root mean square error (RMSE), correlation coefficient (*r*), slope of the linear regression line (slope), and the ratio of the standard deviation of the reanalyses divided by that observed (SD ratio) for the four reanalyses using Ice Station Weddell observations as reference. A positive bias indicates that the reanalysis product has a higher value than the observations. The statistics are presented for the 2 m temperature (T2), skin temperature (SKT), 2 m specific humidity (Q2), 2 m relative humidity with respect to water (RH2w), 2 m relative humidity with respect to ice (RH2i), and 10 m wind speed (U10). In the ranking indicated on each bar, the best reanalysis is given 4 points and the worst 1 point.

has the best *r* (0.82), and both products have relatively little scatter. The worst RMSE and *r* are found in JRA-55 (2.0 m/s and 0.69; Figure 2).

3.1.2. Problems With Relative Humidity

The mean observed 2 m relative humidity with respect to water is close to 85%, and all four reanalyses have rather small negative biases within about −2% to −3%. An exception is JRA-55, where it is −11%. The mean observed 2 m relative humidity with respect to ice indicates super saturation (101.1%), while the reanalyses all have mean values below 100%. For JRA-55 the bias is beyond −15%. Similar conclusions as for the biases can be drawn for the values of RMSE, with worse values with respect to ice than for water, and they are particularly bad with respect to ice for JRA-55 with values in the vicinity of 20%. The correlation coefficients, *r*, are around 0.5 for all reanalyses with respect to water. With respect to ice, however, they are considerably worse and vary between 0 (JRA-55, not significant) to around 0.3 (ERA-Interim and CFSR, significant). The scatterplots of observed and modeled 2 m relative humidity (Figure 3) reveal why the values of *r* are this poor. We can clearly see that the reanalyses struggle in reproducing the highest observed values of relative humidity, which with respect to water implies slopes of the linear regression lines that are between 0.4 (CFSR) and 0.7 (JRA-55 and MERRA-2). With respect to ice, this slope is even down to 0 in JRA-55. A closer look at all data points of relative humidity with respect to ice reveals that none of these are above 100% for JRA-55, while the other products do produce some values above 100%.

All four renalyses struggle in reproducing the 2 m relative humidity, in particular with respect to ice (Figures 2 and 3). Based on among other the near-surface observations from ISW, Andreas et al. (2002)

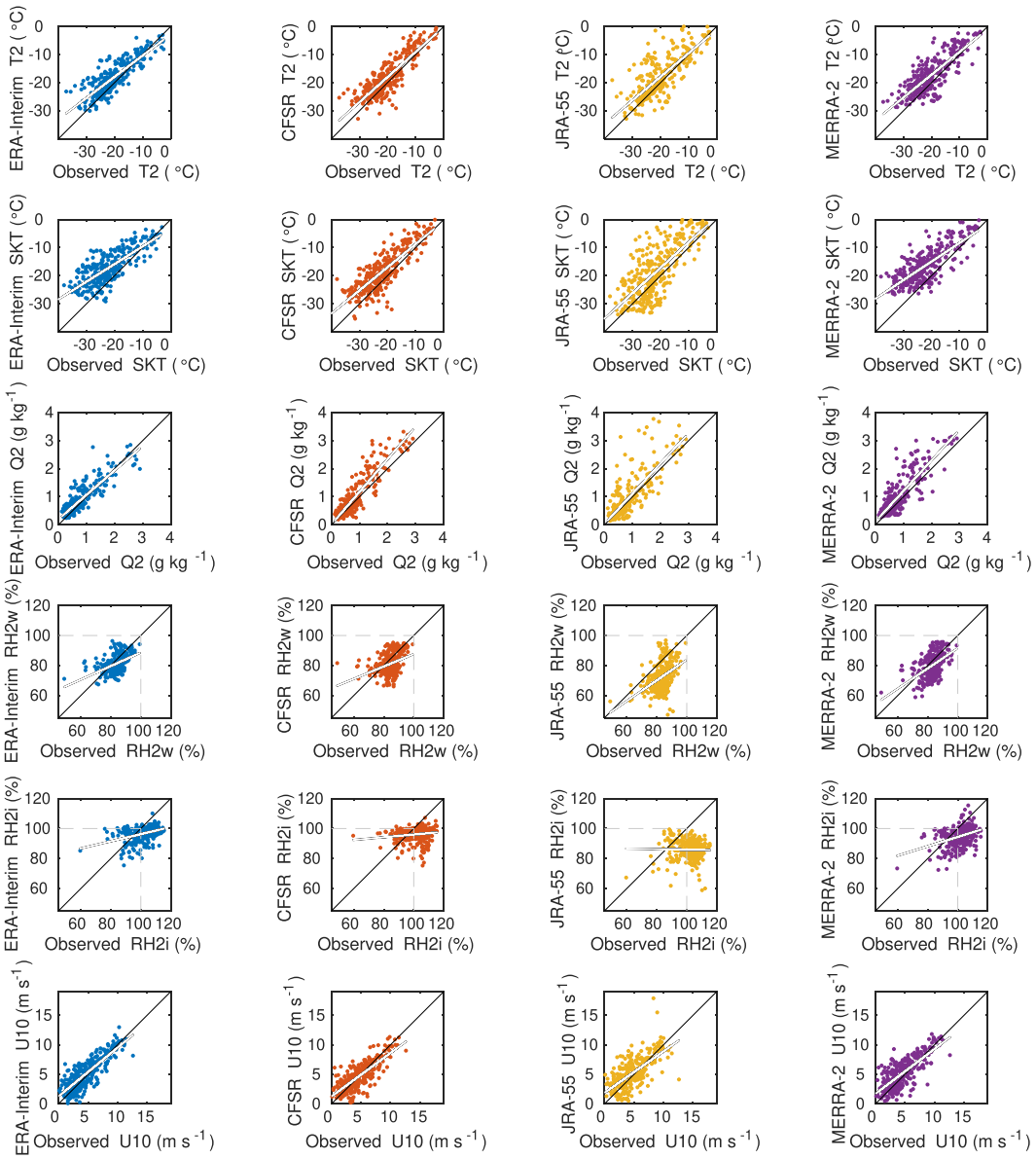

Figure 3. Scatterplots showing the observed 2 m temperature (T2), surface skin temperature (SKT), 2 m specific humidity (Q2), 2 m relative humidity with respect to water (RH2w), 2 m relative humidity with respect to ice (RH2i), and 10 m wind speed (U10) with respect to the same variables in the four reanalysis products.

concluded that water vapor over polar sea ice is nearly always near saturation and often at supersaturation with respect to ice for temperatures between 0 and −25 °C. For temperatures below −25 °C, their results were less robust, due in part to the impaired reliability of the humidity sensors applied for such low temperatures, and this is also a limitation in the ISW observations. We investigate here in further detail how the 2 m relative humidity with respect to ice is reproduced by the reanalyses for temperatures in the range of −35 and 0 °C. Our results (Figure 4) reveal that all four reanalyses do reproduce conditions close to saturation for the entire temperature range, all having relative humidity values of about 95% or higher. JRA-55 is an exception to this, which consistently has values below 95% for any temperature, and for temperatures between −20 and −30 °C, even below 85%. Neither of the products reproduces the observed onset of supersaturation at temperatures below −15 °C, but for temperatures between −35 and −30 °C, CFSR and ERA-Interim do show supersaturation. Though, as we comment on above, and as stated by (Andreas et al., 2002), the observational evidence for supersation based on the ISW data set for temperatures below −25 °C is questionable due to instrument limitations.

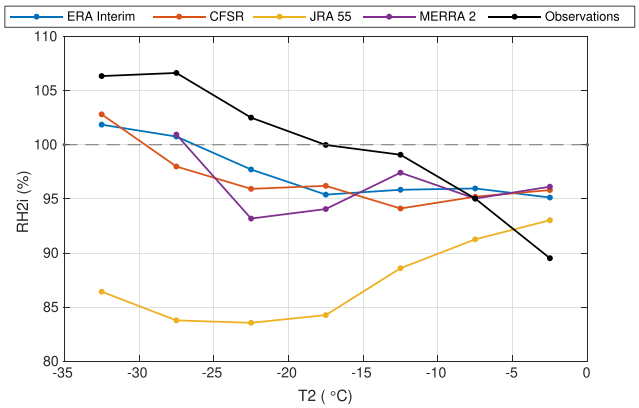

Figure 4. Relative humidity at 2 m with respect to ice (RH2i) from the Ice Station Weddell observations and the four reanalyses averaged in bins of 5 °C based on the observed and modeled 2 m temperature (T2).

