JGR Atmospheres

RESEARCH ARTICLE
10.1029/2020JD033389

Key Points:

- ECMWF is the overall best NWP system when compared against both the near-surface and upper-air observations, followed by AMPS and then GFS
- The models perform the poorest in locations with complex terrain along the coast of the Antarctic continent, and the best over the ocean
- The forecast quality degrades more rapidly with increasing lead time for the wind speed than air temperature and specific humidity

Supporting Information:

- Supporting Information S1

Correspondence to:
M. O. Jonassen,
mariusj@unis.no

Citation:
Jonassen, M. O., Nygård, T., & Vihma, T. (2021). Evaluation of three numerical weather prediction models for the Weddell Sea region for the Austral winter 2013. *Journal of Geophysical Research: Atmospheres*, *126*, e2020JD033389. https://doi.org/10.1029/2020JD033389

Received 30 JUN 2020
Accepted 7 DEC 2020

Evaluation of Three Numerical Weather Prediction Models for the Weddell Sea Region for the Austral Winter 2013

M. O. Jonassen[1,2], T. Nygård[3], and T. Vihma[1,3]

[1]Department of Arctic Geophysics, The University Centre in Svalbard, Longyearbyen, Norway, [2]Geophysical Institute, The University of Bergen, Bergen, Norway, [3]Finnish Meteorological Institute, Helsinki, Finland

Abstract It is widely recognized that numerical weather prediction (NWP) results for the Antarctic are relatively poor compared to the mid-latitudes. In this study, we evaluate output from three operational NWP systems: the ECMWF, Global Forecast System (GFS) and Antarctic Mesoscale Prediction System (AMPS), for the Austral winter (June-August) of 2013 for the Weddell Sea region, paying special attention to regional patterns of error statistics. This is the first evaluation of NWP systems over the Southern Ocean that also addresses the accuracy of forecasted vertical profiles. In the evaluation, we use data from land- and ship-based automatic weather stations (AWS) and radiosoundings. While the ECMWF and AMPS forecasts are on average biased cold and dry near the surface, the GFS forecasts are on average biased warm and moist. The near-surface wind speed is on average overestimated by the AMPS forecasts, whereas it is slightly underestimated by the forecasts of the other two NWP systems. Among the variables investigated, all three NWP systems forecast the near-surface specific humidity most accurately, followed by the temperature and then the wind speed. The forecast quality for the near-surface and upper-air wind speed degrades the most rapidly with increasing lead time, compared to the other variables. ECMWF is the overall best NWP system when compared against both the near-surface and upper-air observations, followed by AMPS and then GFS. The generally poorest model performance is found in locations with complex terrain along the coast of the Antarctic continent, and the best over the ocean.

Plain Language Summary It is widely recognized that weather forecasts are rather poor for the Antarctic compared to the mid-latitudes. In this study, we evaluate output from three weather forecasting systems: the ECMWF, GFS, and AMPS, for the Austral winter (June-August) of 2013 for the Weddell Sea region. This is the first evaluation of weather forecasting systems over the Southern Ocean that also addresses the accuracy of forecasted vertical profiles. In the evaluation, we use data from automatic weather stations and radiosondes. While the ECMWF and AMPS forecasts are on average biased cold and dry near the surface, the GFS forecasts are on average biased warm and moist. The near-surface wind speed is on average overestimated by the AMPS forecasts, whereas it is slightly underestimated by the forecasts of the other two forecasting systems. All three forecasting systems forecast the near-surface specific humidity most accurately, followed by the temperature and then the wind speed. ECMWF is the overall best weather forecasting system when compared against both the near-surface and upper-air observations, followed by AMPS and then GFS. The generally poorest model performance is found in locations with complex terrain along the coast of the Antarctic continent, and the best over the ocean.

1. Introduction

The performance of numerical weather prediction (NWP) systems is generally poorer at higher latitudes compared to other regions (e.g. Jung et al., 2016). This is due to various reasons, including scarcity of in situ observations and challenges related to utilization of satellite remote sensing observations, such as the polar night, lack of data from geostationary satellites, and difficulties in interpreting microwave signals originating from snow, ice, and clouds (Scarlat et al., 2018). Other reasons are complexity of surface boundary conditions due to rapidly changing sea ice concentration (Valkonen et al., 2014), rapidly developing mesoscale systems, such as polar lows and explosive cyclones (Pezza et al., 2016), and complexity of subgrid-scale physical processes related, among others, to stable boundary layer (Nigro et al., 2017), mixed-phase clouds

(Hines et al., 2019), and radiative transfer (Bromwich et al., 2013). All these factors pose challenges different to those at mid- and lower latitudes, which have probably received more attention in NWP model development. However, there is an increasing need for accurate and reliable weather forecasts for the polar regions due to increasing human activity. This need has spurred targeted, internationally coordinated efforts such as the World Meteorological Organization's Polar Prediction Project (PPP). As the major activity of PPP, the Year of Polar Prediction (YOPP; Goessling et al., 2016) has been carried out to enhance our understanding of polar weather forecasting and increase the number of observations in the polar regions.

Operational weather forecasting for the polar regions relies heavily on the use of global NWP forecasting systems in conjunction with limited area models. There is a fair number of studies that have evaluated the quality of NWP systems in detail for the Arctic (Atlaskin & Vihma, 2012; Hines & Bromwich, 2008; Inoue, 2020; Køltzow et al., 2019; Mölders & Kramm, 2010; Müller et al., 2017). For the Antarctic, several studies have evaluated NWP model results over the continent, where in situ observations from manned and automatic weather stations are available as a reference. In addition to presentation of basic error statistics, these studies have had specific objectives, such as evaluation of the added value provided by the Antarctic Mesoscale Prediction System (AMPS; Powers et al., 2003) compared to global model and analysis products (Monaghan et al., 2003; Wille et al., 2017) and on the optimal initial and boundary conditions for AMPS (Bromwich et al., 2013; Hines et al., 2019). Further, numerous studies have addressed the influence of model resolution and complex orography on the model performance with respect to near-surface wind and temperature conditions in the Antarctic coastal zone (Bromwich et al., 2005, 2013; Carrasco et al., 2003; Chen et al., 2014; Monaghan et al., 2003; Speirs et al., 2010; Wille et al., 2016, 2017). However, only a few studies have paid detailed attention to modeled vertical profiles of wind, temperature, and humidity. Coarse-resolution comparisons of model performance at the surface as well as 700 and 500 hPa levels were presented by Monaghan et al. (2003), whereas Wille et al. (2016) and Nigro et al. (2017) evaluated AMPS against detailed profile observations through the lowermost 30 m in the Alexander Tall Tower! (Mateling et al., 2018) on the Ross Ice Shelf and Wille et al., (2017) against unmanned aerial vehicle observations up to the height of 600 m in the same region.

Information on the quality of operational NWP products over the Southern Ocean is very limited. On the basis of a year-round evaluation of the European Centre for Medium-Range Weather Forecasts (ECMWF) operational analyses against drifting buoy observations from the Weddell Sea, Vihma et al. (2002) found that 2 m air temperature and surface temperature had positive biases of 3.5 and 4.4°C, respectively, but the synoptic-scale temperature variations were well represented. The errors were attributed to excessive cloud cover in winter and too thin sea ice without any snow pack. In lieu of in situ observations, Jung & Matsueda (2016) applied model analyses as a reference when evaluating results from several medium-range ensemble forecast systems for the Arctic and the Antarctic (65–90°N/S, i.e. including the Southern Ocean) for the period 2006/2007–2012/2013. They found that the mid-tropospheric flow was less well forecasted for the Antarctic than the Arctic, and that the 2 m air temperature analyses were poor in the Antarctic in winter. Also, Schroeter et al. (2019) applied model analyses as a reference when evaluating the Australian Community Climate and Earth-System Simulator-Global (ACCESS-G: Puri et al., 2013). With focus on the Southern Ocean and Antarctic, south of 50°S in the sector 85–120°E, Schroeter et al., (2019) found, among others, that the model performance weakened toward the high southern latitudes and with increasing forecast lead time.

In addition to evaluation of operational NWP results, NWP systems have also been run for experiments focusing on periods of major field campaigns. Tastula et al. (2012) simulated the period of the Ice Station Weddell (ISW) in February–June 1992 using the Polar WRF model. They found weak correlation between the observed and simulated surface fluxes of latent and sensible heat, and the simulations produced roughly 50% fewer low-level jets than observed. Valkonen et al., (2014) applied Polar WRF for the period of the Ice Station Polarstern, November 2004 to January 2005. The synoptic-scale evolution of the weather was well simulated, but the simulation of clouds, downward longwave radiation, boundary-layer processes as well as snow and sea ice thermodynamics posed major challenges.

To summarize the previous results, we may conclude that the operational NWP products for Antarctica are reasonably good for short-term forecasts of synoptic-scale systems, but they suffer from errors in lower boundary conditions related to snow and sea ice, as well as in subgrid-scale physical processes in the

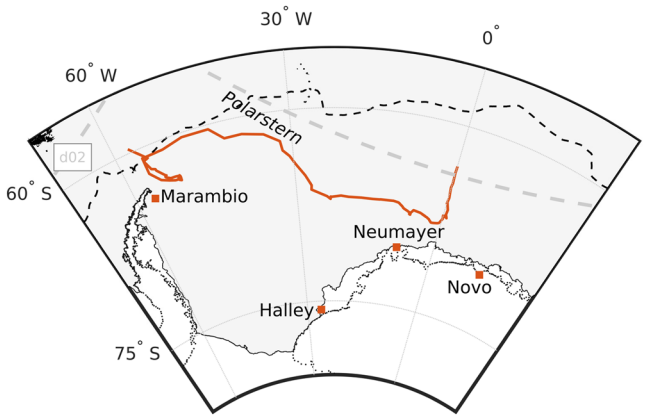

Figure 1. Names and locations of AWS and radiosonde stations (red squares). The Polarstern track from the study period is indicated with a red, solid line. The black, dotted line on the Antarctic continent marks the boundary between the grounding-line and ice shelves, and the black, solid line outside is the ice shelf edge. The horizontal extent of the AMPS second domain (d02, 10 km horizontal resolution) is indicated with a light gray, dashed line over the ocean. The mean sea ice edge (15% concentration) during the study period is indicated with a black, dashed line. AMPS, Antarctic Mesoscale Prediction System; AWS, automatic weather stations.

atmospheric boundary layer and clouds. Not surprisingly, these errors resemble those common in the Arctic (Hines & Bromwich, 2017; Køltzow et al., 2019). The accuracy of NWP products over the Southern Ocean is less well quantified. In particular, we need more information on the general model performance during winter, and for all seasons more details about the model performance with respect to vertical profiles as well as the impact of forecast lead time. Considering the Southern Ocean in winter, the conditions are characterized by a strong zonal circulation with a peak in cyclone occurrence in the belt from 55 to 70°S (Uotila et al., 2011). The June-July-August mean 2 m air temperature decreases from roughly 0°C around 55–60°S to −20°C at the coast of Antarctica. In addition to the strong baroclinicity, the large thermal differences between the areas of open water (leads and coastal polynyas) and near-surface air generate challenges for modeling of the atmosphere.

