Evaluation of the Rapid Refresh Numerical Weather Prediction Model over Arctic Alaska

MATTHEW T. BRAY,[a] DAVID D. TURNER,[b] AND GIJS DE BOER[c,d]

[a] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[b] *Global Systems Laboratory, NOAA, Boulder, Colorado*
[c] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[d] *Physical Sciences Laboratory, NOAA, Boulder, Colorado*

(Manuscript received 15 September 2020, in final form 29 March 2021)

ABSTRACT: Despite a need for accurate weather forecasts for societal and economic interests in the U.S. Arctic, thorough evaluations of operational numerical weather prediction in the region have been limited. In particular, the Rapid Refresh Model (RAP), which plays a key role in short-term forecasting and decision-making, has seen very limited assessment in northern Alaska, with most evaluation efforts focused on lower latitudes. In the present study, we verify forecasts from version 4 of the RAP against radiosonde, surface meteorological, and radiative flux observations from two Arctic sites on the northern Alaskan coastline, with a focus on boundary layer thermodynamic and dynamic biases, model representation of surface inversions, and cloud characteristics. We find persistent seasonal thermodynamic biases near the surface that vary with wind direction, and may be related to the RAP's handling of sea ice and ocean interactions. These biases seem to have diminished in the latest version of the RAP (version 5), which includes refined handling of sea ice, among other improvements. In addition, we find that despite capturing boundary layer temperature profiles well overall, the RAP struggles to consistently represent strong, shallow surface inversions. Further, while the RAP seems to forecast the presence of clouds accurately in most cases, there are errors in the simulated characteristics of these clouds, which we hypothesize may be related to the RAP's treatment of mixed-phase clouds.

SIGNIFICANCE STATEMENT: Human activities continue to expand into northern high latitudes, including shipping and other commercial activities, energy exploration, tourism, and defense. Conducting these activities safely requires accurate weather forecasting tools capable of handling a harsh and complex environment. This work evaluates the performance of one of the primary weather forecasting models used in the United States for short-term decision-making, the Rapid Refresh (RAP) model, run operationally by the National Weather Service. This effort illustrates some promising results, while at the same time bringing some shortcomings to light. Importantly, it highlights some areas in which the RAP can improve to support northern communities and commercial entities through the development of accurate weather forecasts.

KEYWORDS: Arctic; Boundary layer; Model errors; Model evaluation/performance; Numerical weather prediction/forecasting

1. Introduction

Numerical weather prediction (NWP) plays many critical societal roles. For the last several decades, NWP models have played increasingly prominent roles in providing short-term (hours to days) and longer-term (days to seasons) weather forecasts. These forecasts serve as a source of information for industries ranging from agriculture and aviation to energy exploration, shipping, and tourism. As the needs of society have expanded and the role of NWP has become more prominent, there is an enhanced need for evaluation and improvement of the tools used to create these critical forecasts. This fact is particularly true in parts of the world where societal activities and observations have traditionally been limited in scope, but where there are increasing demands for support in the form of weather forecasts to help support increased activity as a result of rapid changes in local climate.

The Arctic is a primary example of a region that has traditionally not seen extensive NWP verification but is seeing elevated activity given warming that has occurred over recent decades. While some Arctic model evaluations (mainly climate models and reanalysis products) have been conducted leveraging limited observations from the region (e.g., Tjernström et al. 2008; Wilson et al. 2011; de Boer et al. 2012, 2014; Lindsay et al. 2014; Wesslén et al. 2014), evaluations of operational weather forecast models have been sparse (e.g., Beesley et al. 2000; Sotiropoulou et al. 2016). This lack of thorough evaluation is despite elevated risk associated with activities undertaken in a region that features extreme meteorological phenomena related to cold weather, aircraft icing, winds, and blowing snow.

The model evaluated in the current study, the Rapid Refresh (RAP) NWP system, has been run operationally at the National Centers for Environmental Prediction (NCEP) since 2012 and serves as a primary source of information for the U.S. National Weather Service and other key entities for hour-to-hour decision making (Benjamin et al. 2016). The rapidly updating RAP forecasts are unique in the Arctic, where other NWP systems update at much slower cadences (e.g., every 6 h instead of RAP's hourly updates). Despite limited evaluation of the RAP in the Arctic, one recent study (Pinto et al. 2020)

Corresponding author: Matthew Bray, matthewbray1@ou.edu

DOI: 10.1175/WAF-D-20-0169.1

evaluated RAP performance in Alaska. In that effort the accuracy of ceiling and visibility forecasts from version 4 of the RAP (hereafter RAPv4) were assessed relative to 160 operational observing sites that measure visibility and ceiling height. They found that RAPv4 overforecast the frequency of ceiling-impacted aviation operations in the coastal regions of Alaska. However, they did not review the accuracy of the thermodynamic profiles predicted by the model, nor the accuracy of the winds or downwelling longwave radiative flux. There is a strong interplay between the surface and the overlying atmosphere in the stable conditions frequently observed in the Arctic (Steeneveld et al. 2006; Holtslag et al. 2013), and errors in model-represented interactions between the surface and boundary layer could easily affect the evolution of clouds and hence the predicted ceiling heights. A separate study looked at the high-resolution cousin of the RAP, comparing the 3-km output from the High-Resolution Rapid Refresh (HRRR) against pressure observations across Alaska (McCorkle et al. 2018). They found that the prediction of surface pressure for all forecast hours matched the unassimilated pressure observations well generally, though there was a general underestimation of rapid increases in temperature and wind speeds during the onset of downslope wind events.

As mentioned earlier, relatively few Arctic locations collect both thermodynamic and cloud observations. A limited consortium of observing sites, the International Arctic Systems for Observing the Atmosphere (IASOA; Uttal et al. 2016) include long-standing observational records of basic atmospheric variables such as temperature, humidity, pressure, winds, cloud cover, radiation, turbulent fluxes and more, both at the surface and throughout the atmospheric column. They span locations in the United States, Canada, Greenland, Russia, and the European Arctic. In the current paper, we use data from two Alaskan IASOA observatories to evaluate the ability of the RAPv4 to simulate local surface meteorology, profiles of temperature, humidity, and wind in the lowest few kilometers near the surface, and downwelling radiative fields, as a function of forecast lead time. This analysis uses nearly two years of RAPv4 data, allowing us to statistically evaluate the ability of the model to capture seasonal dependence of atmospheric evolution. Furthermore, since the two sites are coastal, we also evaluate the ability of the model to predict atmospheric conditions when prevailing winds are from both inland and oceanic regions. To our knowledge, this study provides the first long-term characterization of the thermodynamic structure of the atmosphere in any version of the RAP in the U.S. Arctic.

2. Data and methods

a. Model background

As briefly discussed above, the RAP, developed by the National Oceanic and Atmospheric Administration (NOAA) in association with partner institutes, is initialized hourly, computing 21-h forecasts hourly, and 39-h forecasts initialized at 0300, 0900, 1500, and 2100 UTC (these forecast lengths are for version 4 of the model). The RAP domain covers the entire North American continent, as well as the majority of the Arctic region. The RAP, including its data assimilation system, model physics, and postprocessing components, are fully described in Benjamin et al. (2016).