3.1.3. Summary of Comparison of Surface Layer Data

To summarize and make an overall assessment of the reanalyses' near-surface performance, we applied a ranking system to the error statistics of temperature, humidity, and wind speed (labels on bars in Figure 2). The ranking gives 4 points to the best product and 1 point to the worst product for each of the error metrics and atmospheric parameters considered. In case of a tied ranking between two products, for example, for the first place, both products receive 3.5 points. When all points are summed up for each reanalysis, ERA-Interim has the highest ranking (91.5 points), closely followed by CFSR (85.5 points). MERRA-2 and JRA-55 receive the fewest points (respectively 69 and 54 points).

3.2. Comparison of Radiative and Turbulent Surface Heat Fluxes and Cloud Fraction

3.2.1. Error Statistics

The error statistics of radiative and turbulent surface heat fluxes and cloud fraction for the ISW period are shown in Figure 5, and scatterplots between the different observed and modeled variables are shown in Figure 6.

Considering the downwelling shortwave radiation, the mean observed value is 50.4 W/m^2, and the reanalyses come rather close to this, and only MERRA-2 and JRA-55 feature significant biases of −10.7 and +4.1 W/m^2, respectively (Figure 5). All four reanalyses have linear regression lines with slopes near 1, except MERRA-2, where it is 0.7. The RMSE is the most favorable in CFSR (22.4 W/m^2) and the worst in MERRA-2 (32.4 W/m^2). The values of *r* are very high across all the four reanalyses, with values between 0.95 (MERRA-2) and 0.97 (CFSR). These high values presumably largely reflect the reanalyses' ability to capture the diurnal cycle in shortwave radiation and its seasonal evolution from early autumn to winter.

As to the downwelling longwave radiation, the mean observed value is 204.5 W/m^2. The largest bias is found in JRA-55 (−13.0 W/m^2) and the smallest in ERA-Interim (−2.7 W/m^2, not significant). MERRA-2 is the only product with a positive bias (+9.6 W/m^2). The RMSE lies approximately within 25 and 40 W/m^2 for all four products, which is similar to the values for the downwelling shortwave radiation, JRA-55 having the least favorable value of 37.7 W/m^2 and the three other products all having values around 28 W/m^2. The values of *r*, on the other hand, are notably lower for the downwelling longwave radiation than for the downwelling shortwave radiation, and they are the worst for JRA-55 and MERRA-2 (about 0.68) and best for ERA-Interim and CFSR (0.75). As indicated by the slopes of the linear regression lines all being below 1, all four reanalyses have a tendency to underestimate higher observed values of the downwelling longwave radiation (Figure 6). This tendency is the strongest in MERRA-2 with a slope of only 0.6.

Regarding the turbulent surface flux of sensible heat, the mean observed value is −2.0 W/m^2. The reanalyses also have negative mean values; however, they are larger, ERA-Interim having the largest negative bias of −4.4 W/m^2. MERRA-2 is an exception, with a positive bias of +13.1 W/m^2. The values of RMSE are approximately between 10 and 25 W/m^2. The slopes of the linear regression lines are all markedly below 1, with the lowest value of only 0.2 in MERRA-2 and the highest in CFSR with 0.7. Correspondingly, the correlation coefficients, *r*, are rather poor with values below 0.5 for all products and MERRA-2 having the worst (0.17), indicating a poor match between the data pairs in the linear sense. This poor match is also evident from looking at the scatterplots in Figure 6.

The mean observed turbulent surface flux of latent heat is only 0.3 W/m^2. The latent heat flux is on average slightly positively biased in all four reanalyses, with values between +0.5 W/m^2 (JRA-55) and +4.7 W/m^2 (MERRA-2). All these biases are significant, except for the bias in JRA-55. The RMSE values are also fairly low, between 3.3 W/m^2 in JRA-55 and 10.0 W/m^2 in CFSR. *r* is rather poor in all reanalyses, with values between 0.29 (CFSR) and 0.53 (ERA-Interim). The slopes of the linear regression lines are both above and below 1 in the respective products, with the strongest deviations from 1 found in ERA-Interim (1.2) and JRA-55 (0.7). All products feature a stronger variability in the latent heat flux than the observed one. This

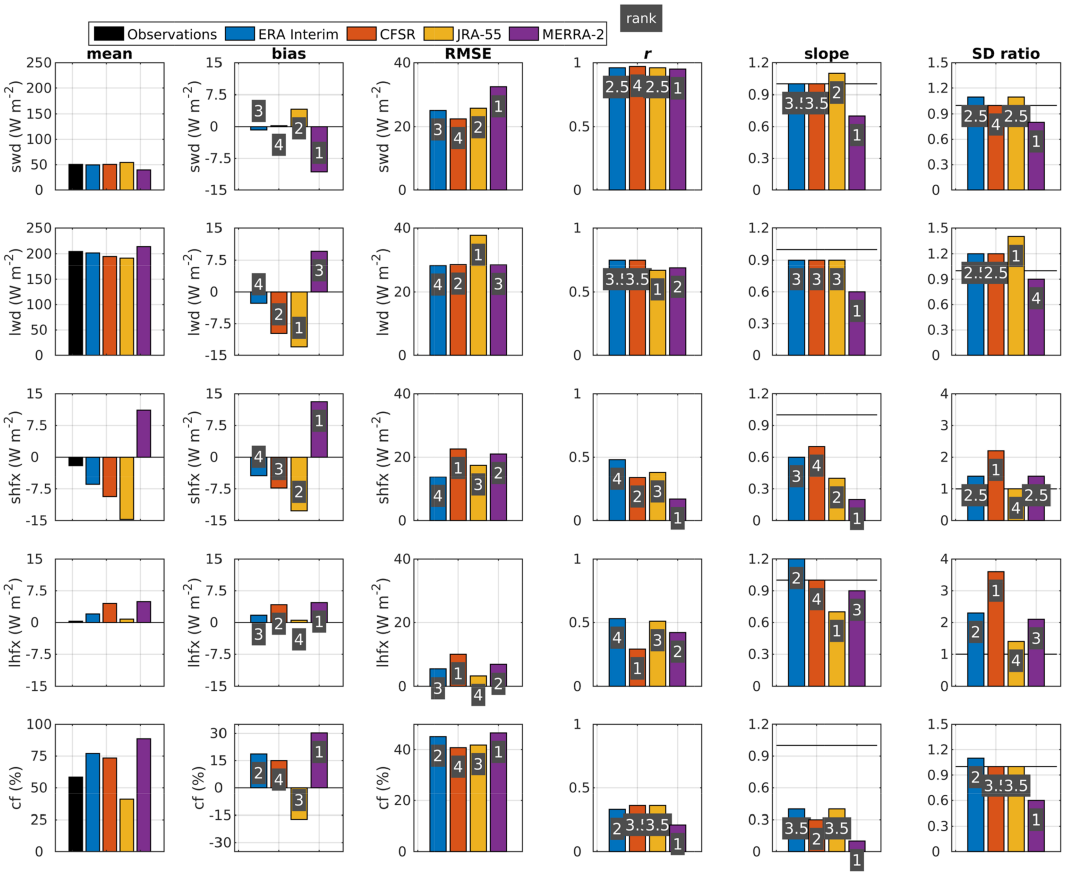

Figure 5. Mean values, mean bias, root mean square error (RMSE), correlation coefficient (*r*), slope of the linear regression line (slope), and the ratio of the standard deviation of the reanalyses divided by that observed (SD ratio) for the four reanalyses using Ice Station Weddell observations as reference. A positive bias indicates that the reanalysis product has a higher value than the observations. The statistics are presented for the downwelling shortwave radiation (swd), downwelling longwave radiation (lwd), turbulent sensible heat flux (shfx), turbulent latent heat flux (lhfx), and cloud fraction (cf). In the ranking indicated on each bar the best reanalysis is given 4 points and the worst 1 point.

is seen in the standard deviation ratios, which are all positive, ranging from 1.4 in JRA-55 all the way up to 3.6 in CFSR.