The purpose of this study is to contribute in filling these knowledge gaps. Our specific objectives are (a) to quantitatively evaluate three NWP systems with respect to both vertical profiles and near-surface variables over the Weddell Sea and surrounding coastal areas in winter, (b) to compare the challenges of NWP in different parts of our study region, and (c) to better understand processes affecting NWP performance. The forecast systems evaluated in this study are (1) the Integrated Forecast System (IFS) of the ECMWF, (2) the Global Forecast System (GFS) of the U.S. National Centers for Environmental Prediction (NCEP), and (3) AMPS. We evaluate them against in situ near-surface and upper-air observations of temperature, humidity, and wind speed.

2. Data and Methods

2.1. Observations

We apply near-surface and upper-air observations from a set of surface- and ship-based automatic weather stations and radiosondes from the Weddell Sea region covering the period June 20-August 9, 2013. The ship-based data, which were obtained from RV Polarstern (hereafter Polarstern), represent a rather unique data set as it was collected over sea ice during the Austral winter, when there have been otherwise very little in situ data available, in particular on vertical profiles. The study period covers the dates during which Polarstern was within the area of the AMPS 10 km domain, and thus when data are available from all three NWP systems. Details about the data sets, such as names and coordinates of the permanent stations, are given in Table S1 and their locations are indicated on the map in Figure 1. Note that in figures and tables, we will for the sake of simplicity refer to the station Novolazarevskaya as "Novo". The near-surface observations at the weather stations were taken at the heights of 2 m (air temperature and specific humidity) and 10 m (wind speed). An exception is Polarstern, where the measurements were made at 29 m and 39 m, respectively. The upper-air data were obtained in the same locations as the surface data, but their temporal data coverage varies. For Polarstern, the number of available soundings is 46, Neumayer 39, Novolazarevskaya 46, Halley 48, and Marambio 19. At Novolazarevskaya, all soundings were made at 00:00 UTC. At the rest of the stations, the majority of soundings were made at 12:00 UTC, with a few exceptions when soundings were made around 06:00 and 18:00 UTC. Both for near-surface and upper-air data, we only consider the 12:00 and 00:00 UTC observations in the presented statistics. However, the diurnal variability in winter in the Antarctic is small, and we assume that the varying observation times only have a minor effect on the results.

2.2. Forecast Systems

The IFS is a global NWP system developed and operated by ECMWF. In the beginning of the study period, the operational model cycle was 38r1, but on June 25, 2013 IFS cycle 38r2 was implemented. In the switch to cycle 38r2, the vertical resolution was increased from 91 levels to 137 levels. About 16 of these levels

were located below 800 hPa in cycle 38r1 and about 25 in cycle 38r2. The horizontal resolution was T1279 (~16 km) in both cycles. Simulations were initialized at 00:00 UTC and 12:00 UTC each day, and were run for 10 days. IFS applies four-dimensional variational data assimilation (4D-Var) with a 12-h time window. Sea surface temperature (SST) and sea ice concentration are based on daily analyses from the U.K. Met Office (Operational Sea Surface Temperature and Sea Ice Analysis, OSTIA).

GFS is a global NWP system developed by NCEP, which provides 16-day forecasts four times per day (00:00, 06:00, 12:00 and 18:00 UTC). In 2013, the horizontal grid spacing was T574 (~27 km) for the first 7 days and T192 (~84 km) for forecasts thereafter. Sea ice is obtained from the daily analysis by the NCEP Marine Modeling Branch. GFS uses a sigma pressure hybrid coordinate system with 64 vertical levels, of which 15 levels are below 800 hPa, and the model top is at 0.3 hPa. We downloaded GFS forecast data through the NOAA National Operational Model Archive and Distribution System (NOMADS).

AMPS is a regional NWP system, which provides 120-h forecasts twice daily. It is based on the Weather Research and Forecasting model in its polar version (Polar WRF), developed by the Byrd Polar and Climate Research Center of The Ohio State University and the National Center for Atmospheric Research (NCAR). The AMPS configuration is updated regularly. The polar modifications of WRF in AMPS include treatment of fractional sea ice, adjustments to the thermal and radiative properties of snow and ice surfaces and selection of subgrid-parameterization schemes best suited for polar conditions. As a regional model, the AMPS takes the first guess/boundary conditions from the global 0.5° GFS model output. As we will see in our results, this makes the analyses from AMPS and GFS very similar. In 2013 (the study period), the vertical resolution of AMPS had recently been increased from 43 to 60 model levels with a model top at 10 hPa. 15 of the model levels are located below 800 hPa. AMPS was run with five two-way nested domains, grid spacing of which were 30, 10, 3, 3, and 1 km. Observational data, including surface data, upper-air soundings, aircraft observations, geostationary, and polar-orbiting satellite atmospheric motion vectors, Constellation Observing System for Meteorology, Ionosphere and Climate (COSMIC) and GPS radio occultations, were assimilated into AMPS using three-dimensional variational data assimilation (3D-Var). In this study, we use the native-format output of the 10 km domain covering the Antarctic continent and some ocean area around it. The AMPS data were downloaded through the Earth System Grid (ESG) data portal.

2.3. Methods

To reach our goals set out in the Introduction, we (1) evaluate the model analyses and forecasts against the five available data sets of near-surface and upper-air observations, (2) compare the mean near-surface analysis fields of the models to get an overview of their spatial distributions of the variables, (3) compare the forecast fields to the analysis fields to assess the spatial distributions of errors, and (4) investigate how rapidly the forecast quality degrades with the lead time.

For the comparison between the models and the observations, data from the model grid points closest to the observation stations were obtained at a 12-hourly resolution (00:00 and 12:00 UTC). Due to substantial smoothing of the model topography, there is for several of the stations a large difference (>100 m) between the model topography and true elevation (Table S1). The GFS grid point closest to Novolazarevskaya is particularly far off with an altitude of 728 m above sea level (ASL), compared to 119 m ASL in reality. For this station, we therefore extracted GFS model data from one grid point farther to the north instead, which has an elevation of 50 m ASL. We note that the model compares slightly less favorably against the observations using this lower-elevation grid point than the closest grid point (see Figure S1 in the supporting information). However, we use the former grid point because of its closer match in elevation and the uncertainties related to adiabatic adjustments of the 2 m temperature, as outlined below. Furthermore, for Neumayer, the closest grid point in GFS is located over the sea, and we therefore extracted the model data from one grid point farther to the south for this station, which is located on the ice shelf. GFS data from this latter grid point compare generally more favorably against the observations than data from the sea-based one (see Figure S1). Moreover, parts of the model topography in the GFS data set have negative height values where steep mountains lie near the continental coastline of Antarctica. This is a common issue in spectral models, such as GFS, and is caused by spectral truncation (Hogan & Rosmond, 1991). Differences between model and true terrain elevation naturally influence the comparison of near-surface temperature. To compensate for this effect the modeled near-surface 2 m air temperature has frequently been adjusted to the height of

the true terrain in model evaluation studies, applying either the dry adiabatic lapse rate (e.g. Bromwich et al., 2005; Tastula & Vihma, 2011) or a moist adiabatic lapse rate (e.g. Jones & Lister 2015). However, we note that these adjustments have their potential deficiencies. Temperature profiles along a sloping surface are not always dry adiabatic, and along the coast, where the land-based stations of our study are all located, the decrease is typically less than dry adiabatic (King & Turner, 1997). Because of the stations' locations in our study, we have chosen to adjust the near-surface temperature by a moist adiabatic lapse rate of 0.006°C/m, which is the same as in Jones and Lister (2015). Though, we note that adjusting the temperatures by choosing a lapse rate typical for dry conditions (0.00981°C/m), does not impact the temperatures to a large degree (see Figure S1).

To allow direct comparison of near-surface data from Polarstern with the model output, we applied the model results from levels above and below the observation heights of 29 m ASL (temperature and humidity) and 39 m ASL (wind speed) and calculated values for these heights using an iterative algorithm taking into account static stability and surface roughness (Launiainen & Vihma, 1990).

For the comparison of upper-air data, we consider data from 10 pressure levels between 925 hPa and 500 hPa. Above the 500-hPa level, the uncertainty of humidity observations increases notably in the polar atmosphere due to cold and dry conditions (e.g. Bromwich et al., 2005), and therefore, we excluded those data from the analysis.

In the model evaluation, we consider data from the lead times 0, 12, 24, 48, 72 and 120 h. Among these, we put the main emphasis on the analyses (0 h) as well as the average of the 24 and 48 h forecasts. In addition, we use data from all above-mentioned lead times averaged over all five stations to evaluate the degradation (or improvement) in model performance with increasing lead time. In this averaging, we first averaged all data from each station separately, and thereafter calculated the average of these again. This way, each station's data set is weighted equally in the final average values, also for the upper-air observations, even though each station has a different number of radiosoundings.

The error statistics that we applied are the bias (modeled value minus observed value), the normalized root mean squared error (NRMSE) and the correlation coefficient (r). The NRMSE is the ratio of the root mean squared error (RMSE) divided by the standard deviation of the observed values of the particular variable at the station. This normalization allows for comparison of RMSE from different sites with different variability, and comparison of the results for different variables. We calculated the significance of the biases and r at the 95% level using the Student t test. The values of r are largely significant throughout the results and only in cases where they are not we will point this out.

It should be noted that the observational data used for comparison with the model data are regularly used for data assimilation. A model analysis is made by correcting a background field (based on the previous short-term forecast) by assimilation of observations. Hence, if the same observations are used in evaluating the analysis, the skill scores are expected to be higher than in the case that the evaluation was made against observations not applied in the data assimilation. Due to temporal autocorrelation of weather variables, the effect is also reflected in the skill scores of short-term forecasts. However, in the case of the observations applied in this study, the same observations were available for data assimilation in all NWP systems evaluated. Hence, the comparison of differences in the performance of the NWP systems is not affected.

3. Results and Discussion

3.1. Comparison of Near-Surface and Upper-Air Data

We here evaluate the NWP systems' near-surface (Figures 2–4) and upper-air (Figures 5–7) performance. We do so by comparing the model analysis and 24–48 h forecast data to observed surface-based (AWS) and profile (radiosondes) data of temperature, specific humidity, and wind speed from the five available observational data sets.