In the current study, we analyze output from RAPv4, which was the operational version of the model when this evaluation effort started. As such, preliminary results of this analysis have been presented to model developers to guide improvements incorporated into version 5 (hereafter RAPv5). RAPv4 is based upon v3.8.1+ of the Advanced Research version of the Weather Research and Forecasting (WRF-ARW) Model (Skamarock et al. 2008). This regional model uses a rotated latitude–longitude grid with 13-km horizontal grid spacing and 50 vertical layers. It uses a sigma-isobaric vertical coordinate, with the first level at (typically) 10 m AGL and model top at 10 hPa; this coordinate results in four model levels below 200 m and eight below 1 km. Model boundary conditions are provided by the operational Global Forecast System (GFS), also run at NCEP. Observations are assimilated into the model upon its hourly reinitializations using the Gridpoint Statistical Interpolation (GSI) hybrid analysis system (Whitaker et al. 2008; Kleist et al. 2009), wherein the ensemble has a weight of 0.85 relative to the static background covariance matrix. A digital filter is used at the end of the data assimilation process to reduce some of the artifacts that result from the assimilation process (Peckham et al. 2016). RAPv4 model physics have been updated as a unified suite for many years, and improvements to the physics scheme are incorporated into the WRF-ARW code repository. The exact components used by the RAPv4 are provided in Table 1.

b. Observations

The observations used in this study were collected at two northern-Alaskan observational sites operated by the U.S. Department of Energy (DOE) Atmospheric Radiation Measurement (ARM) program. These sites include the North Slope of Alaska site (NSA; Verlinde et al. 2016), located in Utqiagvik (71.323°N, 156.609°W, 3 m MSL), and the long-term deployment of the ARM third mobile facility (AMF-3; see de Boer et al. 2018, 2019) at Oliktok Point, Alaska (OLI, 70.49°N, 149.89°W, 4 m MSL). For this evaluation of the RAP, we primarily focused on three main types of observations, including measurements of surface meteorology (temperature and humidity at 2 m AGL, and winds at 10 m AGL), temperature, humidity, and wind profiles from radiosondes launched from these observational sites, and downwelling longwave radiative fluxes. All parts of this study use observations from the period from 1 July 2017 to 31 May 2019.

Radiosondes are launched twice daily at OLI (0000 and 1800 UTC) and four times daily at NSA (0000, 0600, 1200, and 1800 UTC). Downwelling longwave radiation is available from upward-pointing pyrgeometers. The radiation data leveraged in the current study have been subjected to quality control and postprocessing as through the QCRAD ARM Value-Added Product (VAP; Long and Shi 2006). The QCRAD methodology uses climatological analyses of surface radiation to define reasonable limits for testing data for outliers. More information on obtaining the ARM data products used for this analysis—radiosondes, surface meteorology, radiation, and ceilometer data—can be found in the data availability

TABLE 1. RAPv4 physical parameterization details.

Physical component	Parameterization scheme
Cloud microphysics	Thompson aerosol-aware double moment (Thompson and Eidhammer 2014)
Cumulus	Grell and Freitas (2014), shallow component enabled
Planetary boundary layer	Mellor–Yamada–Nakanishi–Niino, modified to include an eddy diffusivity mass flux component (MYNN-EDMF; Olson et al. 2019)
Land surface	Rapid Update Cycle land surface model (RUC LSM; Smirnova et al. 2016), which features nine levels and includes extensive work to treat frozen soils and snow-covered conditions
Radiation	Rapid Radiative Transfer Model (RRTMG; Mlawer et al. 1997; Iacono et al. 2008)
Sea surface temperature and sea ice fraction	GFS analysis

statement (Keeler et al. 2015; Kyrouac and Holdridge 2015; Riihimaki et al. 2019; Morris and Ermold 2015, respectively).

To provide context for the verifications with the ARM data, we will first compare RAP output against operational radiosonde and surface observation station data across Alaska, using the Model Analysis Tool Suite (MATS; Turner et al. 2020). MATS provides the ability to look at the model statistics as a function of time, height, and forecast length; however, due to the need to keep this verification database size manageable, it averages the statistics over all of Alaska (see Fig. 1 for locations included in this average).

c. Data analysis and cloud detection techniques

To compare RAPv4 forecasts and observations directly, observations were averaged to the resolution of the model, using model output from the closest overland model grid point to the observation site. This includes vertical profiles from radiosondes, which were averaged to the RAP's vertical resolution, as well as surface measurements, radiation measurements, and cloud observations, which were averaged over 60-min windows centered around the RAP's hourly output times. Cases were also divided seasonally, with summer defined as May, June, and July and winter defined as November, December, and January; these seasonal definitions were selected so that the winter cases include almost no downwelling shortwave radiation, while summer cases could have up to 24 h of sunlight. For the portion of the analysis focusing on surface temperature inversions, a case was considered an inversion only if the temperature began increasing immediately above the surface.

While the RAPv4 includes the treatment of subgrid-scale clouds (including on the radiative fields), the properties (e.g., height, water content) of these subgrid-scale clouds are not included in the standard output from the model. Due to this gap in available data, downwelling longwave radiation at the surface was used as a proxy for the presence of a cloud in both the model (resolved plus subgrid-scale) and observations. While the RAPv4 output includes the downwelling longwave and shortwave irradiance values for the all-sky condition within the model, clear sky irradiance calculations were needed to help determine the presence of clouds. Thus, for each case, clear sky downwelling longwave radiation values were calculated using RAPv4 profiles of temperature, humidity, and pressure within a stand-alone version of the Rapid Radiative Transfer Model (RRTM; Mlawer et al. 1997). Clouds were determined to exist if the downwelling infrared flux was 10 W m^{-2} greater than the clear sky value; the same logic was used for both the observations and model output. This value is derived from the approximate uncertainty of the pyrgeometers used to take longwave radiation measurements. After determining cloud presence, positive cases (i.e., those where a cloud was present in both the model and the observations) are separated out for further analysis. An approximate cloud base temperature is calculated using radiosonde measurements interpolated to the time of the event and to the height of the cloud base, as measured by a ceilometer.

3. Results

a. Alaska-wide biases

To provide context, a brief analysis of RAP performance across the entire state of Alaska using the MATS output was conducted. We find an average warm bias of less than half a degree is present in the lowest 975 hPa, while cold biases of up to half a degree occur higher in the boundary layer averages (Fig. 2a). To evaluate moisture, we use model relative humidity

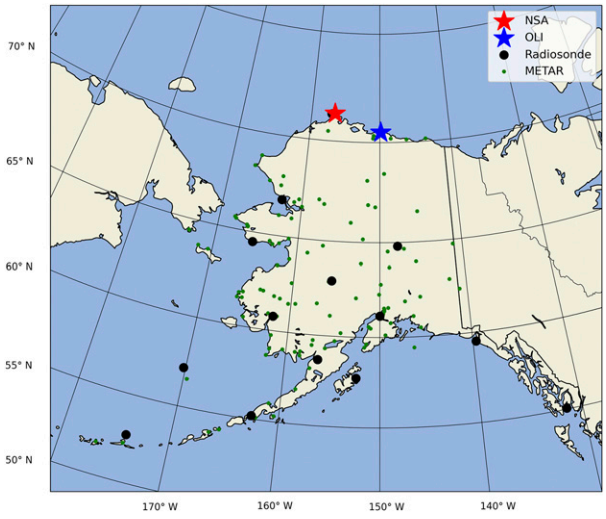

FIG. 1. Locations of the two ARM North Slope observing sites used in this study (stars), along with the radiosonde sites (black) and METAR sites (green) utilized within the MATS Alaska evaluation domain.