Considering the cloud fraction, we see that the mean observed value is 58.4%. ERA-Interim, CFSR, and MERRA-2 all have positive biases, with the latter having the largest of +30.2%. JRA-55, on the other hand, has a negative bias of −17.3%. The RMSE is fairly large in all products, with values between 40.8% (CFSR) and 46.6% (MERRA-2). The mean *r* for all reanalyses (0.32) is the lowest for all variables considered, and for MERRA-2 it is particularly low (0.21). The relatively high RMSE and low *r* values are reflected in the scatterplots (Figure 6), which show a wide spread between the observed and simulated cloud fraction.

3.2.2. Summary of Comparison of Radiative and Turbulent Surface Heat Flux and Cloud Fraction

We apply the same ranking system for the radiative and turbulent surface heat fluxes and cloud fraction as for the near-surface temperature, humidity, and wind (see section 3.1.3). When all points are summed up for each reanalysis, ERA-Interim has the highest ranking (74.5 points), followed by CFSR (69.5 points), JRA-55 (64.5 points), and last MERRA-2 (41.5 points). We note that *r* and the slope of the regression line are especially poor for the cloud fraction, confirming that caution is needed when applying reanalysis products for cloud fraction.

3.3. Comparison With Upper-Air Observations

In the following, we evaluate the reanalyses' upper-air performance between 200 and 4,000 m ASL with respect to potential temperature, specific humidity, relative humidity, and wind speed. We pay particular attention to differences in performance with respect to height ASL, geographical location, and data set

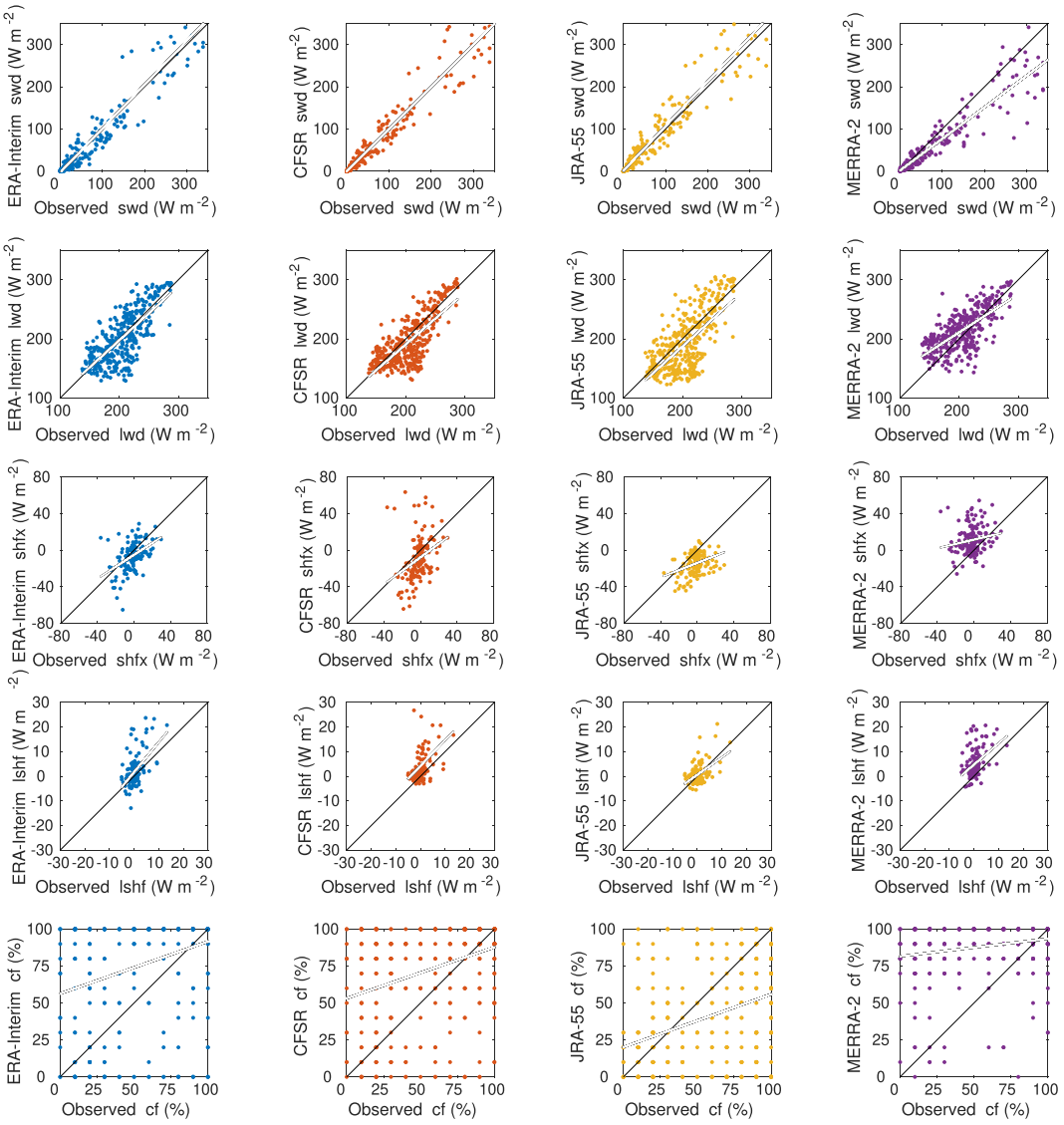

Figure 6. Scatterplots showing the observed downwelling shortwave radiation (swd), downwelling longwave radiation (lwd), turbulent sensible surface heat flux (shfx), turbulent latent surface heat flux (lhfx) and cloud fraction (cf) with respect to the same variables in the four reanalysis products.

independence. Regarding locations, we note that the ISW airsonde and tethersonde soundings were obtained solely over the Antarctic sea ice pack in autumn and winter, whereas the *Aranda* and *Fedorov* soundings were ship-based and were thus made partly over the open ocean and partly over the sea ice (Figure 1) in summer and autumn. The error statistics results for the *Aranda* and *Fedorov* data sets, and the ISW airsondes and ISW tethersondes data sets are presented in Figures 7 and 8, respectively.

3.3.1. Temperature

The mean potential temperature profiles from the ship-based *Aranda* and *Fedorov* data sets (Figure 7) closely resemble each other, both in terms of their values at the lowest levels (about 270 K at 200 m ASL) and their shape (variation with height), although the observations cover different years. The mean ISW profiles (Figure 8), all observed over a compact sea ice field, deviate from these by being substantially colder (~255 K at 200 m ASL) and more statically stably stratified in the lower hundreds of meters (~14 K/km vs. ~7 K/km). The mean reanalyses profiles capture the general shapes of these observed profiles quite well, though there are some biases in the reanalyses. When compared against the *Aranda* and *Fedorov* data sets, the sign of the reanalysis biases depends on the altitude. In the lowest few hundred meters, there are warm biases of +0.1 to 1.4 K across all the four reanalyses. Compared against the *Aranda* data set, these warm