3.1.1. Temperature

The lowest mean 2 m air temperature was observed at Halley (−24.7°C, Figure 2), which is located on the Brunt Ice Shelf. Here, cold near-surface air masses prevailed during the study period, which are seen in the

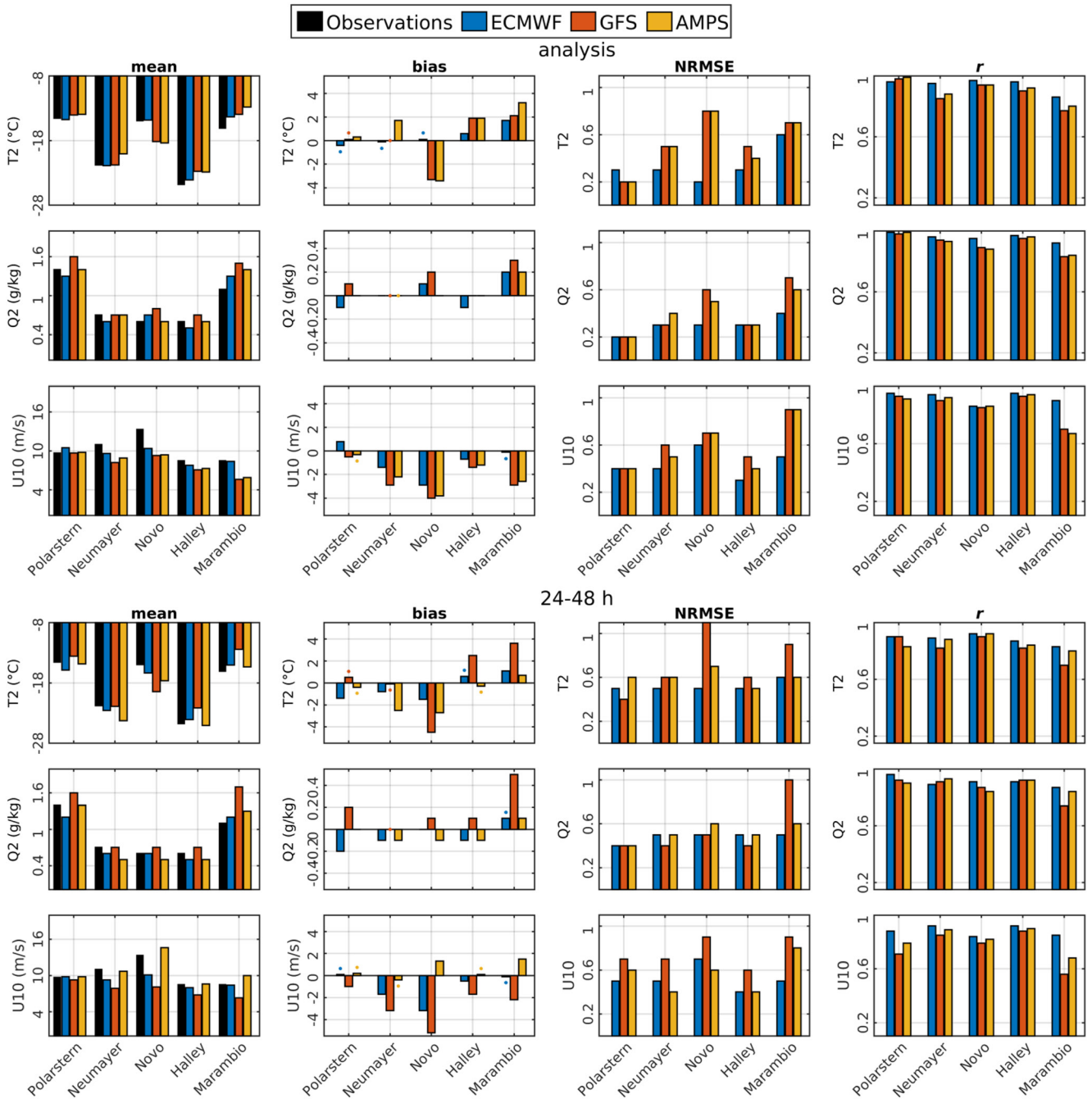

Figure 2. Mean values, bias, NRMSE, and correlation coefficient (*r*), for the ECMWF, GFS, and AMPS analyses and 24–48 h forecast data sets using AWS data as reference. A positive bias indicates that the NWP system has a higher value than the observations. The statistics are presented for the 2 m air temperature (T2), 2 m specific humidity (Q2), and 10 m wind speed (U10). Dots above or below the bias bars indicate non-significant biases. AMPS, Antarctic Mesoscale Prediction System; AWS, automatic weather stations; ECMWF, European Center for Medium-Range Weather Forecasts; GFS, global forecast system; NRMSE, normalized root mean squared error; NWP, numerical weather prediction.

mean model analysis fields (Figure 3). The highest mean 2 m air temperature was observed at Polarstern (−14.5°C, Figure 2), which is natural given its location in the sea ice zone.

Among the more prominent biases are the cold biases in the NWP systems' analysis and 24–48 h forecast data for Novolazarevskaya, and these are particularly large (up to ∼−3.4°C) in the GFS near-surface (Figure 2) and lower atmosphere (900–850 hPa, Figure 5) 24–48 h forecasts. ECMWF has the smallest biases for

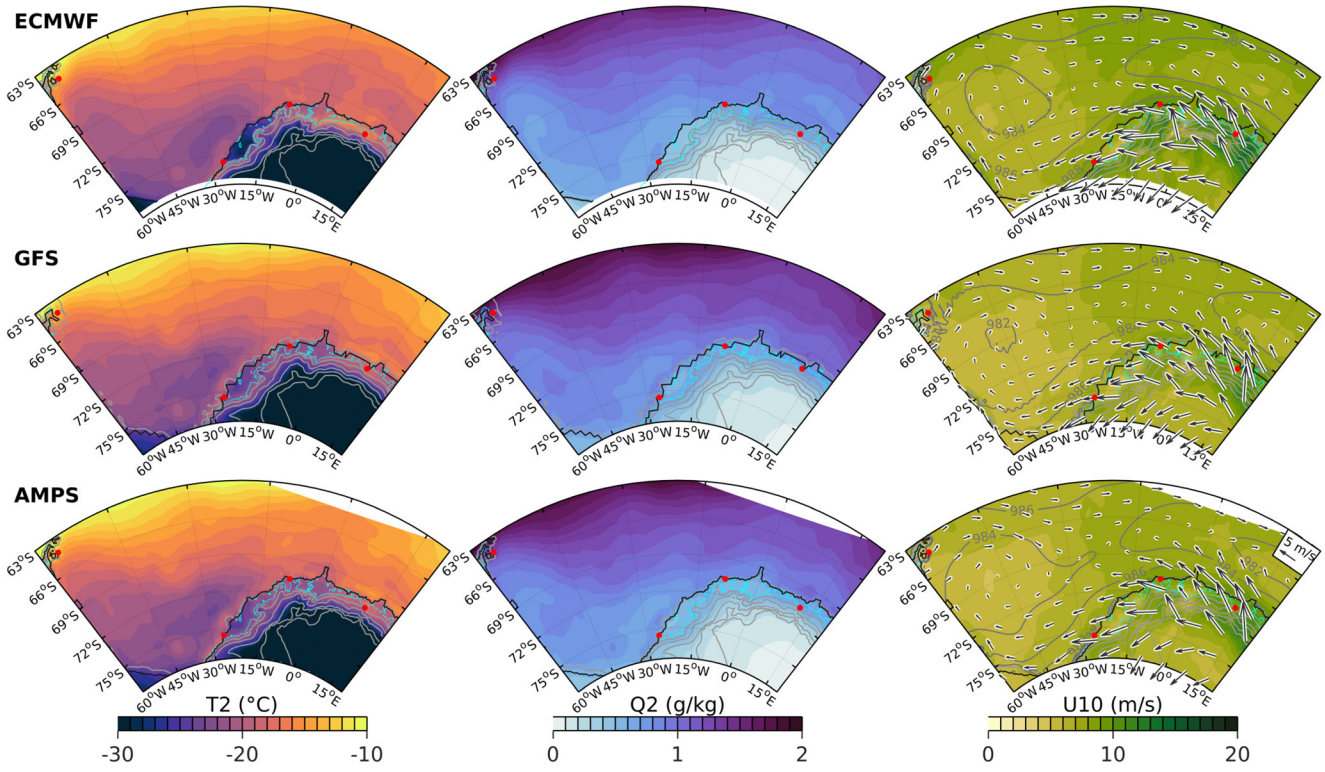

Figure 3. Mean 2 m air temperature (T2, left column), 2 m specific humidity (Q2, middle column), and 10 m wind speed (U10, right column) from the analysis fields. Wind vectors are given in the U10 panels and the ice shelf grounding line is indicated with a cyan contour line in each panel. Mean sea level pressure (in hPa) is indicated by contours in the U10 panels. AWS and radiosonde station locations are marked with red dots (see Figure 1 for names) and surface elevation contours are given with gray lines in the T2 and Q2 panels. AWS, automatic weather stations.

this station. The mean near-surface temperature model fields (Figure 3) help put these results in context. The ECMWF analysis evidently reveals more spatial detail than the other two analyses in the areas dominated by complex topography. One example is the band of relatively high temperatures (up to about −20°C) extending between Novolazarevskaya and Halley, at the southern (upwind) side of the ice shelf grounding line (cyan line in Figure 3). This band is more distinct in the ECMWF analysis, with sharper gradients toward the colder, immediate surroundings, where the temperature is down to about −28°C. In contrast to the other two NWP systems' analyses, the ECMWF analysis is biased slightly warm at Novolazarevskaya (~+0.1°C), suggesting that the warm temperature band is more accurately represented by the ECMWF analysis. In its 24–48 h forecast data, ECMWF is also biased cold, which goes in hand with its near-surface forecast being colder than the analysis in a larger area along the coast between Neumayer and Novolazarevskaya (Figure 4). Other marked biases are the warm near-surface biases (~+2°C) in the GFS and AMPS analyses when compared against the Halley observations (Figure 2). In contrast, the ECMWF analysis has a bias of only about +0.5°C here. For the 24–48 h forecasts, only GFS retains a significant bias at Halley, which is similar in magnitude to its analysis bias. Also, we note that AMPS has smaller biases in its 24–48 h forecasts for this station than in its analyses. Compared against the upper-air Halley observations (Figure 5), all three model analyses perform remarkably well, with biases smaller than +/−0.5°C for all the height levels. In the 24–48 h forecast data, however, both GFS and AMPS suffer from warm biases at most levels at this station, peaking at +1.5°C at 800 hPa. We note that the mean 2 m air temperature over the Brunt Ice Shelf is lower in the ECMWF analysis (~−28°C) than in the other two analyses (~−24°C, Figure 3), indicating that the ECMWF analysis more accurately represents the cold temperatures here.