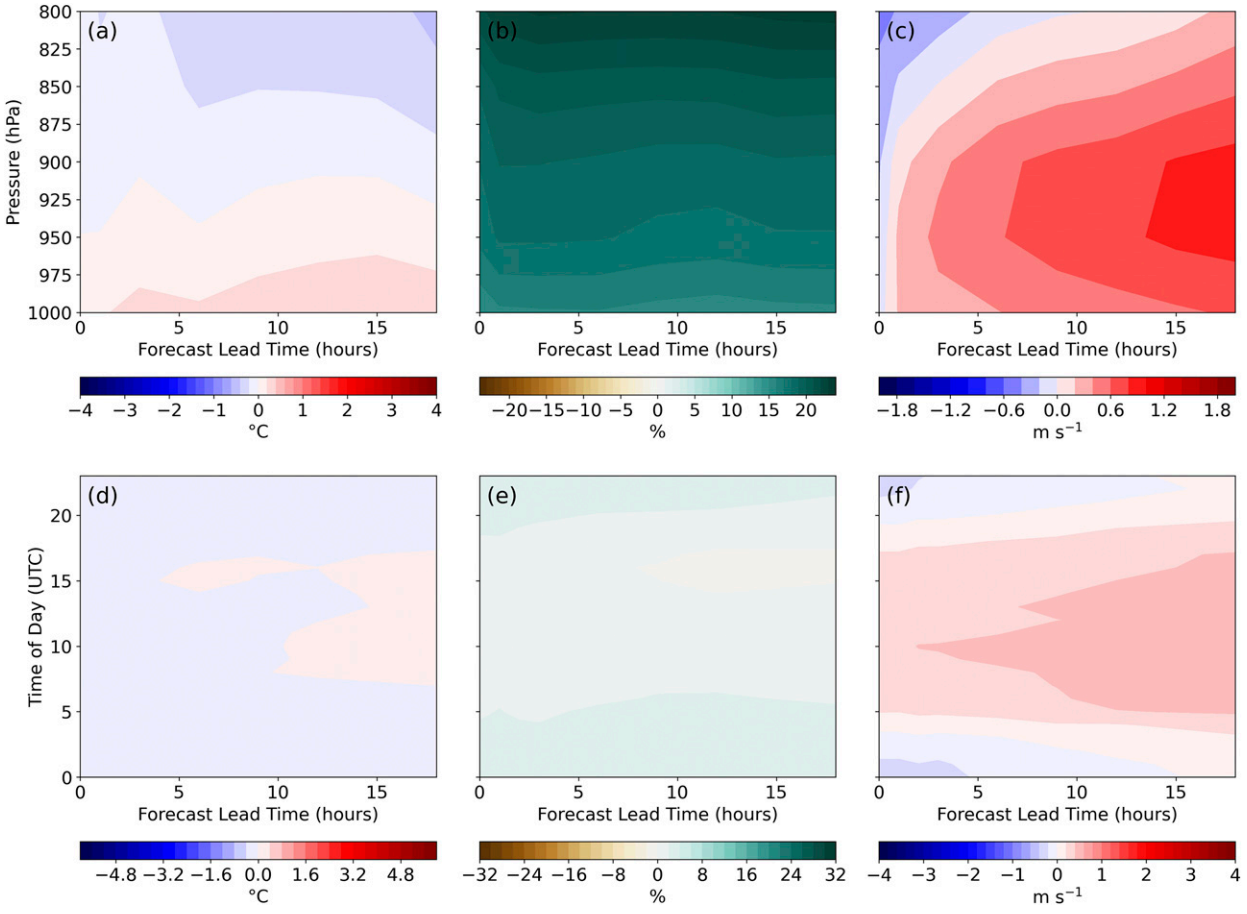

FIG. 2. (a) RAP temperature bias (°C) relative to radiosondes averaged across all observation sites in Alaska from August 2018 through July 2019 as a function of height and forecast lead time. (b) As in (a), but for RHobT bias (%). (c) As in (a), but for wind bias (m s^{-1}). (d) RAP bias relative to observations in the 2-m temperature (°C) as a function of time of day (UTC; *y* axis) and forecast lead time (h; *x* axis) averaged across all observing sites in Alaska from August 2018 through July 2019. (e) As in (d), but for 2-m RHobT (%). (f) As in (d), but for 10-m wind speed (m s^{-1}). Note that during the summer, the mean solar noon occurs around 2200 UTC.

output calculated using the model specific humidity and the observed temperature profile (also known as RHobT), which is designed to exclude an interconnection with temperature biases while providing physically interpretable results. Using RHobT, a moist bias ranging from around 10% near the surface to over 20% at 800 hPa is clear; given that this is averaged across the all of the varied Alaska observation sites, this may represent a systematic bias within the RAP output (Fig. 2b). Winds show a positive speed bias of up to 1 m s^{-1}, maximized around 925 hPa, which could indicate a persistent overestimation of low-level jets across the Alaska domain (Fig. 2c). In both the temperature and wind cases, biases increase over time, as would be expected from errors arising due to model physics errors.

At the surface, fairly small biases across Alaska are present across all three evaluated model fields. Averaged temperature and humidity errors at the surface are negligibly small when averaged across the Alaska domain, though a hint of a diurnal cycle in the model biases is evident, which becomes larger with the forecast length (Figs. 2d,e). This diurnal cycle is clearer in the wind biases, which reach a maximum of just over 0.5 m s^{-1} when averaged (Fig. 2f). This maximum occurs at around 1000 UTC, or around 0100 in the morning local time.

Of course, these boundary layer errors (Fig. 2) are averaged across all Alaska observing locations, which includes a variety of geographical conditions, so one would not necessarily expect strong bias patterns; however, it is evident from this analysis that systematic biases are not generally present in the RAP (outside of, perhaps, the boundary layer moisture profiles shown in Fig. 2b) across all of Alaska, so that any errors are more likely attributable to local conditions. Indeed, comparing bias profiles at a forecast lead time of six hours for the averaged Alaska domain and the observations from the NSA and OLI observing sites alone, significant differences in the boundary layer temperature, humidity, and wind bias patterns appear (Fig. 3). For this reason, a comprehensive spatial analysis of model biases would be very difficult, and so focusing on specific subregions of the Alaska domain will produce more useful results.

For the remainder of this study, we focus on model biases along the North Slope of Alaska. As discussed above, model

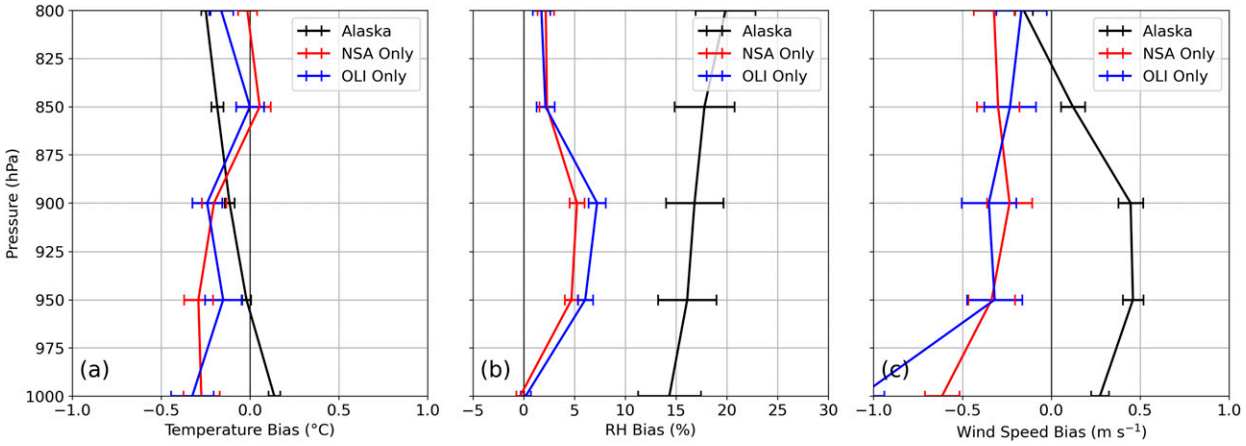

FIG. 3. (a) RAP temperature bias (°C) relative to radiosondes averaged across all observation sites in Alaska (black), across all NSA observations (red), and across all OLI observations (blue) from August 2018 through July 2019 as a function of height at a lead time of 6 h. (b) As in (a), but for RHobT (%). (c) As in (a), but for wind bias (m s^{-1}). Statistical confidence of 95% calculated with a *t* test is indicated by the horizontal error bars.