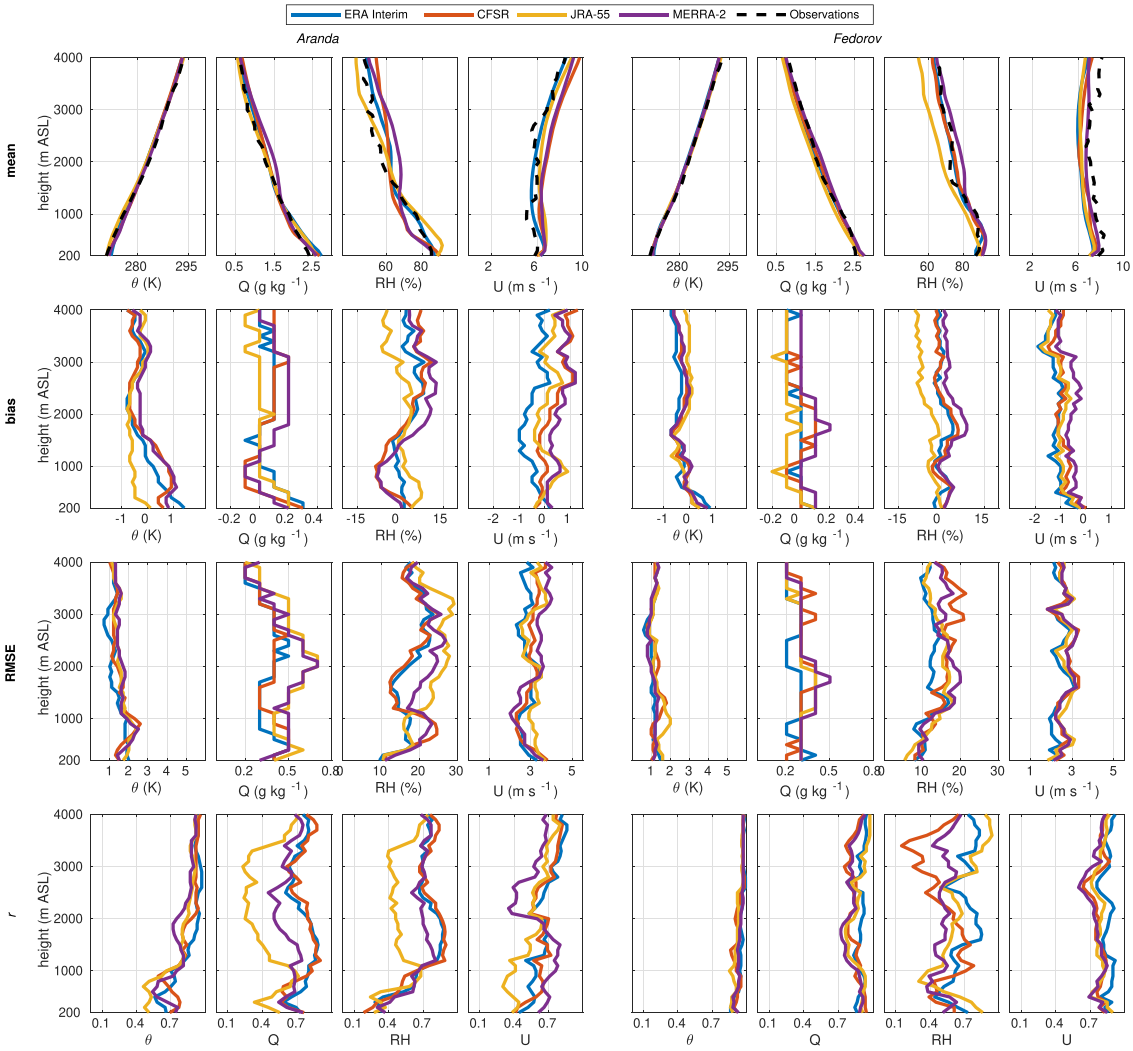

Figure 7. Profiles of mean, bias, root mean square error (RMSE), and correlation coefficient (*r*) for potential temperature (θ), specific humidity (Q) and wind speed (U) for the reanalyses using the *Aranda* and *Fedorov* profile data sets as reference.

biases are significant in all four reanalyses, except JRA-55, and they cover a relatively deep layer of up to about 1,500 m ASL. Compared against the *Fedorov* data set, the low-level warm biases are generally smaller and occur in a shallower layer, and none of them are significant. Farther aloft, cold biases down to about −1 K dominate in both data sets. The spread in the biases between the reanalyses is generally larger for the independent *Aranda* and ISW sounding data sets than for the *Fedorov* data set, especially at the lower levels. In the ISW soundings, the biases also differ in sign from the other two data sets; while there on average is a warm bias in ERA-Interim and JRA-55 in the lower hundreds of meters, there is a cold bias in MERRA-2 and CFSR for the same layer. We note that only the cold biases are statistically significant.

The highest RMSE for all four reanalyses is found in the lowest few hundred meters when compared against the independent *Aranda* and ISW airsonde data sets. The latter data set reveals the highest RMSE at 200–300 m ASL, with the best RMSE found in CFSR (3.8 K) and the worst in JRA-55 (5.5 K). The *Fedorov* data set does not reveal a clear maximum in the RMSE of the reanalyses at the lower levels, and the spread in RMSE is lower than in the independent data sets.

The least favorable *r* is found for the *Aranda* data set between 200 and 1,000 m ASL, and JRA-55 has the lowest *r* of about 0.45. Farther aloft, *r* is around or higher than 0.7 in all reanalyses and data sets investigated.

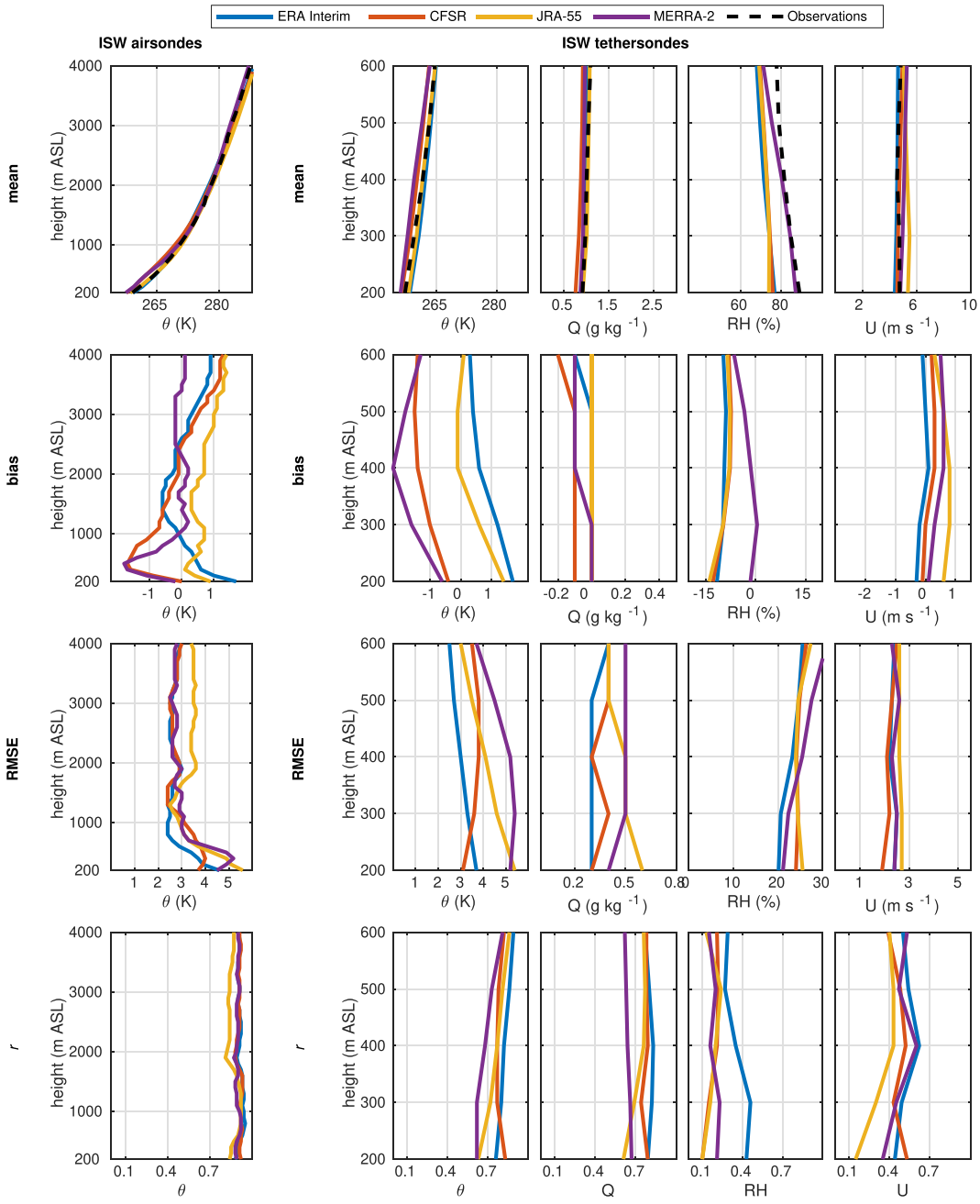

Figure 8. Same as in Figure 7 but for the ISW airsondes and ISW tethersondes data sets.