We find particularly large differences between the near-surface AMPS analysis and 24–48 h forecasts (Figure 4), and this is especially evident for the Antarctic interior where there are differences of up to −7°C. There are also large differences in central and northern parts of the Weddell Sea and over the complex topography near the coast. The reason for these large differences is not clear, but we do note that the AMPS

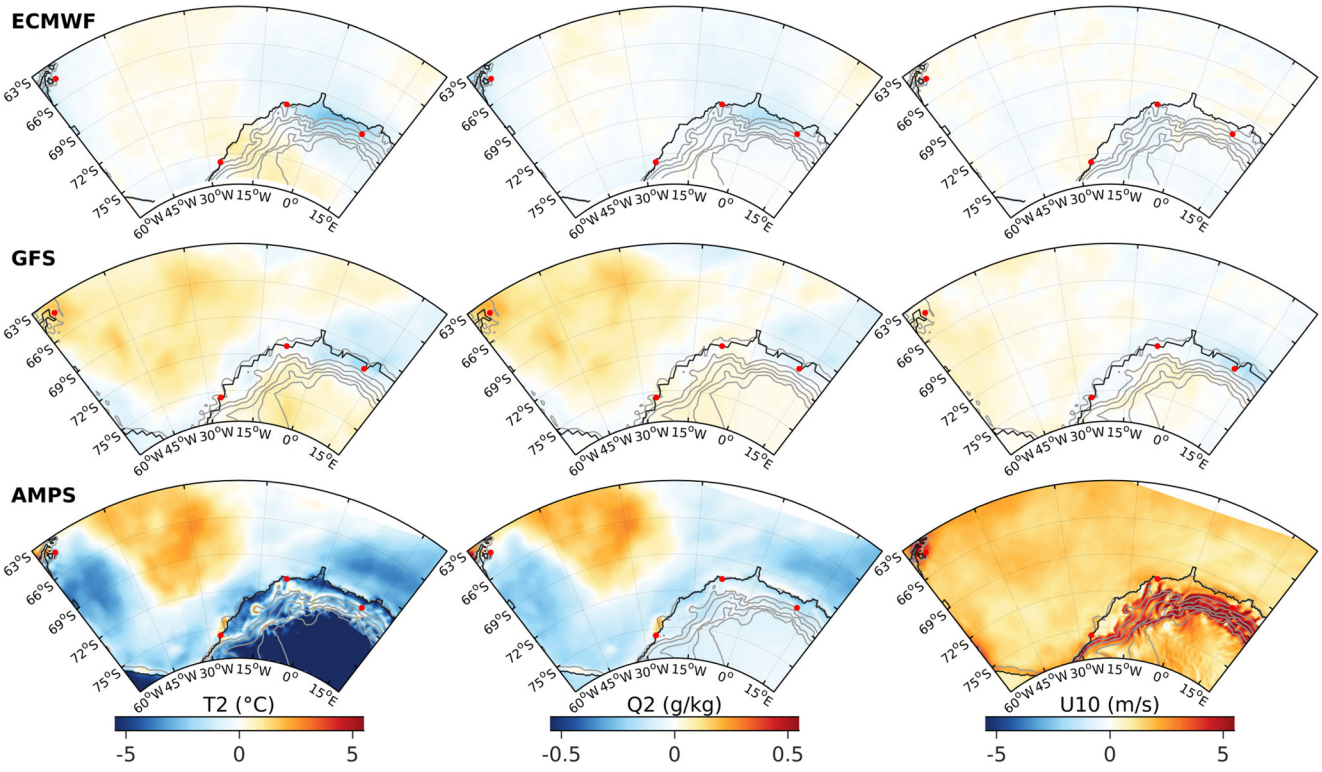

Figure 4. Mean difference between analyses and 24–48 h forecast fields of 2 m air temperature (T2), 2 m specific humidity (Q2), and 10 m wind speed (U10) fields from the ECMWF, GFS, and AMPS. Positive (negative) differences indicate higher (lower) values in the 24–48 h forecast fields. AWS and radiosonde station locations are marked with red dots (see Figure 1 for names) and topography height contours are given with gray lines. AMPS, Antarctic Mesoscale Prediction System; AWS, automatic weather stations; ECMWF, European Center for Medium-Range Weather Forecasts; GFS, global forecast system.

and GFS models feature different physics. The results for NRMSE and *r* indicate the overall poorest performance across the models for Marambio, both in the near-surface (Figure 2) and upper-air data (Figure 5). We find the least favorable values of NRMSE (~1) and *r* (~0.6) around 900 hPa in the GFS 24–48 h forecasts. In contrast, the models compare generally the most favorably against the Polarstern observations, where GFS performs particularly well among the near-surface 24–48 h forecast data, with an NRMSE of 0.4 and an *r* of 0.9.

3.1.2. Humidity

Regarding the specific humidity, we find the highest mean observed value in the near-surface data from Polarstern (~1.4 g/kg, Figure 2), which was situated in a relatively moist, marine environment during the study period. In contrast, Halley is notably drier and has a mean observed value of only around 0.6 g/kg in the near-surface data, which is congruent with the low temperatures at this station. As can be expected, because the near-surface relative humidity is always near to ice-saturation over polar sea ice (Andreas, 2002), the spatial patterns in the 2 m specific humidity fields are rather similar to those for the 2 m air temperature. This is true both for the mean analysis fields (Figure 3) and the difference between these and the mean 24–48 h forecast fields (Figure 4). We will therefore describe these in less detail than for the 2 m air temperature.

We find the largest specific humidity biases in the GFS 24–48 h forecasts for Marambio, using the near-surface (Figure 2) and upper-air observations (Figure 6) as references. At 2 m, the bias is +0.5 g/kg, and in the upper-air data it reaches a maximum of +0.5 g/kg between 700 and 600 hPa. As can be seen in Figure 4, a larger area around Marambio has higher 2 m specific humidity values in the 24–48 h GFS forecasts than in the analysis. Also worth noting are the dry biases dominating in the ECMWF 24–48 h forecasts, which we see at most stations from 2 m and up to between 850 hPa (Halley) and700 hPa (Neumayer). In terms of NRMSE and *r,* the models' near-surface performance is particularly poor for Marambio, and for this station the GFS 24–48 h forecasts have values of 1 and 0.74, respectively. In contrast, the three NWP systems

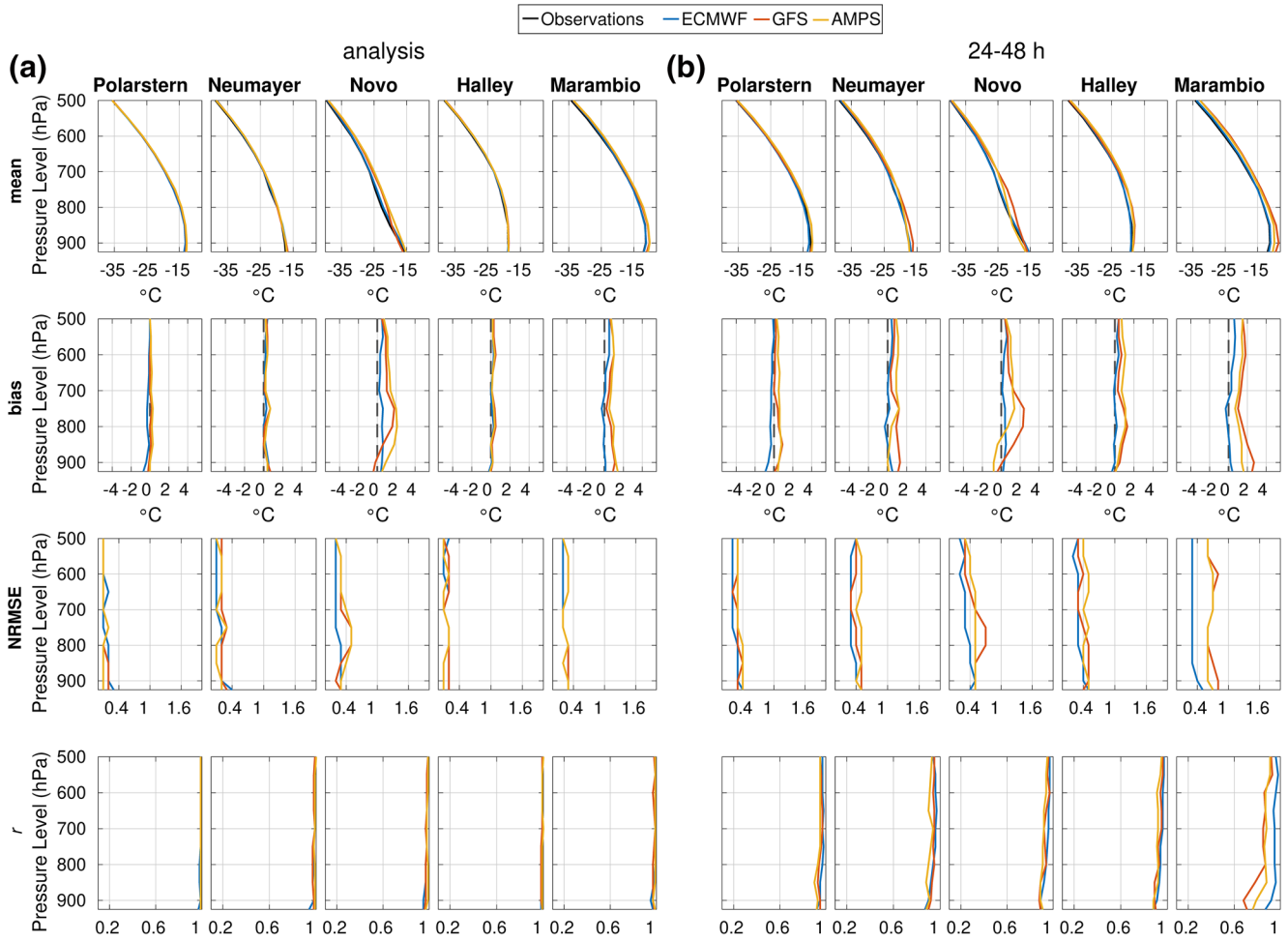

Figure 5. Profiles of mean, bias, NRMSE, and correlation coefficient (r) for temperature for the analyses (panels a) and 24–48 h forecasts (panels b).

compare particularly well against the near-surface data from Polarstern. For this station, the NWP systems have the same NRMSE values (0.4) and ECMWF has the highest r (0.96) in the 24–48 h forecasts.

3.1.3. Wind

Concerning the wind speed, the highest near-surface mean value was observed at Novolazarevskaya (13.3 m/s) and the lowest at Halley and Marambio (8.5 m/s). In terms of upper-air data (Figure 7), the highest mean wind speeds (up to 17 m/s) in the lower layers (below 700 hPa) are also found at Novolazarevskaya, within a structure reminiscent of a low-level jet. The mean near-surface model analysis fields (Figure 3) help us to put these values into a spatial context. We see strong winds along and down the slopes of the ice sheet, in particular close to Neumayer and Novolazarevskaya. At Halley, the cyclone activity is weaker (Uotila et al., 2011) and the katabatic winds typically do not reach the cold boundary layer over the ice shelf (Renfrew & Anderson, 2002). Marambio is located on the lee side of the Antarctic Peninsula mountains, and therefore sheltered from the strongest zonal winds. Further, there is no strong local katabatic flow on the island.