errors across the Alaska domain are likely influenced predominately by local factors, and the North Slope region has several geographical features unique to the Arctic. The North Slope sites are along the coast of the Arctic Ocean, which varies between frozen and unfrozen states throughout the year. Moreover, previous field campaigns have shown a large number of mixed-phase clouds that contain supercooled liquid water across the North Slope (e.g., Verlinde et al. 2007), which have proven difficult for modeling systems to represent properly (e.g., Klein et al. 2009). Each of these factors present modeling challenges unique to the Arctic, and the availability of high-resolution observations of radiative fluxes make these sites especially valuable in verifying the performance of the RAP.

In addition to these purely scientific considerations, the North Slope of Alaska warrants detailed study due to the increasing need for accurate weather forecasts in the region. With the projected expansion of the energy and fishing industries, along with increased tourism in the coming years, reliable forecast products are essential to the safe execution of these activities in future years. Combined with the specific challenges of modeling weather in the region described above, this makes the North Slope an excellent candidate for a detailed bias analysis.

b. *Lower atmospheric biases*

Although observations from both NSA and OLI were analyzed, henceforth only results from the OLI site are shown. The performance of the RAPv4 was consistent across both locations (see the bias profiles in Fig. 4, for example), and so the results provided for OLI in the next three subsections may be taken as representative of the both sites. Additionally, while data from the NSA site is assimilated into the RAP (see Fig. 1), soundings from the OLI site are not. Thus, OLI provides a more independent evaluation of model performance. The observations from the OLI site (using both the 1800 and 0000 UTC soundings) are compared with the closest RAP grid point that is over land, unless otherwise mentioned.

1) SEASONAL MODEL ERRORS

An initial overview of model performance in the lower atmosphere (defined here as up to 1500 m above ground level) reveals several patterns of model bias. In both summer and winter, temperature root-mean-square error (RMSE) in the RAPv4 is largest near the surface. While this near-surface error remains roughly constant with respect to forecast lead time in the summer months, the magnitude of the RMSE increases rapidly with lead time in the winter, pointing to a seasonal cause for these errors (Figs. 4a,d). Above 400 m, summer RMSE decreases with height for all forecast lead times, although the longer lead times have larger errors overall. In the winter, RMSEs above 400 m are roughly constant around 2°C, with slightly larger values at longer lead times. In short, forecast lead time seems to have a greater influence on RMSE aloft than near the surface for summer months, and near the surface than aloft for winter months.

On the other hand, the RMSE in relative humidity (RHobT) tends to increase with height in both seasons, though this pattern is especially pronounced in the summer (Figs. 4b,e). This humidity error is present from initialization and only increases by a few percent with forecast lead time, and, in the summer cases in particular, peaks at just under 1000 m AGL. This lack of forecast hour dependence is likely indicative of a model initialization problem or could be related to model errors in the representation of low-level clouds, which will be explored in more detail in section 3d below. Wind errors below 1500 m are fairly consistent across the seasons, with RMSE values between 2 and 4 m s^{-1} (increasing with forecast lead time) in both summer and winter (Figs. 4c,f).

2) SEASONAL WIND REGIME BIASES

Both the OLI and NSA sites are located on the northern coast of Alaska, with the Arctic Ocean to the north. While there are some coastal differences between the OLI and NSA sites, generally northerly winds will be from over the ocean (which is generally covered with sea ice between late October

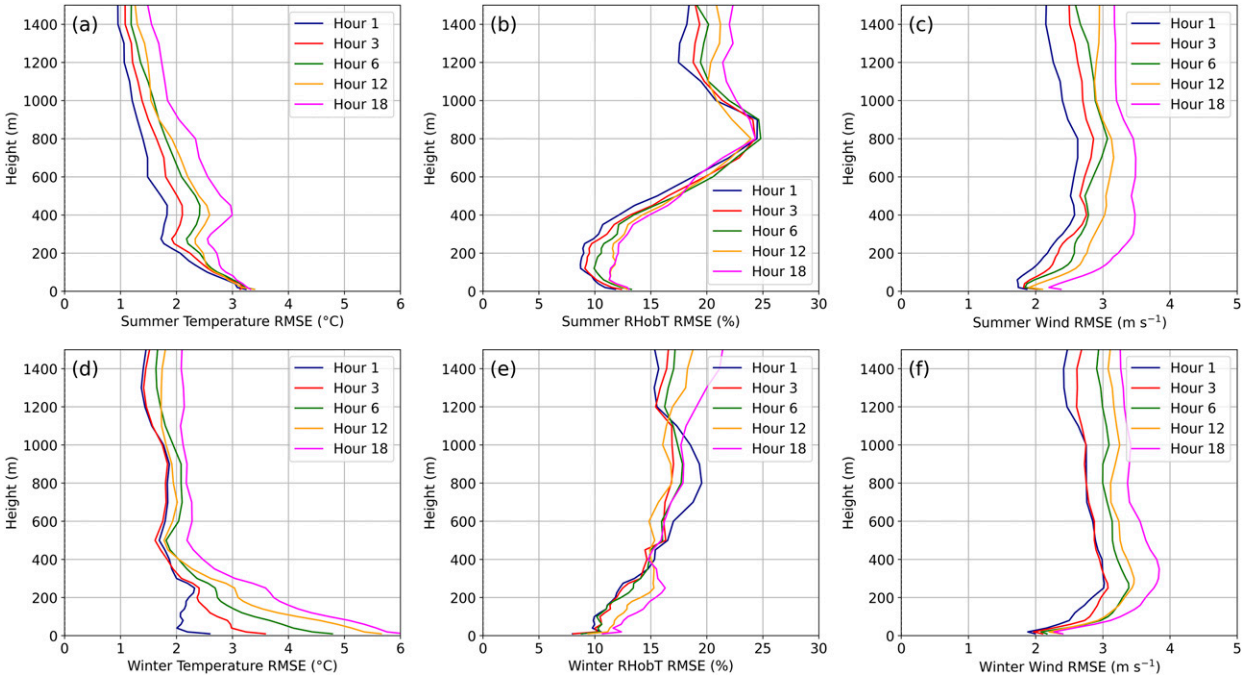

FIG. 4. (a) Profiles of temperature root-mean-squared error (°C) between the RAP and radiosonde observations at Oliktok Point (OLI) for several forecast lead times, limited to summer cases (defined here as May, June, and July). (b) As in (a), but for relative humidity (%), calculated with respect to the observed temperature (RHobT) for both the model and observations. (c) As in (a), but for wind speed ($m\ s^{-1}$). (d)–(f) As in (a)–(c), respectively, but for the winter cases (defined here was November, December, and January).

and mid-June), while southerly winds will be from over land (which may be snow covered between October and May). In cases where observed winds are northerly (from over the ocean), the RAPv4 exhibits a warm bias near the surface in the summer and a cold bias near the surface in the winter (Figs. 5a,b). The summer warm bias, generally confined to the lowest 500 m, is strongest directly above the surface (around 2°C) and remains roughly constant from model initialization through forecast hour 18 (Fig. 5a). However, the cold bias near the surface in the winter is small at model initialization but steadily intensifies and deepens throughout the forecast, eventually reaching a magnitude of around 4°C and a depth of around 500 m at forecast hour 18 (Fig. 5b). Similar patterns are present for each season in the humidity profiles. In the summer, a moist bias is present from initialization onward throughout the entire layer, though it is strongest at up to 10% near the surface (Fig. 5c). On the other hand, in the winter, a dry bias emerges near the surface during the model run and intensifies over time to a maximum of almost 20% (Fig. 5d).