The overall highest values of *r* and the smallest spread between the reanalyses are found in *Fedorov* and the ISW airsondes.

3.3.2. Specific Humidity

The mean specific humidity profiles in the *Aranda* and *Fedorov* data sets have similar shapes and values, ranging from about 2.5 g/kg or slightly above at the lowest levels down to about 0.5 g/kg at the highest levels. In comparison, the ISW tethersonde profiles are substantially drier with values between 0.6 and 1 g/kg up to 600 m ASL, which corresponds to the lower temperatures in these profiles. The largest bias is found for ERA-Interim and CFSR in the *Aranda* data set at 200 m ASL (+0.3 g/kg). Farther aloft, MERRA-2 has the highest positive bias with values up to +0.2 g/kg between 2,000 and 3,000 m ASL. Both the spread in bias across the reanalyses and their absolute values are lower in the *Fedorov* data set than in the *Aranda* data set. Also, the

biases are not statistically significant for the *Fedorov* data set, except in JRA-55, which has a significant dry bias at most height levels peaking at −0.2 g/kg at 900 and 3,100 m ASL. The ISW tethersonde data set shows small biases in the reanalyses, and only for CFSR it has a significant value at all height levels of down to −0.2 g/kg.

Both the largest RMSE (about 0.7 g/kg in JRA-55 and MERRA-2) and largest spread in RMSE between reanalyses are found at 2,000 m ASL in the Aranda reanalyses profiles. This is close to where the highest RMSE is found for the *Fedorov* data set too, but the absolute values and their variation with height are lower in all four reanalyses for *Fedorov*. In the ISW tethersonde data set, there is little variation with height in the RMSE across the reanalyses, and ERA-Interim has the lowest values (below 0.3 g/kg for most height levels).

The reanalyses' correlation coefficient, *r*, features large variability with height for the *Aranda* data set, in particular in JRA-55, ranging from only 0.25 (not significant) between 2,000 and 3,000 m ASL to 0.7 (significant) at 4,000 m ASL. *r* is generally much higher (all above 0.7) for the *Fedorov* than for the *Aranda* data set, and the spread in *r* between the reanalyses is lower. Also, in the ISW tethersonde data set the spread in *r* is fairly low between the reanalyses and the values are all 0.65 or above.

3.3.3. Relative Humidity (With Respect to Water)

The mean relative humidity profiles in the *Aranda* and *Fedorov* data sets feature the same variability with height as the specific humidity; that is, the highest values are found at the lowermost levels (about 85%) and the lowest values at the uppermost levels (about 55%). The ISW tethersonde profiles also have values around 85% for the lower altitudes covered by that data set (up to 600 m ASL). Looking at the reanalyses biases, these are the highest for the *Aranda* data set between 2,000 and 3,000 m, MERRA-2 having highest values at close to +15% between 2,000 and 3,000 m ASL. A similar, but somewhat smaller, bias is seen in MERRA-2 for the *Fedorov* data set. The reanalyses display negative biases throughout with values down to almost −15% at 200 m ASL for the ISW tethersonde data set. An exception is MERRA-2, which has a bias very close to 0% for most altitudes.

As is generally the case for the other variables investigated, both the largest RMSE (up to 30% in JRA-55 at 3–4,000 m ASL), and the largest spread in RMSE between the reanalyses are found when compared with the *Aranda* data set. The ISW tethersonde data set displays little variation with height in the RMSE across the reanalyses and ERA-Interim has the lowest values (below 25% for most height levels).

The correlation coefficients, *r*, feature a very similar variability with height to the ones for specific humidity in the *Aranda* data set. For the *Fedorov* data set, however, there is a relatively much stronger variability in *r* with height for relative than for specific humidity, CFSR having the lowest values down to almost 0.1 at about 3,500 m ASL. For the ISW tethersondes, the values of *r* are fairly poor with none of the reanalyses having values above 0.5 for any height level.

3.3.4. Wind Speed

While the mean wind speed generally increases with height in the *Aranda* data set, there is an overall, slight decrease with height in the *Fedorov* profiles. This variation with height is seen by the reanalyses, but they do feature biases. We note that none of the biases in the *Aranda* data set are significant. For the *Fedorov* data set, however, ERA-Interim features significantly negative biases, peaking at about −1.9 m/s at 3,300 m ASL. The absolute values and variability in the wind speed bias across the reanalyses are larger in the *Aranda* data set than in the *Fedorov* data set. The mean ISW tethersonde reanalyses profiles follow very closely the observed wind, and the agreement is particularly good for CFSR.

The RMSE is roughly between 2 and 4 m/s in all three wind speed data sets. ERA-Interim has marginally the lowest RMSE values for the *Aranda* and *Fedorov* data sets, and the spread in RMSE between the reanalyses is somewhat smaller for the *Fedorov* than for the other data sets.

The highest *r* (at least 0.6) across all products and height levels are found for the *Fedorov* data set. The reanalyses also feature the smallest variation in *r* for this data set. The lowest, and also insignificant value of *r* (only 0.1) for any height level, is found at 200 m ASL in JRA-55 when compared against the ISW tethersonde profiles.

3.3.5. Influence of Temperature Inversions on the Skill Scores

We have calculated separate error statistics for the lowest 1,500 m ASL for two subsets of profiles, with and without inversions in the lowermost 1,500 m ASL. Following R. W. Jones et al. (2016a), we define an inversion as an atmospheric layer in which the temperature increases with height by 2 K or more. We only use the

Aranda and ISW airsonde data sets in the calculations of these statistics because of specific characteristics of the other data sets, outlined in the following. The *Fedorov* data set is not independent, and, compared to this data set, the reanalyses do not have any significant temperature biases nor do they have substantially increased temperature RMSE in the lower layers as the *Aranda* and ISW airsonde reanalyses profiles do. The ISW tethersonde data set has a limited vertical range (up to 600 m ASL) and is therefore also excluded from this analysis.

In the *Aranda* data set, we identified at least one temperature inversion in 16 out of 34 profiles, and in the ISW airsonde data the corresponding numbers were 32 out of 40 profiles. Among the latter, the temperature data were of poor quality in the lower 1,500 m ASL in one profile, and we therefore set the resulting number of ISW airsonde profiles containing no inversion to 7. The results from the *Aranda* reanalysis profile error statistics (see Table S1 in supporting information) reveal that when temperature inversions are present, the mean potential temperature bias (mean of all reanalyses biases) in the lowermost 1,500 m layer is +0.5 K, and for profiles without any inversion, it is +0.1 K. For mean RMSE the values are 2.1 K (inversion) and 1.5 K (no inversion), and for the mean r they are 0.59 (inversion) and 0.78 (no inversion). Thus, there is on average a consistent degradation in model performance for potential temperature when the profiles include inversions, and this result is also largely valid for when the statistics for each individual reanalysis are considered. The increased positive temperature biases with inversions present alludes to a general underestimation of the inversion strength. Furthermore, we also note that the error statistics for specific humidity are also generally worse for profiles with than without temperature inversions for the *Aranda* data set and the same is seen for relative humidity.

For the ISW airsonde data set, the situation is somewhat different with respect to the potential temperature profile error statistics (see Table S2); while the mean RMSE for all reanalyses degrades from the noninversion (1.7 K) to the inversion profiles (3.6 K), the bias and r improve from +1.1 to −0.5 K and 0.88 to 0.90, respectively. For ERA-Interim, the bias is reduced from 0.7 to 0.0 K when going from noninversion to inversion profiles, and the corresponding change for JRA-55 is from +1.3 to +0.3 K. For CFSR and MERRA-2, on the other hand, there is a change in sign in the biases, and they go from +0.7 to −1.4 K and +1.7 to −1.0 K, respectively, when the noninversion profiles are compared with the inversion profiles. The above suggests a general overestimation of the inversion strengths by CFSR and MERRA-2, and an underestimation by ERA-Interim and JRA-55.