Negative near-surface wind speed biases dominate in the data from all three NWP systems' analyses, and in the 24–48 h forecasts from ECMWF and GFS (Figure 2). These biases in the ECMWF and GFS forecasts are particularly large at Neumayer and Novolazarevskaya, where strong winds prevailed, and where we find negative biases also in the lower parts of the upper-air forecast data (Figure 7). Unlike the ECMWF and GFS 24–48 h forecasts, AMPS has mostly no or even positive wind speed biases compared against both the near-surface (Figure 2) and upper-air (Figure 7) observations. AMPS has also everywhere higher near-surface wind speeds in the 24–48 h forecasts than in its GFS-based analysis (Figure 4).

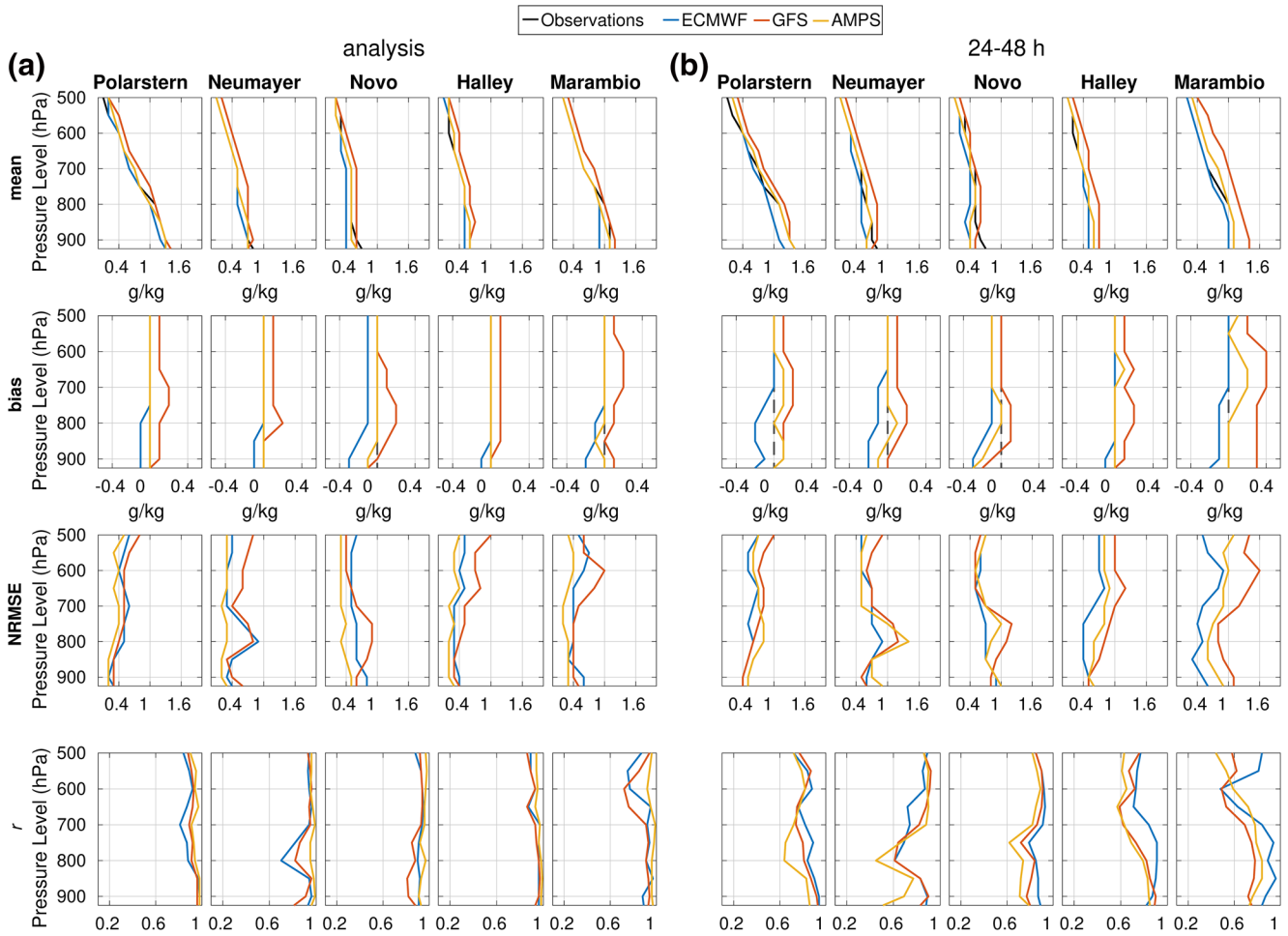

Figure 6. As in Figure 5, but for specific humidity.

As for the temperature and specific humidity, also for wind speed the models' values of NRMSE and *r* are particularly poor for Marambio and Novolazarevskaya. At Marambio, we find the least favorable values in the 24–48 h GFS forecasts at 925 hPa, where the NRMSE is 1.3 and the *r* is below 0.2. The model systems perform on average the best when compared against the near-surface observations from Polarstern, with relatively small and even insignificant biases in all three models' analyses and 24–48 h forecasts. Among the 24–48 h near-surface forecasts, ECMWF has the best performance for this station with an NRMSE of 0.5 and an *r* of 0.88.

Although we have not specifically calculated error statistics for mean sea level pressure (mslp) in this study, we find that the models compare qualitatively well when considering their average mslp fields for the study period (Figure 3, right column). This suggests that the large-scale pressure gradients and corresponding atmospheric flow are fairly similar in all of the models. However, on average, GSF has approximately 1 hPa lower mslp than ECMWF and AMPS.

3.2. Discussion of Prominent Model Errors

We now discuss the model error statistics, paying particular attention to the most prominent and consistent model errors. A good example of these are the relatively large near-surface cold biases dominating in the forecasts for Novolazarevskaya, in particular in GFS. As we will argue in the following, these can be better understood by consulting the mean near-surface model analysis fields (Figure 3), and especially the temperature and wind fields. The latter resembles rather closely the climatological (long-term mean) wind field

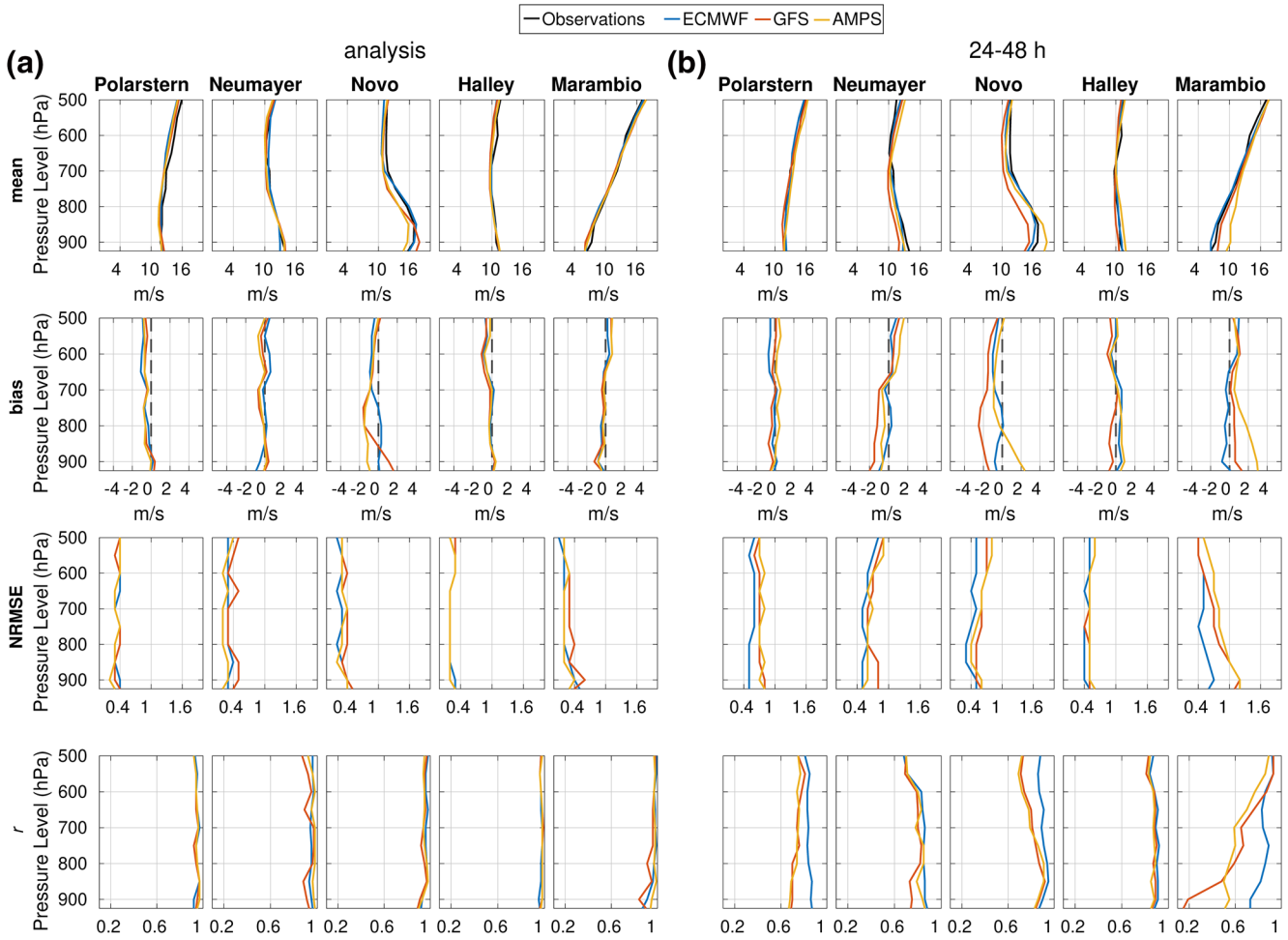

Figure 7. As in Figure 6, but for wind speed.