In cases where the prevailing winds are from the south (over land), the biases exhibit similar vertical patterns to the northerly wind cases but with a reduced magnitude (Fig. 6). A warm, moist bias is present from initialization in these cases in the summer but the bias in both temperature and humidity is about half the size (1°C and 5%) of the bias as in the summertime northerly wind cases (cf. Figs. 6a,c against Figs. 5a,c). Similarly, the model develops a cold, dry bias gradually with forecast lead time in the winter but the maximum magnitude when the winds are from the south (Figs. 6b,d) is only one fourth of the magnitude (1°C and 5%) of the bias of the northerly wind counterpart (Figs. 5b,d). Outside of these notable near-surface patterns, lower atmospheric thermodynamic biases are fairly small (less than 1°C or 5%) across both seasons and wind directions, including temperature and humidity above 500 m (Figs. 5 and 6).

A pattern of wind biases, though generally small in magnitude ($<1\ m\ s^{-1}$), is present for cases with northerly winds in both seasons (Figs. 5e,f). In the summer, a slight negative bias of around $0.6\ m\ s^{-1}$ is present between 50 and 200 m from forecast hour 1 onward (Fig. 5e). This slow bias near the surface is also present in the winter, though it is weaker at $0.4\ m\ s^{-1}$, along with a $0.6\ m\ s^{-1}$ positive bias from 200 to 500 m and a negative bias throughout the rest of the lower atmosphere (Fig. 5f). In cases with southerly winds, the seasonal dependence is smaller, as both summer and winter tend to exhibit negative wind biases of around $0.5\ m\ s^{-1}$ throughout the lower atmosphere (Figs. 6e,f). Interestingly, in the winter, the model begins with a $0.2\ m\ s^{-1}$ positive wind speed bias under 500 m, before transitioning to a negative bias, indicating that the model physics likely play a role in development of these biases (Fig. 6f).

3) A COMPARISON OF RAPV4 AND RAPV5

The differences in the temperature, humidity, and wind biases between the northerly wind cases versus southerly wind cases, compounded with the seasonal differences, suggests that biases observed during times with northerly winds are associated with the ocean surface. In particular, it appears that errors

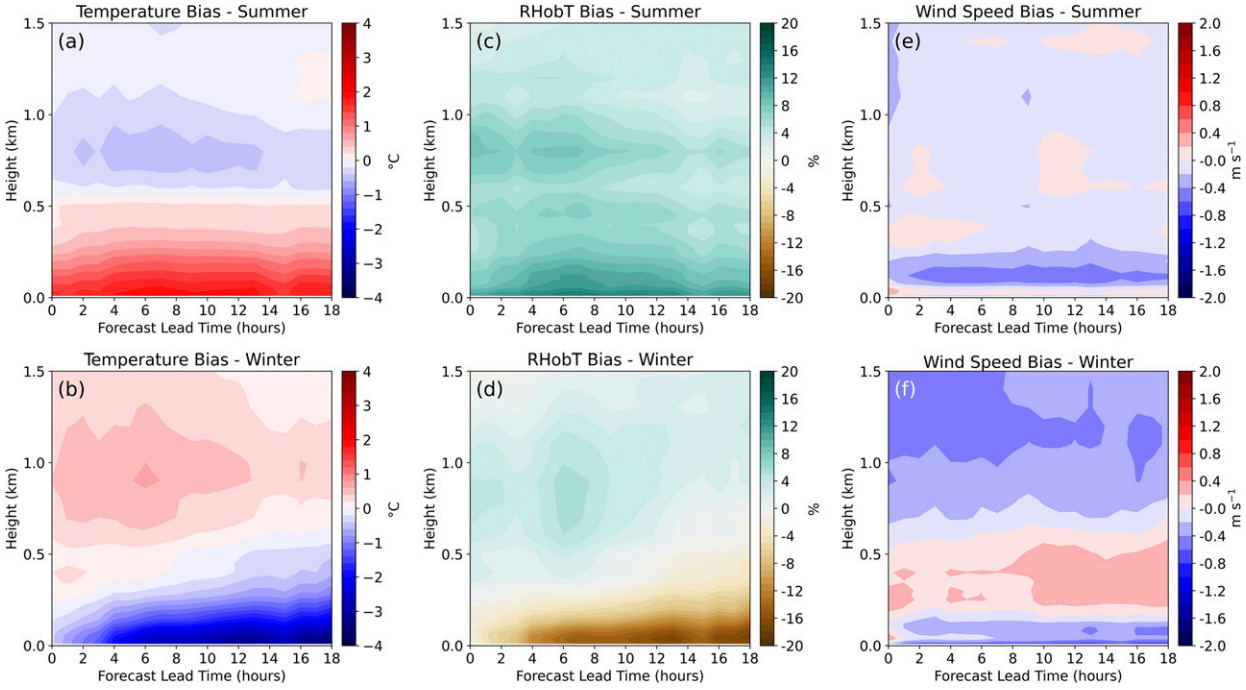

FIG. 5. RAP profiles of (a),(b) bias in temperature (°C), (c),(d) RHobT (%), and (e),(f) wind speed (m s^{-1}) relative to radiosondes at Oliktok Point (OLI) as a function of height and forecast lead time for observed northerly wind events during the (top) summer and (bottom) winter.

in the sea ice treatment within the RAPv4 are contributing to the biases seen over the lowest 500 m, given their strong correlation with times when winds are from the north during winter. Our analysis was performed while RAPv5 was being developed, and although there were several updates associated with the new version, one of the most pertinent changes for the Arctic is RAPv5's support for fractional sea ice coverage. To compare the two versions and investigate the potential impact