3.3.6. Summary of the Reanalyses' Upper Air Performance

In order to summarize the upper-air performance of the four reanalyses, we have calculated vertically averaged error statistics for each of the atmospheric parameters considered and applied the same ranking system as for the near-surface data. The results from the *Aranda* and ISW airsonde sounding data sets are presented in Tables S1 and S2, and the results from the *Fedorov* and ISW tethersonde sounding data sets can be found in Tables S3 and S4.

The results of the ranking reveal that ERA-Interim has the highest ranking scores, except for the ISW airsonde data set where CFSR has marginally the highest score. JRA-55 and MERRA-2 have either the lowest or the second lowest ranking scores for all four data sets. An exception is the *Fedorov* data set where MERRA-2 has the second highest score. In terms of the vertically averaged values of RMSE and r, ERA-Interim is the best, with only a few exceptions, for all data sets and all parameters considered. Considering biases, however, ERA-Interim has the worst or second-worst values for several data sets and parameters, such as for the potential temperature and wind speed for the *Aranda* and *Fedorov* data sets.

4. Discussion and Conclusions

In this study, we used surface layer and upper-air data from two research cruises and an ice station in the Weddell Sea from 1992 and 1996 to evaluate four current reanalyses: ERA-Interim, CFSR, JRA-55, and MERRA-2.

In terms of surface layer performance, we find that CFSR and ERA-Interim generally perform the best when the ranking scores are summed, and ERA-Interim has the overall highest scores. We note that CFSR performs particularly well for the 2 m temperature, for which it has the best bias, RMSE, and r, and it also performs well for 10 m wind speed, for which it is the only product with a nonsignificant bias. We do not know why CFSR performs this well for these variables, but note that among the reanalyses evaluated here, CFSR is

the only product based on a coupled atmosphere - sea ice - ocean model (Saha et al., 2010). This is expected to yield advantage close to the air - sea and air- ice interphases. ERA-Interim, on the other hand, performs very well when it comes to the relative humidity, both with respect to water and with respect to ice. It is also the product that in general comes the closest to reproducing the observed saturation and supersaturation with respect to ice.

All four reanalyses feature warm biases in their 2 m temperature, ranging from +2.0 K in CFSR to +2.8 K in MERRA-2. There are few other evaluation studies of reanalyses covering the Antarctic sea ice zone, largely because few observations exist, and those that do cover rather short time periods. This is particularly true for the current reanalyses as evaluated herein. A study that does evaluate current reanalyses over the Antarctic sea ice is that by R. W. Jones et al. (2016b). Based on data from three research vessel cruises, they documented near-surface cold biases in ERA-Interim, CFSR, JRA-55, and MERRA in the Amundsen Sea Embayment in West Antarctica, though these biases were dominated by strong negative values close to the coast (approaching -6 K). Farther offshore, they found data points with weaker and even positive temperature biases in all four reanalyses. Regarding older reanalyses, the NCEP-National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al., 1996) has been found to have a cold bias by Vihma (2002) and Vancoppenolle et al. (2011), who used buoy data and data from SIMBA and ISPOL in their evaluations. Using ERA-Interim data from 1979 to 2013, P. D. Jones and Lister (2014) also found cold biases along the coast, whereas they also documented prominent warm biases in the Antarctic interior. Using satellite data, Fréville et al. (2014) also found widespread warm biases in the Antarctic interior.

Reliable near-surface variables in reanalyses are important for climate research, where variables such as 2 m air temperature and 10 m wind speed often receive much attention, and for usage of reanalyses in driving ocean and sea-ice models. For example, too high near-surface wind speeds, as we document for ERA-Interim, JRA-55, and MERRA-2, would lead to overestimation of the wind stress and its curl, with implications to, for example, ocean dynamics and transport, and sea-ice drift (Saha et al., 2010; Uotila et al., 2014).

Considering the error statistics for surface heat flux and cloud fraction, the relatively largest model errors are found in MERRA-2 and JRA-55 for the latter variable, with biases of respectively +30.2% and −17.3%. Previous studies have found connections between biases in longwave and shortwave radiation and biases in cloud fractions (Walsh et al., 2009; Zib et al., 2012). Such an investigation is, however, beyond the scope of our study. Radiative fluxes are indeed not simply explained by the cloud fraction, and the cloud liquid water, ice, and aerosol contents are more important for the radiative transfer (Vancoppenolle et al., 2011; Walsh et al., 2009; Zib et al., 2012).

Regarding the upper-air performance, ERA-Interim outperforms all the other reanalyses when the rank scores are summed up for all data sets, except for the ISW airsondes where CFSR has the marginally highest score. However, ERA-Interim does suffer from some prominent biases, including a significant warm bias of up to +1.4 K compared with the *Aranda* soundings. This low-level warm bias in ERA-Interim is consistent with findings in more recent data from the eastern side of the Antarctic peninsula (Nygård et al., 2016) and also with data from the Arctic (de Boer et al., 2014; Jakobson et al., 2012; Liu et al., 2008; Lüpkes et al., 2010; Wesslén et al., 2014). Corresponding to this warm bias, there is a significant moist bias in the lowermost layers of the reanalyses, when compared against the *Aranda* soundings and, to a lesser degree, the *Fedorov* soundings. At higher levels (above 1,500 m ASL), all four reanalyses have significant cold biases compared with both *Fedorov* and *Aranda* profiles. Similar cold biases were found by Nygård et al. (2016) for reanalysis data from the eastern side of the Antarctic Peninsula. They pointed out that such cold biases are consistent with biases found in satellite data (Boylan et al., 2015; J. Wang et al., 2013) and reasoned that this might be a source for these cold biases. However, none of the satellite data sets addressed in those studies were available for 1992 and 1996.

Regarding the data independency, the *Fedorov* data set was the only data set made available for assimilation through the GTS. Assuming that the *Fedorov* data were used in data assimilation, this is probably a major reason why the error statistics are generally more favorable and there is mostly a smaller spread between the reanalyses for this data set than for the other upper-air data sets. Still, significant biases remain, as, for example, upper-level cold biases down to −0.7 K in ERA-Interim. Hence, there are indications that the data assimilation system does not fully utilize available soundings, for example, by giving them too little weight. This is consistent with previous studies, for example, by de Boer et al. (2014), Jakobson et al. (2012), Liu et al.

(2008), Lüpkes et al. (2010), and Wesslén et al. (2014), who found biases similar in nature to these data in spite of the fact that the radiosonde data used for validation were sent to GTS and were thus available for assimilation.

Considering spatial variability in reanalysis performance, previous validation studies for the Antarctic have revealed substantial differences between different products (R. W. Jones et al., 2016b; Nygård et al., 2016). In our study, we identified strong spatial differences with respect to lower-level potential temperature. Under the very stable stratification in the lower layers of the ISW soundings over sea ice, JRA-55 and ERA-Interim both feature warm biases (though not significant), just like in the ship-based *Aranda* and *Fedorov* soundings under less stable stratification. CFSR and MERRA-2, however, feature significant cold biases in the lower levels of the ISW soundings. Temperature inversions are known to be difficult to represent appropriately in models, and several previous studies have found larger errors in conditions of strong inversions (Harden et al., 2011; R. W. Jones et al., 2016b; Lüpkes et al., 2010; Pavelsky et al., 2010). It is therefore not surprising that such differences may occur. We do see that the reanalyses error statistics for the lower 1,500 m ASL degrade when going from the noninversion profiles to profiles containing inversions for the *Aranda* data set, and to some degree also for the ISW airsonde data set.

Finally, it is worth noting that even though the observations considered in this study are older than in several recent evaluation studies (R. W. Jones et al., 2016b; Lüpkes et al., 2010; Nygård et al., 2016), the documented biases are largely similar in nature, including dominant warm biases in the ABL. As the amount of observations has varied depending on the decade, season, and region, the consistency of the warm bias in the ABL suggests a need to improve ABL and surface energy budget parameterizations. In addition to reanalyses, a warm near-surface bias in conditions of a stable boundary layer is a common feature in numerical weather prediction and often attributed to excessive heat and momentum fluxes in the stable boundary layer (Cuxart et al., 2006; Vihma et al., 2014).