for the region, as deduced by for example Van Den Broeke and Van Lipzig, (2013) and Vignon et al. (2019). These studies found the wind field along the Antarctic coast line and nearby mountains to be dominated by (1) katabatic wind and (2) wind induced by large-scale pressure gradients connected to synoptic scale depressions. Taking Novolazarevskaya as an example, Turner and Pendlebury (2004) confirm the prevalence of these flow types there. According to them, these flow types are typically associated with winds from (1) the south-southeast to south-southwest and (2) the east-southeast at this station. We find that the time series of both the observed and modeled wind data from this station (see Figure S2) are congruent with this. Considering both the mean wind and temperature fields in Figure 3, we see that the aforementioned band of high temperatures upstream of the ice shelf grounding line coincides with relatively low wind speeds adjacent to it. Such an interaction between the near-surface wind and temperature fields is what one would expect with katabatic winds over slopes of the dimensions here considered (Vihma et al., 2011), where katabatic winds cause adiabatic warming and vertical mixing in the stable boundary layer. Although cold biases prevail at Novolazarevskaya in the GFS analysis and 24–48 h forecasts throughout the study period, they appear to be enhanced during katabatic winds (see Figures S2–S4), presumably due to the under-representation of the adiabatic warming. This could explain why GFS resolves the mentioned warm band more poorly, possibly due to a coarser horizontal grid resolution. It would also be in line with findings by (Tastula & Vihma, 2011) on cold biases in Polar WRF at Rothera and Casey, who argued that these biases could be enhanced by subsidence heating due to local downslope winds not being resolved by their simulations. The role of katabatic or other downslope winds in generation of the cold biases is also supported by the fact that studies evaluating ECMWF, GFS and Polar WRF results over flat surfaces in the Arctic and Antarctic have

mostly revealed warm near-surface biases, as in Atlaskin and Vihma, (2012) for the Finnish Arctic in winter and Jakobson et al. (2012) for the central Arctic Ocean in summer. Cold biases at coastal Antarctic stations have also been found in other studies evaluating NWP forecasts, for example Bromwich et al. (2005) (using AMPS and Polar MM5), and in studies evaluating reanalyzes for Antarctica (P. D. Jones & Lister, 2015; R. W. Jones et al., 2016; Vancoppenolle et al., 2011; Vihma et al., 2002). A few of these studies have been conclusive as to alternative reasons for such cold biases. Bromwich et al. (2005) attributed coastal cold biases partly to topographic smoothing, which they argued makes the stations being placed further inland than in reality, thus giving the stations a more continental environment than observed. In our investigation of the models' land-sea mask (not shown), however, the stations appear to be at about the same distance to the actual coast line as the model coast line. Hence, this effect seems not to be important in our results. Negative wind speed biases dominate in all three model analyses, and in the 24–48 h forecasts by ECMWF and GFS. A closer look at time series of near-surface wind speed data (see Figure S2 for data from Novolazarevskaya) reveals that the underestimations generally increase with increasing wind speed. This is true for all stations in the GFS data and also in the ECMWF data for Neumayer and Novolazarevskaya. Furthermore, we find that the strongest winds along the Antarctic continent in our study area were induced by large-scale pressure gradients connected to synoptic scale depressions (mechanism 2 outlined above). Such strong wind is known to be enhanced at the mesoscale when stably stratified flow impinges on and is forced to flow along mountains, forming a phenomenon known as barrier flow (e.g. Jonassen et al., 2020). The wind field along the slopes in the region (Figure 3) does resemble barrier flow during strong synoptic flow induced by mechanism 2 under the situations with the strongest winds (see also Figures S5 and S6). It is probable that an underestimation of the enhancement from barrier flow, and/or not proper representation of local effects from terrain, contribute to the negative wind speed biases that we find in ECMWF and GFS, especially at Neumayer and Novolazarevskaya. Underestimation of near-surface winds at Novolazarevskaya by ECMWF, in particular in winter, was also detected in the year round comparison by Zhang et al., (2015).

Positive wind speed biases dominate in AMPS, and such biases have also been found in previous studies for both the Antarctic (Bromwich et al., 2005, 2013; Tastula et al., 2012; Tastula & Vihma, 2011; Valkonen et al., 2014; Wille et al., 2017) and the Arctic (Wilson et al., 2011). From the results of these studies, it appears that these positive biases are related to the Mellor-Yamada-Janjić (MYJ) boundary layer scheme. Another factor contributing to the positive wind bias in AMPS may be the lack of parameterization of subgrid-scale orography generated by nunataks, which tend to slow down winds in the marginal zone of the ice sheet (Zhang et al., 2015). Evaluated against observations from nine coastal stations (including Neumayer and Novolazarevskaya), Zhang et al. (2015) found that the ECMWF near-surface winds had much better skill scores than those of AMPS. We note that wind speed biases, both negative and positive ones, might have unfavorable downstream effects on ocean and sea ice models using winds from NWP models as forcing. For example, Zhang et al. (2015) found that coastal ice drift, concentration, and thickness were highly sensitive to wind forcing in their simulations. We find the generally poorest model performance for the Marambio station for all variables considered, both in terms of the near-surface and upper-air data. There may be several reasons for this. A closer look at the model physiographic fields (not shown) reveals that none of the NWP systems resolve the Marambio Island on which the station is situated. Thus, in the land-sea masks of all three models (not shown), the station is located over the (mostly sea-ice covered) ocean. In order to investigate the influence from the model surface cover on the near-surface model results for Marambio, we extracted data from the closest land-based grid point in each model for this station. The results (Figure S1), show a generally poorer model performance when using data from these grid points rather than the original grid points, in particular with respect to near-surface temperature and wind speed. The differing model topography between these grid points might have an influence on the results, but this analysis nevertheless indicates that the models' surface cover is not the main source for the dominating model errors at Marambio.

The large model errors at this station could also be related to the difficulties in simulating the flow over the Antarctic Peninsula. Poor representation of the Peninsula orography is known to generate major errors in the wind field downstream of the Peninsula (Stössel et al., 2011) also seen in several atmospheric reanalyses as larger errors in air temperature and specific humidity in the eastern than western side of the Peninsula (Nygård et al., 2016). Furthermore, we note that Marambio had clearly less frequent radiosoundings than the other stations during the study period (see Section 2.1), and this could affect the forecast quality negatively through less data available for assimilation.

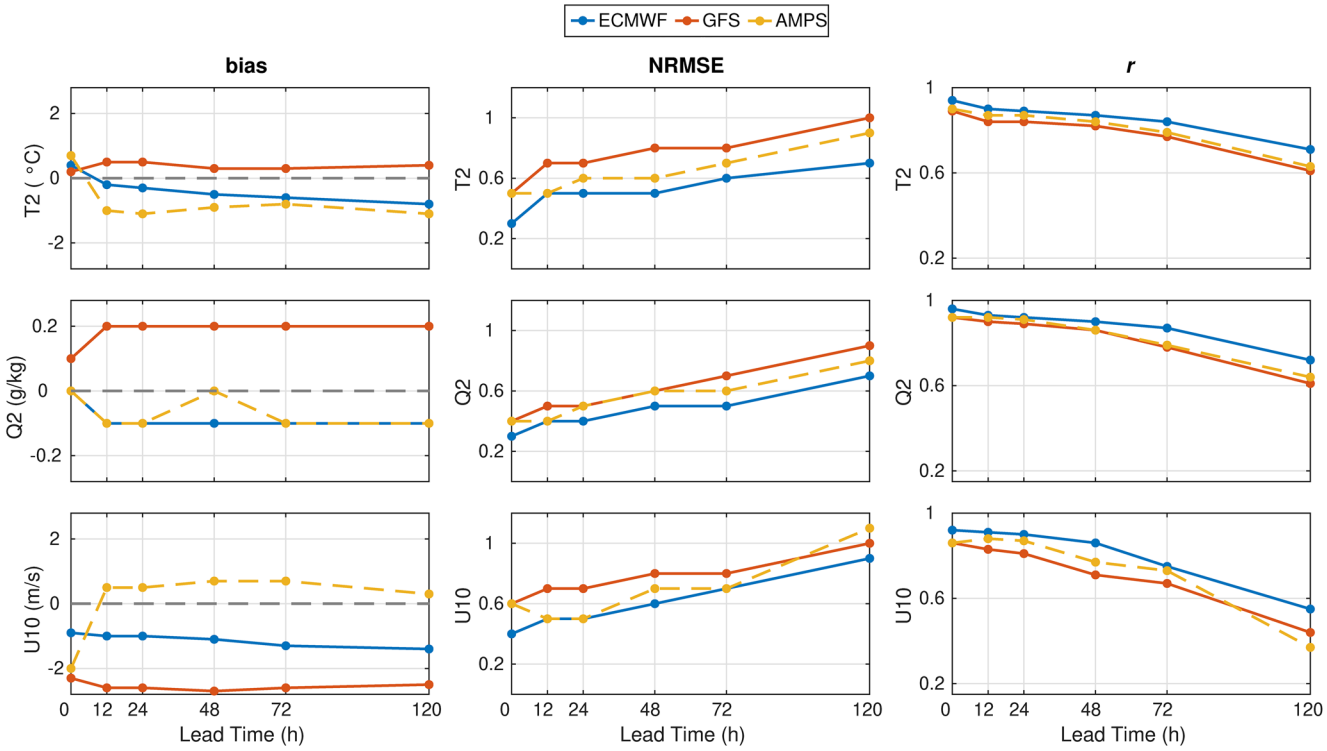

Figure 8. Mean bias, NRMSE and correlation coefficient (*r*) versus lead time (h) averaged over all five AWS. A positive bias indicates that the NWP system has a higher value than the observations. The statistics are presented for the 2 m air temperature (T2), 2 m specific humidity (Q2), and 10 m wind speed (U10).

3.3. Overall Model Performance and Effect of Lead Time

We here present the near-surface error statistics for different lead times. We have averaged the error statistics for each meteorological variable over all five available data sets (Figure 8). This lets us: (1) summarize the near-surface performance of each NWP system and compare how they perform with respect to each other, (2) judge which near-surface meteorological variable is forecasted the best, and (3) investigate how the error statistics are affected by lead time for each NWP system and meteorological variable.

The near-surface model biases for the different lead times indicate that GFS is on average slightly too warm (∼+0.4°C), too moist (mostly +0.2 g/kg) and the wind speed is too low (∼−2.1 m/s). The warm biases are dominated by the stations Halley and Marambio, both for the shorter (24–48 h, Figure 2) and longer lead times (not shown). The moist biases are seen across most of the stations and the same is the case for the negative wind speed biases. ECMWF, in contrast, is on average too cold (down to −0.8°C, at 120 h) and slightly too dry (mostly −0.1 g/kg) and the wind speed is too low (down to −1.4 m/s at 120 h). These biases are largely consistent across all stations and lead times, except for the analysis. The AMPS analysis (0 h) is on average biased warm (∼+0.7°C) and the wind speed is biased low (∼−2 m/s). Naturally, these values are close to the GFS analysis, which the AMPS analysis is based on. For longer lead times, AMPS is on average biased cold (down to ∼−1.1°C, at 24 h) and dry (mostly −0.2 g/kg) and the wind speed is, unlike for the other NWP systems, biased slightly high (up to ∼+0.8 m/s at 48 and 72 h). Apart from the analysis and a few stations with non-significant biases, these biases are mostly consistent across the different stations and lead times.