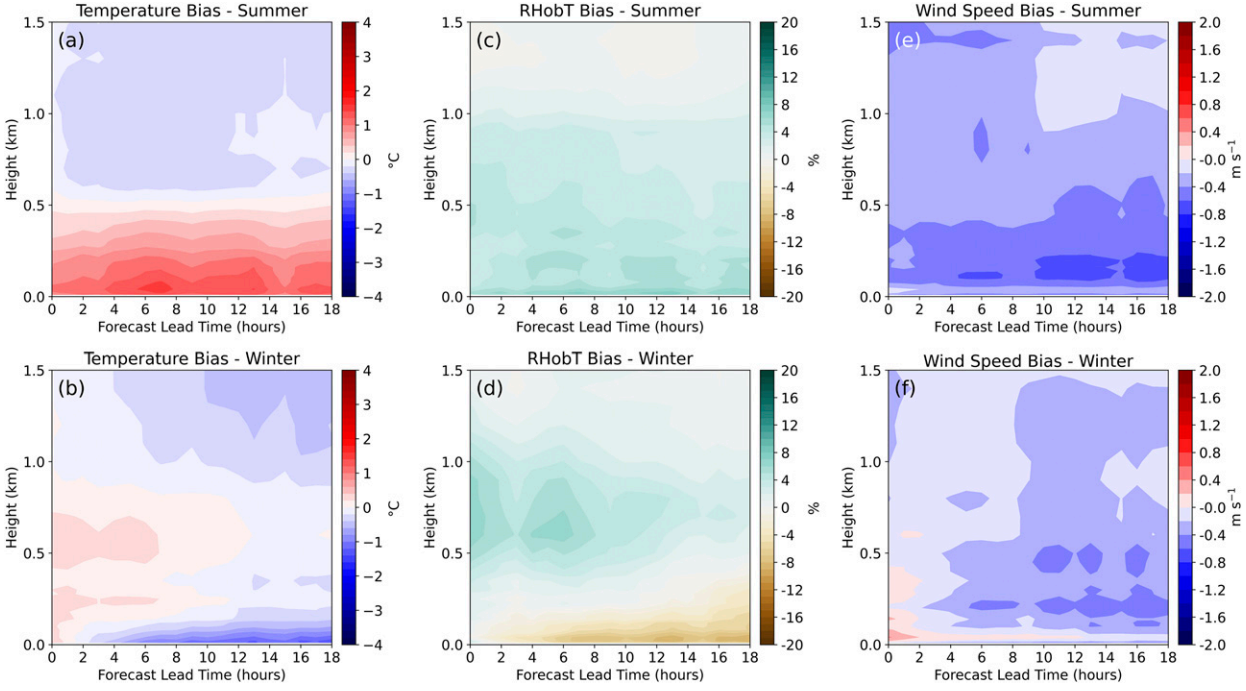

FIG. 6. As in Fig. 5, but for observed southerly wind events.

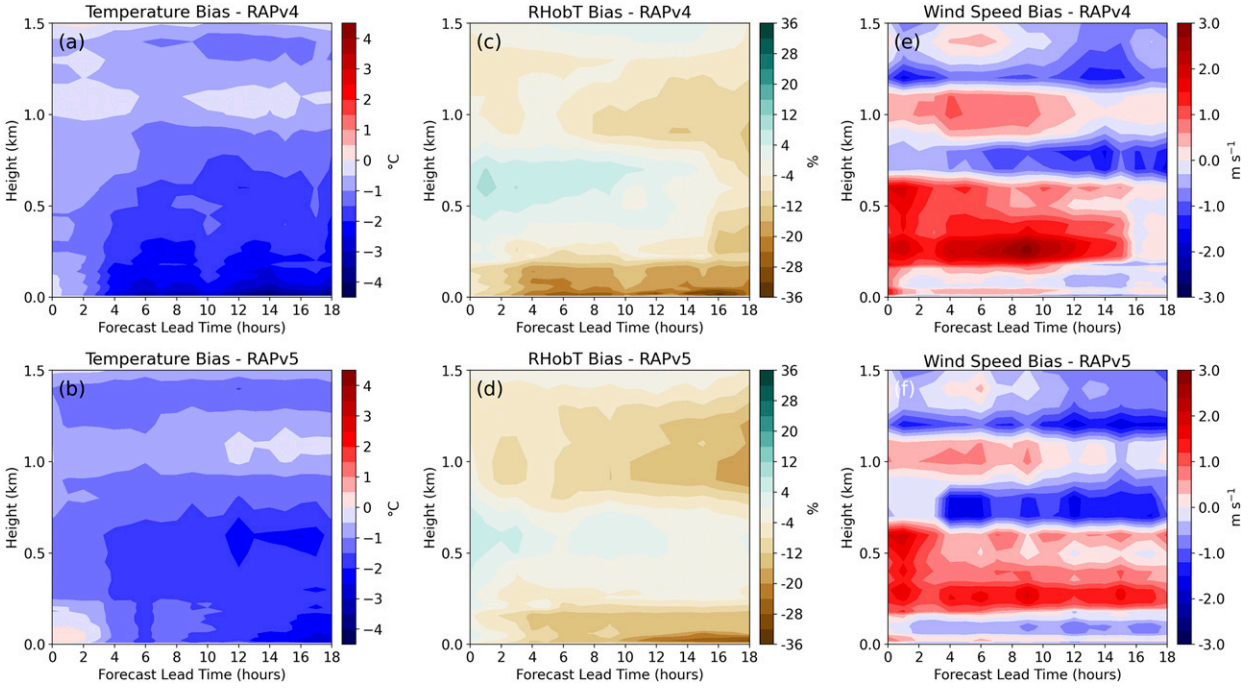

FIG. 7. The comparison of (a),(b) RAP temperature (°C), (c),(d) RHobT (%), and (e),(f) wind speed (m s^{-1}) for (top) RAPv4 and (bottom) RAPv5 for northerly wind events at Oliktok Point (OLI) for February 2019.

of fractional sea ice, we examine data from a retrospective run[1] of RAPv5 for February 2019 during times when winds were from the north (Fig. 7). Close to the surface, RAPv4 reveals patterns consistent with those provided in Figs. 5b, 5d, and 5f, with a cold, dry bias that increases with time up to around 4°C and 35% (Figs. 7a–c). Above 500 m, the February 2019 remains cold and dry, in contrast to the warm, moist biases noted above this level in Figs. 5b and 5d, which is likely a product of the specific weather conditions that occurred during this one month. Although the RAPv5 exhibits the same bias patterns, the magnitudes of the biases are reduced in the northerly wind cases, with the magnitude of the near-surface temperature bias closer to 2°C (Fig. 7). The dry RHobT bias in the lowest 200 m is also reduced to around 10% in RAPv5 compared to RAPv4 (Figs. 7c,d). This suggests that the inclusion of sparse openings in the sea ice to the north of Oliktok likely improves the representation of heat and moisture fluxes over this region. This, in turn, allows warmer, moister air to be resolved over the ice and helps to alleviate downwind (inland) biases during periods of onshore flow. Moreover, the average modeled upward sensible heat flux over all of February 2019 at the ocean grid point nearest to Oliktok Point is 2.3 W m^{-2} greater in the RAPv5 than in the RAPv4, a 15% increase between model versions. This lends additional support to the fractional sea ice hypothesis, as pockets of open water in the RAPv5 would allow for a greater upward sensible heat flux. Wind biases for the February 2019 northerly wind cases follow a layered structure, similar to the pattern seen in Fig. 5f, with magnitudes of up to 2 m s^{-1} (Figs. 7e,f). These biases show at best modest reduction in the RAPv5, indicating some lingering dynamical errors during periods of onshore flow (Figs. 7e,f).

Cases with southerly winds, however, exhibit a warm bias in the lower boundary layer (Fig. 8a), unlike the general wintertime period shown in Fig. 6b, which revealed a strengthening cold bias near the surface. The biases present in the two model versions are almost identical in the southerly wind cases, with a reduction in the magnitude of the warm bias in RAPv5, by at most 1°C (Figs. 8a,b) and a reduction in the low wind speed bias in the lowest 400 m of about 1 m s^{-1} (Figs. 8e,f). Humidity biases in these southerly wind cases are relatively minor for both model versions (less than 10% in general), leaving little room for improvement in the RAPv5 (Figs. 8c,d). These slightly reduced low-level biases may be explained by a change in land use dataset from a 30-s MODIS to 15-s MODIS (used in RAPv4 and v5, respectively), which greatly increased the spatial variation of land use categories over Alaska. Overall, the RAPv5 appears to match or improve upon the RAPv4 across the board for the February 2019 period, with especially notable improvements in thermodynamic biases during periods of onshore flow.

c. *Surface meteorology and surface inversion biases*

To further investigate model biases, it is important to look at them as a function of both time-of-day and forecast hour. To do

[1] Due to the computational expense of the RAP, limited duration retrospective runs are performed after the model code is frozen as part of the NWS evaluation process. February 2019 was the selected wintertime retrospective period for this model.