Acknowledgments
We express our deep gratitude for the work of the late Edgar L. Andreas in Ice Station Weddell. This study was supported by the Academy of Finland (Contract 304345, ASPIRE project) and the EC Marie Curie Support Action LAWINE (Grant 707262). A. P. Makshtas was supported by the Ministry of Education and Science of the Russian Federation (Project RFMEFI61617X0076). The observational data sets used in this study are available at the websites indicated in the supporting information. This is a contribution to the Year of Polar Prediction (YOPP), a flagship activity of the Polar Prediction Project (PPP), initiated by the World Weather Research Programme (WWRP) of the World Meteorological Organisation (WMO).

References

Andreas, E. L. (1995). Air-ice drag coefficients in the western Weddell Sea: 2. A model based on form drag and drifting snow. *Journal of Geophysical Research*, *100*(C3), 4833.

Andreas, E. L. (2002). Parameterizing scalar transfer over snow and ice: A review. *Journal of Hydrometeorology*, *3*(4), 417–432.

Andreas, E. L., & Claffey, K. J. (1995). Air-ice drag coefficients in the western Weddell Sea: 1. Values deduced from profile measurements. *Journal of Geophysical Research*, *100*(C3), 4821.

Andreas, E. L., Claffy, K. J., & Makshtas, A. P. (2000). Low-level atmospheric jets and inversions over the Western Weddell Sea. *Boundary-Layer Meteorology*, *97*(3), 459–486.

Andreas, E. L., Guest, S. P., Persson, P. O. G., Fairall, C. W., Horst, T. W., Moritz, R. E., & Semmer, S. R. (2002). Near-surface water vapor over polar sea ice is always near ice saturation. *Journal of Geophysical Research*, *107*(C10), 8033. https://doi.org/10.1029/2000JC000411

Andreas, E. L., Jordan, R. E., & Makshtas, A. P. (2004). Simulations of snow, ice, and near-surface atmospheric processes on Ice Station Weddell. *Journal of Hydrometeorology*, *5*(4), 611–624.

Andreas, E. L., Jordan, R. E., & Makshtas, A. P. (2005). Parameterizing turbulent exchange over sea ice: The ice station weddell results. *Boundary-Layer Meteorology*, *114*(2), 439–460.

Assmann, K. M., Jenkins, A., Shoosmith, D. R., Walker, D. P., Jacobs, S. S., & Nicholls, K. W. (2013). Variability of circumpolar deep water transport onto the Amundsen Sea Continental shelf through a shelf break trough. *Journal of Geophysical Research: Oceans*, *118*, 6603–6620. https://doi.org/10.1002/2013JC008871

Bareiss, J., & Görgen, K. (2008). ISPOL weather conditions in the context of long-term climate variability in the north-western Weddell Sea. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *55*(8-9), 918–932.

Boylan, P., Wang, J., Cohn, S. A., Fetzer, E., Maddy, E. S., & Wong, S. (2015). Validation of AIRS version 6 temperature profiles and surface-based inversions over Antarctica using Concordiasi dropsonde data: AIRS v6 Antarctic Surface Inversions. *Journal of Geophysical Research: Atmospheres*, *120*, 992–1007. https://doi.org/10.1002/2014JD022551

Bracegirdle, T. J., & Marshall, G. J. (2012). The reliability of antarctic tropospheric pressure and temperature in the latest global reanalyses. *Journal of Climate*, *25*(20), 7138–7146.

Bromwich, D. H., Fogt, R. L., Hodges, K. I., & Walsh, J. E. (2007). A tropospheric assessment of the ERA-40, NCEP, and JRA-25 global reanalyses in the polar regions. *Journal of Geophysical Research*, *112*, D10111. https://doi.org/10.1029/2006JD007859

Bromwich, D. H., Nicolas, J. P., Monaghan, A. J., Lazzara, M. A., Keller, L. M., Weidner, G. A., & Wilson, A. B. (2013). Central West Antarctica among the most rapidly warming regions on Earth. *Nature Geoscience*, *6*(2), 139–145.

Chung, C. E., Cha, H., Vihma, T., Räisänen, P., & Decremer, D. (2013). On the possibilities to use atmospheric reanalyses to evaluate the warming structure in the Arctic. *Atmospheric Chemistry and Physics*. https://doi.org/10.5194/acp-13-11209-2013

Claffey, K. J., Edgar, L. A., & Makshtas, A. (1994). Upper-air data collected on Ice Station Weddell. *United States. Army. Corps of Engineers & National Science Foundation (U.S.)*. Retrieved from https://catalogue.nla.gov.au/Record/4106075

Copernicus Climate Change Service (C3S) (2017). ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS). (n.d.). Retrieved April 16, 2019, from https://cds.climate.copernicus.eu/cdsapp#!/home

Cuxart, J., Holtslag, A. A. M., Beare, R. J., Bazile, E., Beljaars, A., Cheng, A., et al. (2006). Single-column model intercomparison for a stably stratified atmospheric boundary layer. *Boundary-Layer Meteorology*, *118*(2), 273–303. https://doi.org/10.1007/s10546-005-3780-1

de Boer, G., Shupe, M. D., Caldwell, P. M., Bauer, S. E., Persson, O., Boyle, J. S., et al. (2014). Near-surface meteorology during the Arctic Summer Cloud Ocean Study (ASCOS): Evaluation of reanalyses and global climate models. *Atmospheric Chemistry and Physics*, *14*(1), 427–445. https://doi.org/10.5194/acp-14-427-2014

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Dutrieux, P., de Rydt, J., Jenkins, A., Holland, P. R., Ha, H. K., Lee, S. H., et al. (2014). Strong sensitivity of Pine Island ice-shelf melting to climatic variability. *Science*, *343*(6167), 174–178. https://doi.org/10.1126/science.1244341

Fréville, H., Brun, E., Picard, G., Tatarinova, N., Arnaud, L., Lanconelli, C., et al. (2014). Using MODIS land surface temperatures and the Crocus snow model to understand the warm bias of ERA-Interim reanalyses at the surface in Antarctica. *The Cryosphere Discussions*, *8*(1), 55–84. https://doi.org/10.5194/tcd-8-55-2014

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., et al. (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Goessling, H. F., Jung, T., Klebe, S., Baeseman, J., Bauer, P., Chen, P., et al. (2016). Paving the way for the year of polar prediction. *Bulletin of the American Meteorological Society*, *97*(4), ES85–ES88. https://doi.org/10.1175/BAMS-D-15-00270.1

Gordon, A. L., & Ice Station Weddell Group of Principal Investigators and Chief Scientists (1993). Weddell Sea exploration from ice station. *Eos, Transactions American Geophysical Union*, *74*(11), 121–126.

Harden, B. E., Renfrew, I. A., & Petersen, G. N. (2011). A climatology of wintertime barrier winds off southeast Greenland. *Journal of Climate*, *24*(17), 4701–4717.

Jakobson, E., Vihma, T., Palo, T., Jakobson, L., Keernik, H., & Jaagus, J. (2012). Validation of atmospheric reanalyses over the central Arctic Ocean. *Geophysical Research Letters*, *39*, L10802. https://doi.org/10.1029/2012GL051591

Jones, P. D., & Lister, D. H. (2014). Antarctic near-surface air temperatures compared with ERA-Interim values since 1979. *International Journal of Climatology*, *35*(7), 1354–1366.