Considering the NRMSE and *r*, all three NWP systems have a near consistent degradation in values with increasing lead time for all near-surface parameters considered. However, the performance of AMPS improves during the first 12 h, probably due to a decreasing influence with increasing lead time from the initial conditions which are based on GFS. ECMWF is, with a few exceptions, the best NWP system judged by these error statistics for each individual near-surface parameter and lead time, followed by AMPS and thereafter GFS. Exact reasons for the good performance of ECMWF are difficult to identify. However, among the three NWP systems evaluated, it is the only that employs a 4D-Var assimilation scheme, which is expected

to yield advantage over the other systems' 3D-Var schemes (e.g. Yang et al., 2009). Further, the ECMWF model has a much higher number of vertical model levels than GFS and AMPS and a higher horizontal resolution than GFS. Regarding ECMWF, we note that the model was updated to a new cycle (from 38r1 to 38r2) during the study period, among others increasing the number of vertical model levels from 91 to 137. It is difficult to assess the impact of this change on our results, as the two periods are relatively short and they do not overlap in time. With a short time period the signal from the potential model improvement is not distinguishable from the random noise of variability in the error statistics. Though, we note that time series of the near-surface variables at Novolazarevskaya (Figure S2 in the supporting information) do not reveal any obvious discontinuities on the 25 June, when the new cycle was implemented. In any case, we assume that the differences between these two model cycles are (much) smaller than the differences between the three NWP systems.

The most accurately forecasted near-surface variable is specific humidity, as judged by the values of NRMSE and *r*. Temperature is second and wind is third in this regard. Using the 48 h forecast statistics averaged over the three model products as an example, the values of NRMSE are 0.57, 0.63, and 0.70 for humidity, temperature and wind speed. For *r*, the corresponding values are 0.87, 0.84, and 0.78. Furthermore, we note that the forecast performance decreases more rapidly with increasing lead time for the wind speed than for the other parameters (followed by specific humidity and temperature), as seen by faster degrading values of NRMSE and *r*. Considering the values of NRMSE averaged over all three models, we find mean increases of 0.0039, 0.0034, and 0.0031 per lead time hour for wind, humidity, and temperature, respectively. For *r*, the corresponding values are −0.0036, −0.0023, and −0.0020.

The faster degradation in model performance with respect to wind speed than the other variables is in line with (Haiden et al., 2019), who show in their Figure 31 that in Europe the ECMWF operational model forecast accuracy decreases faster with increasing lead time (from 24 to 120 h) for extreme 10 m wind speed than for extreme 2 m air temperature. However, Haiden et al. (2019) do not mention this aspect when referring to their figure. Relatively poor model performance for wind speed in the Antarctic has also been found for AMPS by Bromwich et al. (2005), who attributed this finding mainly to complex topography. They found better model performance for wind for areas with less complex topography and also in the free atmosphere, where the topographic influence is reduced. Indeed, all stations in our study are located in or near complex topography (Turner & Pendlebury, 2004) and this could presumably at least in part explain the relatively poor performance for this variable. An exception is Polarstern, which was located over the sea-ice covered Weddell Sea. For this station, the NWP systems' 24–48 h forecasts generally perform better than for the other stations when considering temperature and specific humidity, but worse for wind speed, at least in terms of *r* (Figure 2). The same is largely true when comparing the values of *r* for Polarstern for each individual forecast hour to the corresponding values averaged over the other (coastal) stations (see Figure S7).

This would seemingly contradict the argument about the dominating effect of complex topography. In the case of wind speed, the dominating effect may be the higher cyclone activity in the Weddell Sea compared to its coasts (Uotila et al., 2011; their Figures 3 and 4). Interestingly, Køltzow et al. (2019) obtained opposite results for the Atlantic sector of the Arctic: the short-term forecasts for 10 m wind were better over the ocean (evaluated against ASCAT data) than over land (against SYNOP). They attributed the difference to representativeness issues of SYNOP wind observations.

The main results of the upper-air error statistics and their sensitivity to forecast lead time are summarized in Figures S8–S10.

4. Conclusions

In this study, we have presented the first evaluation of NWP systems over the Southern Ocean, which also addresses the accuracy of forecasted vertical profiles, while paying particular attention to regional patterns of error statistics. We have compared the analyses and 12–120 h forecasts produced by three NWP systems, ECMWF, GFS, and AMPS, against in situ near-surface and upper-air observations of air temperature, specific humidity, and wind speed. These reference data are from four coastal and one ship-based AWS and radiosonde stations in the Weddell Sea area from the period June 20 to August 9, 2013.

In terms of near-surface biases, we found that ECMWF and AMPS are generally biased cold and dry, and GFS is biased warm and moist. ECMWF and GFS on average underestimate the near-surface wind speed, while AMPS overestimates it. The quality of the forecasts varies spatially. Across all three NWP systems and all five stations, we found the largest cold biases when comparing GFS against the near-surface data from Novolazarevskaya. We attribute this to under-representation of adiabatic heating during events of katabatic wind. Furthermore, we found relatively large underestimations of the near-surface and lower atmosphere wind speed when comparing ECMWF and GFS against the stations Novolazarevskaya and Neumayer. We relate these biases to under-representation of barrier flow. However, there are also negative wind speed biases, in particular in GFS, at other stations, where effects such as barrier flows are not as pronounced. This points to additional error sources, related for example to boundary layer parameterizations. Regarding the positive wind speed biases in AMPS, previous studies have found similar biases and have attributed this rather systematic model error to the Mellor-Yamada-Janjić (MYJ) boundary layer scheme.

In terms of both near-surface and upper-air data, we found the generally poorest model performance for Marambio, probably caused by the Marambio Island not being resolved by the NWP systems' physiographic fields and challenges in representing the flow around the Antarctic Peninsula. Furthermore, fewer radiosoundings were available for assimilation by the model systems from Marambio than the other stations during the study period. The NWP systems compare generally the most favorably against the observations from Polarstern made in the sea ice zone, where there is no complex terrain.

Considering overall performance of the NWP systems, ECMWF has generally the most favorable error statistics for near-surface temperature, humidity, and wind speed for the analysis and each lead time hour. AMPS is second best in this regard and GFS is third. The same conclusions can largely be drawn based on the upper-air data. The good performance of ECMWF is probably partly related to its 4D-Var assimilation scheme, which is expected to yield advantage over the other models' 3D-Var schemes. Moreover, the ECMWF model has a much higher number of vertical model levels than GFS and AMPS, and a higher horizontal resolution than GFS, which is important for regions of complex orography.

Among the near-surface variables, specific humidity is on average forecast the best in all three NWP systems, temperature is forecast second best, and wind speed third. We also find that the error statistics (r and NRMSE) for wind speed degrade the most rapidly with increasing lead time.

We acknowledge that the model products evaluated herein are already some years old. Being operational systems, they have undergone regular updates and improvements. Since 2013, highlights from these updates include numerical grids with higher horizontal resolution in ECMWF (~16 km vs. ~9 km), GFS (~23 km vs. ~13 km), and also for the AMPS domains, including the second domain that we have used data from (10 km vs. 8 km). Also, the GFS assimilation scheme has been upgraded to 4D-Var. We note that the vertical resolution has remained the same in all three model systems. Whether the recent model updates have notably decreased the model errors, remains to be studied.

All in all, in particular taking the challenging conditions provided by the study area into account, we conclude that the error statistics presented herein are rather favorable for all of the models. It is non-trivial to make good forecasts for this region of Antarctica in winter, and the same applies to the observations.

Some of the documented model errors can be fairly robustly attributed for certain conditions, such as complex orography. We point out that even if the observation site is located on a flat ice shelf, as Halley and Neumayer, the forecast accuracy may still be affected by orography of the surroundings in spatial scales of tens of km. Some aspects in our results remain open. These include (i) reasons for the better forecast skills for 2 m specific humidity than 2 m air temperature, (ii) the large spatial patterns in the differences between the analyses and forecasts in AMPS, and (iii) the contributions of various factors (such as orography, cyclone activity, and spatio-temporal variations in the surface boundary conditions) to differences in forecast skills between the sea ice station (Polarstern) and the other stations. Further work is needed to solve these questions, and to improve NWP model forecasts for Antarctic winter conditions.

Data Availability Statement

The data sets used in this study are available at the websites indicated in the supporting information.

Acknowledgments
This study was supported by the Academy of Finland (Contracts 304345 and 308441).

References

Andreas, E. L. (2002). Near-surface water vapor over polar sea ice is always near ice saturation. *Journal of Geophysical Research*, *107*(C10), https://doi.org/10.1029/2000jc000411

Atlaskin, E., & Vihma, T. (2012). Evaluation of NWP results for wintertime nocturnal boundary-layer temperatures over Europe and Finland. *Quarterly Journal of the Royal Meteorological Society*, *138*(667), 1440–1451. https://doi.org/10.1002/qj.1885

Bromwich, D. H., Monaghan, A. J., Manning, K. W., & Powers, J. G. (2005). Real-time forecasting for the Antarctic: An evaluation of the Antarctic Mesoscale Prediction System (AMPS). *Monthly Weather Review*, *133*(3), 579–603. https://doi.org/10.1175/mwr-2881.1

Bromwich, D. H., Otieno, F. O., Hines, K. M., Manning, K. W., & Shilo, E. (2013). Comprehensive evaluation of polar weather research and forecasting model performance in the Antarctic. *Journal of Geophysical Research: Atmosphere*, *118*(2), 274–292. https://doi.org/10.1029/2012jd018139

Carrasco, J. F., Bromwich, D. H., & Monaghan, A. J. (2003). Distribution and characteristics of mesoscale cyclones in the Antarctic: Ross Sea Eastward to the Weddell Sea*. *Monthly Weather Review*, *131*(2), 289–301. https://doi.org/2.0.co;2">10.1175/1520-0493(2003)131<0289:dacomc>2.0.co;2

Chen, S.-Y., Wee, T.-K., Kuo, Y.-H., & Bromwich, D. H. (2014). An impact assessment of GPS radio occultation data on prediction of a rapidly developing cyclone over the Southern Ocean. *Monthly Weather Review*, *142*(11), 4187–4206. https://doi.org/10.1175/mwr-d-14-00024.1

Haiden, T., Janousek, M., Vitart, F., Ferranti, L., & Prates, F. (2019). *Evaluation of ECMWF forecasts, including the 2019 upgrade*. Technical Memorandum (Vol. 853). Reading, UK: ECMWF. https://doi.org/10.21957/mlvapkke

Hines, K. M., & Bromwich, D. H. (2008). Development and testing of Polar Weather Research and Forecasting (WRF) Model. Part I: Greenland Ice Sheet Meteorology. *Monthly Weather Review*, *136*(6), 1971–1989. https://doi.org/10.1175/2007mwr2112.1

Hines, K. M., & Bromwich, D. H. (2017). Simulation of late summer Arctic clouds during ASCOS with Polar WRF. *Monthly Weather Review*, *145*(2), 521–541. https://doi.org/10.1175/mwr-d-16-0079.1