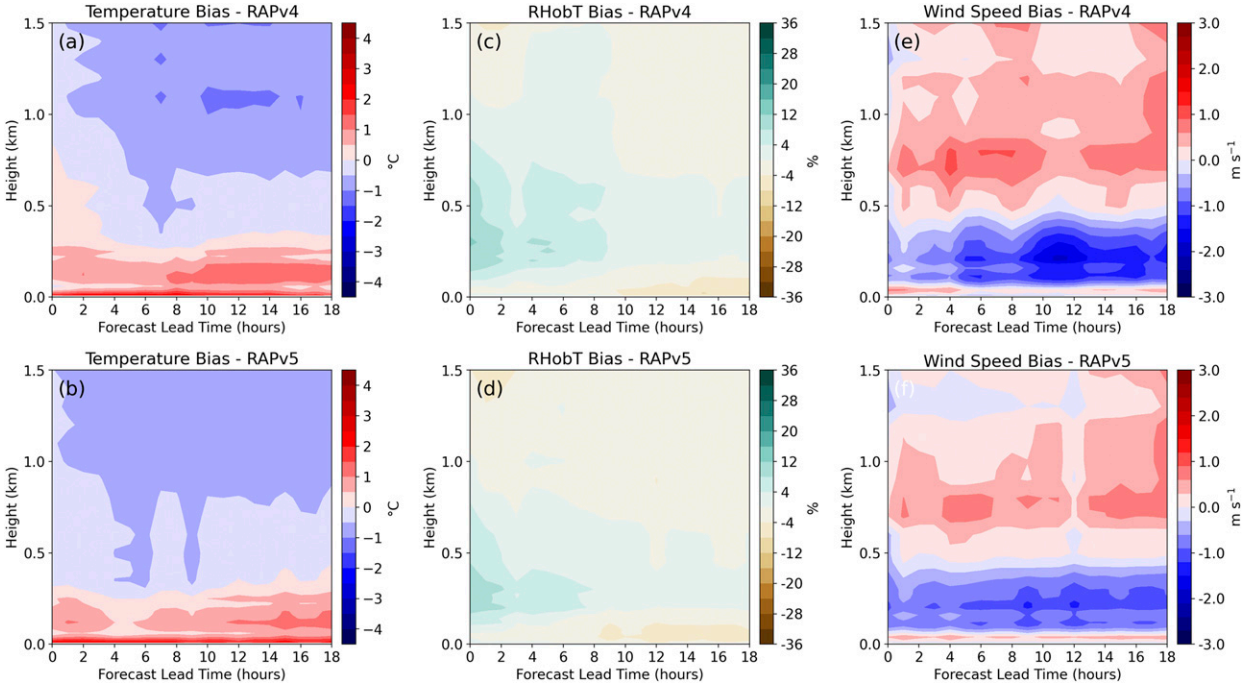

FIG. 8. As in Fig. 7, but for southerly wind events.

so, biases in temperature, RHobT, and wind speed were calculated against surface meteorological observations for observed northerly wind cases occurring both in summer and winter (Fig. 9). This comparison with the surface observations agrees well with results in Fig. 5, showing a warm and moist bias in summer and a cold and dry bias in the summer with magnitudes of up to 5°C and 20%. The surface observations show that RAPv4 winds are consistently too fast (1–2 m s^{-1}) for all forecast times (Fig. 9e) during the summer; this pattern is not seen as clearly in the comparisons with the radiosondes (Fig. 5e). However, the surface observations reveal that the biases depend on the time of day when the forecast would be valid. In the summer, biases in temperature and humidity exhibit a diurnal pattern, with more intense warm and moist biases in the afternoon local time (2200–0400 UTC), and less biased forecasts in the early morning local time (1100–1700 UTC) (Fig. 9). Potential causes for this diurnal pattern include issues with downwelling solar radiation related to cloud properties (which will be discussed more in section 3d) or to errors in the representation of nocturnal inversions, which then carry into the daytime. This diurnal pattern is not present in the winter, likely because of the lack of a daily cycle in solar radiation. The impact of data assimilation cycling on model biases also emerges when viewing surface errors. Individual RAPv4 runs (which appear as lower-left to upper-right diagonals on Fig. 9) directly following the introduction of new observations of fields like sea surface temperature (assimilated prior to the 0100 UTC run) seem to feature a brief decrease in biases before those biases return in future runs. So, while a correction toward the true state is provided by the data assimilation step, the model physics reintroduce biases shortly thereafter.

While the discussion of thermodynamic biases above helps to explain some of the lower atmospheric errors exhibited by the RAPv4, we have yet to address another key feature of the Arctic atmosphere: surface temperature inversions. We also note that for simplicity, only results for forecast hour six (using soundings valid at both 0000 and 1800 UTC) are shown in the next two sections; though not shown, the bias patterns remain qualitatively consistent throughout the forecast while varying in magnitude. Because Arctic inversions tend to be quite shallow, lapse rates below 950 hPa can serve as a metric for whether an inversion is present in the atmosphere. Using this metric, the RAPv4 seems able to simulate low-level lapse rates quite well in a probabilistic sense, especially during the summer, when compared with radiosonde observations (Fig. 10a). In particular, the more extreme ends of the spectrum (both very strong inversions and near superadiabatic layers with lapse rate magnitudes of greater than 10°C km^{-1}) are captured quite well. In the winter, the RAPv4 seems to overrepresent inversions weaker than around 20°C km^{-1} relative to the observations, though the strongest inversions are generally modeled well (Fig. 10b). This ability to model the strongest inversions and mixed layers (which would tend to occur in clear conditions) well may indicate that the model has the most trouble simulating inversions during cloudy periods, which would often feature the misrepresented weak inversions.

Still, biases emerge in the precise characteristics of the RAPv4-simulated surface inversions compared to observed inversions. On average, RAPv4-simulated inversions are not as