Jones, R. W., Renfrew, I. A., Orr, A., Webber, B. G. M., Holland, D. M., & Lazzara, M. A. (2016a). Evaluation of four global reanalysis products using in situ observations in the Amundsen Sea Embayment, Antarctica. *Journal of Geophysical Research: Atmospheres*, *121*, 6240–6257. https://doi.org/10.1002/2015JD024680

Jones, R. W., Renfrew, I. A., Orr, A., Webber, B. G. M., Holland, D. M., & Lazzara, M. A. (2016b). Evaluation of four global reanalysis products using in situ observations in the Amundsen Sea Embayment, Antarctica: Amundsen Sea Reanalyses Evaluation. *Journal of Geophysical Research: Atmospheres*, *121*, 6240–6257. https://doi.org/10.1002/2015JD024680

Jung, T., Gordon, N. D., Bauer, P., Bromwich, D. H., Chevallier, M., Day, J. J., et al. (2016). Advancing polar prediction capabilities on daily to seasonal time scales. *Bulletin of the American Meteorological Society*, *97*(9), 1631–1647. https://doi.org/10.1175/BAMS-D-14-00246.1

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., et al. (1996). The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*, *77*(3), 437–471. https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

Kobayashi, S., Yukinari, O. T. A., Harada, Y., Ebita, A., Moriya, M., Onoda, H., et al. (2015). The JRA-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan. Series II*, *93*(1), 5–48. https://doi.org/10.2151/jmsj.2015-001

Launiainen, J., & Vihma, T. (1990). Derivation of turbulent surface fluxes—An iterative flux-profile method allowing arbitrary observing heights. *Environmental Software*, *5*(3), 113–124.

Lindsay, R., Wensnahan, M., Schweiger, A., & Zhang, J. (2014). Evaluation of seven different atmospheric reanalysis products in the Arctic*. *Journal of Climate*, *27*(7), 2588–2606.

Liu, J., Zhang, Z., Hu, Y., Chen, L., Dai, Y., & Ren, X. (2008). Assessment of surface air temperature over the Arctic Ocean in reanalysis and IPCC AR4 model simulations with IABP/POLES observations. *Journal of Geophysical Research*, *113*, D10105. https://doi.org/10.1029/2007JD009380

Lüpkes, C., Vihma, T., Jakobson, E., König-Langlo, G., & Tetzlaff, A. (2010). Meteorological observations from ship cruises during summer to the central Arctic: A comparison with reanalysis data. *Geophysical Research Letters*, *37*, L09810. https://doi.org/10.1029/2010GL042724

Nicolas, J. P., & Bromwich, D. H. (2014). New reconstruction of Antarctic near-surface temperatures: Multidecadal trends and reliability of global reanalyses*. *Journal of Climate*, *27*(21), 8070–8093.

Nygård, T., Vihma, T., Birnbaum, G., Hartmann, J., King, J., Lachlan-Cope, T., et al. (2016). Validation of eight atmospheric reanalyses in the Antarctic Peninsula region. *Quarterly Journal of the Royal Meteorological Society*, *142*(695), 684–692. https://doi.org/10.1002/qj.2691

Pavelsky, T. M., Boé, J., Hall, A., & Fetzer, E. J. (2010). Atmospheric inversion strength over polar oceans in winter regulated by sea ice. *Climate Dynamics*, *36*(5-6), 945–955.

Perez, J., Menendez, M., Mendez, F. J., & Losada, I. J. (2014). Evaluating the performance of CMIP3 and CMIP5 global climate models over the north-east Atlantic region. *Climate Dynamics*, *43*(9-10), 2663–2680.

Rinke, A., Dethloff, K., Cassano, J. J., Christensen, J. H., Curry, J. A., Du, P., et al. (2006). Evaluation of an ensemble of Arctic regional climate models: Spatiotemporal fields during the SHEBA year. *Climate Dynamics*, *27*(4), 433–435. https://doi.org/10.1007/s00382-006-0157-1

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., et al. (2010). The NCEP Climate Forecast System Reanalysis. *Bulletin of the American Meteorological Society*, *91*(8), 1015–1058. https://doi.org/10.1175/2010BAMS3001.1

Saha, S., Nadiga, S., Thiaw, C., Wang, J., Wang, W., Zhang, Q., et al. (2006). The NCEP Climate Forecast System. *Journal of Climate*, *19*(15), 3483–3517. https://doi.org/10.1175/JCLI3812.1

Screen, J. A., & Simmonds, I. (2011). Erroneous Arctic temperature trends in the ERA-40 reanalysis: A closer look. *Journal of Climate*, *24*(10), 2620–2627.

Steig, E. J., Schneider, D. P., Rutherford, S. D., Mann, M. E., Comiso, J. C., & Shindell, D. T. (2009). Warming of the Antarctic ice-sheet surface since the 1957 International Geophysical Year. *Nature*, *457*(7228), 459–462.

Sturaro, G. (2003). A closer look at the climatological discontinuities present in the NCEP/NCAR reanalysis temperature due to the introduction of satellite data. *Climate Dynamics*, *21*(3-4), 309–316.

Tastula, E.-M., Vihma, T., & Andreas, E. L. (2012). Evaluation of polar WRF from modeling the atmospheric boundary layer over Antarctic Sea Ice in autumn and winter. *Monthly Weather Review*, *140*(12), 3919–3935.

Tastula, E.-M., Vihma, T., Andreas, E. L., & Galperin, B. (2013). Validation of the diurnal cycles in atmospheric reanalyses over Antarctic sea ice. *Journal of Geophysical Research: Atmospheres*, *118*, 4194–4204. https://doi.org/10.1002/jgrd.50336

Uotila, P., Holland, P. R., Vihma, T., Marsland, S. J., & Kimura, N. (2014). Is realistic Antarctic sea-ice extent in climate models the result of excessive ice drift? *Ocean Modelling*, *79*, 33–42.

Vancoppenolle, M., Timmermann, R., Ackley, S. F., Fichefet, T., Goosse, H., Heil, P., et al. (2011). Assessment of radiation forcing data sets for large-scale sea ice models in the Southern Ocean. *Deep-Sea Research. Part II, Topical Studies in Oceanography*, *58*(9-10), 1237–1249. https://doi.org/10.1016/j.dsr2.2010.10.039

Vihma, T. (2002). Surface heat budget over the Weddell Sea: Buoy results and model comparisons. *Journal of Geophysical Research*, *107*(C2), 3013. https://doi.org/10.1029/2000JC000372

Vihma, T., Johansson, M. M., & Launiainen, J. (2009). Radiative and turbulent surface heat fluxes over sea ice in the western Weddell Sea in early summer. *Journal of Geophysical Research*, *114*, C04019. https://doi.org/10.1029/2008JC004995

Vihma, T., Launiainen, J., Uotila, J., & Kotro, A. (1997). FINNARP Air-Sea-Ice interaction experiment in the Weddell Sea in 1996-1997. *Antarctic Reports of Finland 7,* 30.

Vihma, T., Pirazzini, R., Fer, I., Renfrew, I. A., Sedlar, J., Tjernström, M., et al. (2014). Advances in understanding and parameterization of small-scale physical processes in the marine Arctic climate system: A review. *Atmospheric Chemistry and Physics*, *14*(17), 9403–9450. https://doi.org/10.5194/acp-14-9403-2014

Walsh, J. E., & Chapman, W. L. (1998). Arctic cloud–radiation–temperature associations in observational data and atmospheric reanalyses. *Journal of Climate*, *11*(11), 3030–3045.

Walsh, J. E., Chapman, W. L., & Portis, D. H. (2009). Arctic cloud fraction and radiative fluxes in atmospheric reanalyses. *Journal of Climate*, *22*(9), 2316–2334.

Wang, J., Hock, T., Cohn, S. A., Martin, C., Potts, N., Reale, T., et al. (2013). Unprecedented upper-air dropsonde observations over Antarctica from the 2010 Concordiasi Experiment: Validation of satellite-retrieved temperature profiles. *Geophysical Research Letters*, *40*, 1231–1236. https://doi.org/10.1002/grl.50246

Wang, Y., Zhou, D., Bunde, A., & Havlin, S. (2016). Testing reanalysis data sets in Antarctica: Trends, persistence properties, and trend significance. *Journal of Geophysical Research: Atmospheres*, *121*, 12,839–12,855. https://doi.org/10.1002/2016JD024864

Wesslén, C., Tjernström, M., Bromwich, D. H., de Boer, G., Ekman, A. M. L., Bai, L.-S., & Wang, S.-H. (2014). The Arctic summer atmosphere: an evaluation of reanalyses using ASCOS data. *Atmospheric Chemistry and Physics*, *14*(5), 2605–2624.

Zib, B. J., Dong, X., Xi, B., & Kennedy, A. (2012). Evaluation and intercomparison of cloud fraction and radiative fluxes in recent reanalyses over the Arctic using BSRN surface observations. *Journal of Climate*, *25*(7), 2291–2305.