Hines, K. M., Bromwich, D. H., Wang, S.-H., Silber, I., Verlinde, J., & Lubin, D. (2019). Microphysics of summer clouds in central West Antarctica simulated by the Polar Weather Research and Forecasting Model (WRF) and the Antarctic Mesoscale Prediction System (AMPS). *Atmospheric Chemistry and Physics*, *19*(19), 12431–12454. https://doi.org/10.5194/acp-19-12431-2019

Hogan, T. F., & Rosmond, T. E. (1991). The description of the navy operational global atmospheric prediction system's spectral forecast model. *Monthly Weather Review*, *119*(8), 1786–1815. https://doi.org/2.0.co;2">10.1175/1520-0493(1991)119<1786:tdotno>2.0.co;2

Inoue, J. (2020). Review of forecast skills for weather and sea ice in supporting Arctic navigation. *Polar Science*, 100523. https://doi.org/10.1016/j.polar.2020.100523

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*(10), https://doi.org/10.1029/2012gl051591

Jonassen, M. O., Chechin, D., Karpechko, A., Lüpkes, C., Spengler, T., Tepstra, A., et al. (2020). Dynamical processes in the Arctic atmosphere. In: A. Kokhanovsky, C. Tomasi (eds) *Physics and Chemistry of the Arctic Atmosphere*. Springer Polar Sciences (pp. 1–51) Cham. Switzerland: Springer. https://doi.org/10.1007/978-3-030-33566-3_1

Jones, P. D., & Lister, D. H. (2015). Antarctic near-surface air temperatures compared with ERA-Interim values since 1979. *International Journal of Climatology*, *35*(7), 1354–1366. https://doi.org/10.1002/joc.4061

Jones, R. W., Renfrew, I. A., Orr, A., Webber, B. G. M., Holland, D. M., & Lazzara, M. A. (2016). Evaluation of four global reanalysis products using in situ observations in the Amundsen Sea Embayment, Antarctica. *Journal of Geophysical Research: Atmosphere*, *121*(11), 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

Jung, T., & Matsueda, M. (2016). Verification of global numerical weather forecasting systems in polar regions using TIGGE data. *Quarterly Journal of the Royal Meteorological Society*, *142*(695), 574–582. https://doi.org/10.1002/qj.2437

King, J. C., & Turner, J. (1997). *Antarctic meteorology and climatology*, New York, NY: Cambridge Univ. Press. https://doi.org/10.1017/cbo9780511524967

Køltzow, M., Casati, B., Bazile, E., Haiden, T., & Valkonen, T. (2019). An NWP model intercomparison of surface weather parameters in the European Arctic during the year of polar prediction special observing period Northern hemisphere 1. *Weather and Forecasting*, *34*(4), 959–983. https://doi.org/10.1175/waf-d-19-0003.1

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. https://doi.org/10.1016/0266-9838(90)90021-w

Mateling, M. E., Lazzara, M. A., Keller, L. M., Weidner, G. A., & Cassano, J. J. (2018). Alexander Tall Tower! A study of the boundary layer on the Ross Ice Shelf, Antarctica. *Journal of Applied Meteorology and Climatology*, *57*(2), 421–434. https://doi.org/10.1175/jamc-d-17-0017.1

Mölders, N., & Kramm, G. (2010). A case study on wintertime inversions in Interior Alaska with WRF. *Atmospheric Research*, *95*(2-3), 314–332. https://doi.org/10.1016/j.atmosres.2009.06.002

Monaghan, A. J., Bromwich, D. H., Wei, H.-L., Cayette, A. M., Powers, J. G., Kuo, Y.-H., & Lazzara, M. A. (2003). *Performance of weather forecast models in the rescue of Dr. Ronald Shemenski from the south Pole in April 2001**, *18*(2), 142–160. https://doi.org/2.0.co;2">10.1175/1520-0434(2003)018<0142:powfmi>2.0.co;2

Müller, M., Batrak, Y., Kristiansen, J., Køltzow, M. A. Ø., Noer, G., & Korosov, A. (2017). *Characteristics of a convective-scale weather forecasting system for the European Arctic. Monthly weather review*, *145*(12), 4771–4787. https://doi.org/10.1175/mwr-d-17-0194.1

Nigro, M. A., Cassano, J. J., Wille, J., Bromwich, D. H., & Lazzara, M. A. (2017). A self-organizing-map-based evaluation of the Antarctic mesoscale prediction system using observations from a 30-m instrumented tower on the Ross ice shelf, Antarctica [Data set]. *Weather and Forecasting*, *32*(1), 223–242. https://doi.org/10.1175/WAF-D-16-0084.1

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

Pezza, A., Sadler, K., Uotila, P., Vihma, T., Mesquita, M. D. S., & Reid, P. (2016). Southern Hemisphere strong polar mesoscale cyclones in high-resolution datasets. *Climate Dynamics*, *47*(5-6), 1647–1660. https://doi.org/10.1007/s00382-015-2925-2

Powers, J. G., Monaghan, A. J., Cayette, A. M., Bromwich, D. H., Kuo, Y.-H., & Manning, K. W. (2003). Real-time mesoscale modeling over Antarctica: The Antarctic Mesoscale Prediction System*. *Bulletin of the American Meteorological Society*, *84*(11), 1533–1546. https://doi.org/10.1175/bams-84-11-1533

Puri, K., Dietachmayer, G., Steinle, P., Dix, M., Rikus, L., Logan, L., et al. (2013). Implementation of the initial ACCESS numerical weather prediction system. *Australian Meteorological and Oceanographic Journal*, *63*, 265–284. https://doi.org/10.22499/2.6302.001

Renfrew, I. A., & Anderson, P. S. (2002). The surface climatology of an ordinary katabatic wind regime in Coats Land, Antarctica. *Tellus A*, *54*(5), 463–484. https://doi.org/10.1034/j.1600-0870.2002.201397.x

Scarlat, R. C., Melsheimer, C.. & Heygster, G. (2018). *Retrieval of total water vapor in the Arctic using microwave humidity sounders. Atmospheric Measurement Techniques*, *11*(4), 2067–2084. https://doi.org/10.5194/amt-11-2067-2018

Schroeter, B. J. E., Reid, P., Bindoff, N. L., & Michael, K. (2019). Antarctic verification of the Australian numerical weather prediction model. *Weather and Forecasting*, *34*(4), 1081–1096. https://doi.org/10.1175/waf-d-18-0171.1

Speirs, J. C., Steinhoff, D. F., McGowan, H. A., Bromwich, D. H., & Monaghan, A. J. (2010). Foehn Winds in the McMurdo Dry Valleys, Antarctica: The origin of extreme warming events. *Journal of Climate*, *23*(13), 3577–3598. https://doi.org/10.1175/2010jcli3382.1

Stössel, A., Zhang, Z., & Vihma, T. (2011). The effect of alternative real-time wind forcing on Southern Ocean sea ice simulations. *Journal of Geophysical Research*, *116*(C11), https://doi.org/10.1029/2011jc007328

Tastula, E.-M., & Vihma, T. (2011). WRF model experiments on the Antarctic Atmosphere in winter. *Monthly weather review*, *139*(4), 1279–1291. https://doi.org/10.1175/2010mwr3478.1

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. https://doi.org/10.1175/mwr-d-12-00016.1

Turner, J., & Pendlebury, S. (2004). *The international Antarctic weather forecasting handbook*. British Antarctic Survey.

Uotila, P., Vihma, T., Pezza, A. B., Simmonds, I., Keay, K., & Lynch, A. H. (2011). Relationships between Antarctic cyclones and surface conditions as derived from high-resolution numerical weather prediction data. *Journal of Geophysical Research*, *116*(D7), https://doi.org/10.1029/2010jd015358

Valkonen, T., Vihma, T., Johansson, M. M., & Launiainen, J. (2014). Atmosphere-sea ice interaction in early summer in the Antarctic: evaluation and challenges of a regional atmospheric model: Antarctic Atmosphere-Sea Ice Interaction. *Quarterly Journal of the Royal Meteorological Society*, *140*(682), 1536–1551. https://doi.org/10.1002/qj.2237

Van Den Broeke, M. R., & Van Lipzig, N. P. M. (2013). Response of wintertime Antarctic temperatures to the Antarctic oscillation: Results of a regional climate model. In E. Domack, A. Levente, A. Burnet, R. Bindschadler, P. Convey, & M. Kirby (Eds.), *Antarctic Peninsula Climate Variability: Historical and Paleoenvironmental Perspectives*. Antarctic Research Series (pp 43–58). Washington, DC: AGU. https://doi.org/10.1029/ar079p0043

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

Vignon, É., Traullé, O., & Berne, A. (2019). On the fine vertical structure of the low troposphere over the coastal margins of East Antarctica. *Atmospheric Chemistry and Physics*, *19*(7), 4659–4683. https://doi.org/10.5194/acp-2018-1197-supplement

Vihma, T., Tuovinen, E., & Savijärvi, H. (2011). Interaction of katabatic winds and near-surface temperatures in the Antarctic. *Journal of Geophysical Research: Atmospheres*, *116*(D21), https://doi.org/10.1029/2010jd014917

Vihma, T., Uotila, P., Cheng, B., & Launiainen, J. (2002). Surface heat budget over the Weddell Sea: Buoy results and model comparisons. *Journal of Geophysical Research*, *107*(C2). https://doi.org/10.1029/2000jc000372

Wille, J. D., Bromwich, D. H., Cassano, J. J., Nigro, M. A., Mateling, M. E., & Lazzara, M. A. (2017). Evaluation of the AMPS boundary layer simulations on the Ross Ice Shelf, Antarctica, with unmanned aircraft observations. *Journal of Applied Meteorology and Climatology*, *56*(8), 2239–2258. https://doi.org/10.1175/jamc-d-16-0339.1

Wille, J. D., Bromwich, D. H., Nigro, M. A., Cassano, J. J., Mateling, M., Lazzara, M. A., & Wang, S.-H. (2016). Evaluation of the AMPS boundary layer simulations on the Ross Ice Shelf with tower observations. *Journal of Applied Meteorology and Climatology*, *55*(11), 2349–2367. https://doi.org/10.1175/jamc-d-16-0032.1

Wilson, A. B., Bromwich, D. H., & Hines, K. M. (2011). Evaluation of Polar WRF forecasts on the Arctic System Reanalysis domain: Surface and upper air analysis. *Journal of Geophysical Research*, *116*(D11), https://doi.org/10.1029/2010jd015013

Yang, S.-C., Corazza, M., Carrassi, A., Kalnay, E., & Miyoshi, T. (2009). Comparison of local ensemble transform Kalman filter, 3DVAR, and 4DVAR in a quasigeostrophic model. *Monthly weather review*, *137*(2), 693–709. https://doi.org/10.1175/2008mwr2396.1

Zhang, Z., Vihma, T., Stössel, A., & Uotila, P. (2015). The role of wind forcing from operational analyses for the model representation of Antarctic coastal sea ice. *Ocean Modelling*, *94*, 94–111. https://doi.org/10.1016/j.ocemod.2015.07.019