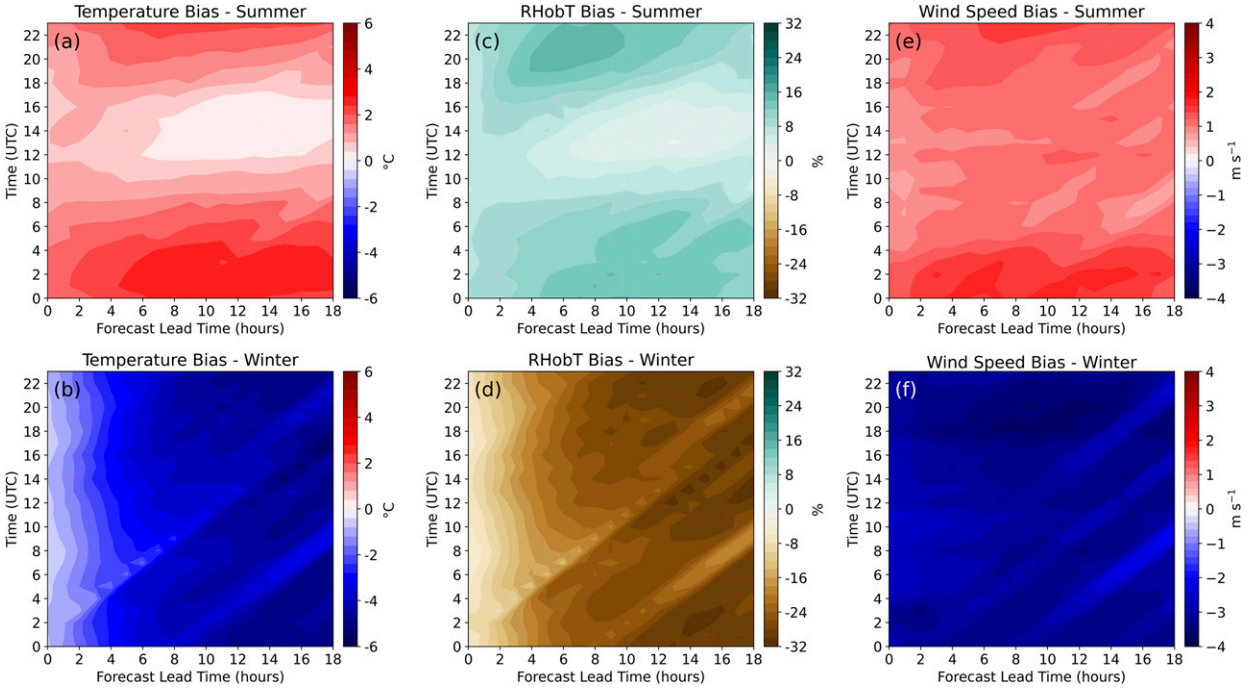

FIG. 9. RAP near-surface bias relative to observations in the (a),(b) 2-m temperature (°C), (c),(d) 2-m RHobT (%), and (e),(f) 10-m wind speed (m s^{-1}) as a function of time of day (UTC; *y* axis) and forecast lead time (h; *x* axis) for conditions with northerly winds for both (a),(c),(e) summer and (b),(d),(f) winter at Oliktok Point (OLI). Note that during the summer, the mean solar noon occurs around 2200 UTC.

steep as their observed counterparts (Fig. 10c). So, for a given inversion strength, the model will on average produce inversions that are a few hundred meters deeper than would generally be observed. Moreover, the modeled inversions exhibit far more spread within their characteristics, especially depth (Fig. 10c). While there is a very low probability of an observed inversion deeper than 600 m, there is a notable probability of modeled inversions up to 1000 m deep.

d. *Cloud cover and cloud property biases*

Finally, we evaluate RAPv4's ability to accurately simulate Arctic clouds. This ability is critical and may ultimately affect

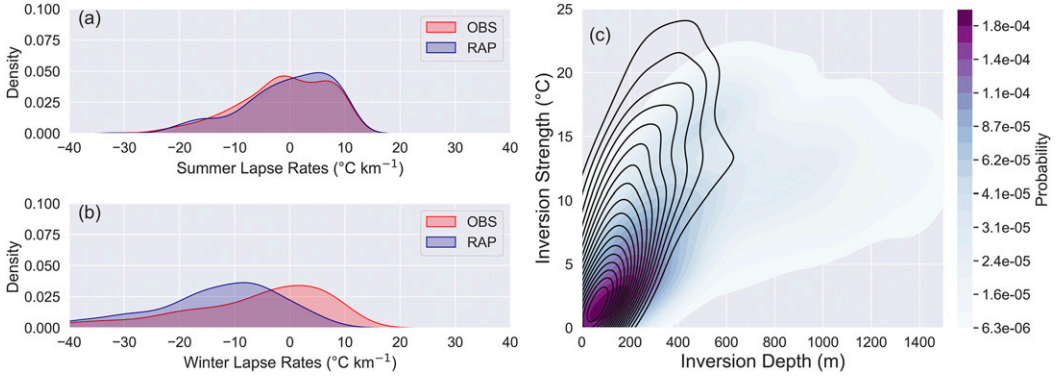

FIG. 10. (a) Probability distribution of surface to 950-hPa lapse rates (defined as positive when temperature decreases away from the surface; °C km^{-1}) for all summer cases at Oliktok Point (OLI) at forecast hour 6, both observed (red) and modeled (blue). (b) As in (a), but for all winter cases. (c) Probability density distribution of modeled (color fill) and observed (black contour) inversion strengths (difference in the maximum temperature in the inversion and the surface temperature; °C; *y* axis) and inversion depths (distance from the surface to the height of the maximum temperature; m; *x* axis) for all matched inversions (inversion present in both the model and the observations) for forecast hour 6 at Oliktok Point (OLI). That is, the probability that an observed inversion or a modeled inversion would have the given strength and depth combination. The observed probability contours are at the same numerical levels as the model.

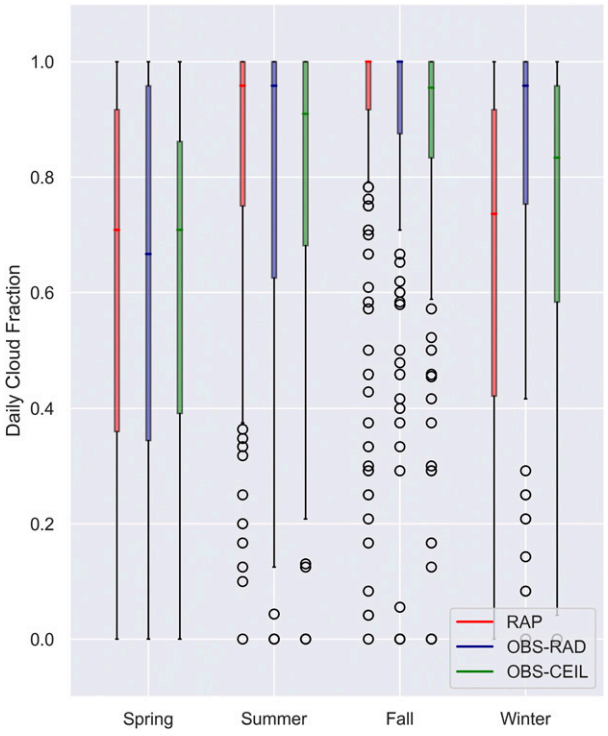

FIG. 11. Seasonal daily cloud fraction box-and-whisker plots calculated from RAP downwelling longwave radiation values for forecast hour 6 (RAP; red), observed downwelling longwave radiation values (OBS-RAD; blue), and ceilometer observations (OBS-CEIL; green) at Oliktok Point (OLI). Medians are represented by solid horizontal colored lines, while the boxes and whiskers represent the 25th–75th and 5th–95th percentiles, respectively, and open dots represent outliers.

and be affected by the thermodynamic errors described in the section above. We classify cloudy time points in the data using downwelling longwave radiative flux density, as measured by surface instruments and calculated from the RAPv4. Doing so allows us to use a consistent variable between observations and simulations that offers a signal between cloudy and clear conditions as well as between different types of clouds; however, because other factors such as the liquid water path and base height of any clouds present would also affect downwelling longwave radiation, care must be taken when evaluating cloud presence on this metric alone. While there are some discrepancies between this radiative technique and one that leverages active remote sensing (ceilometer), it seems that the radiation method serves as an acceptable proxy for determining cloud cover, with errors in the median generally less than 10% between the methods (Fig. 11, Table 2 row 3). In general, the radiative technique identifies more cloudy periods than the ceilometer, evidenced by the frequency bias values greater than 1 on all ceilometer entries of Table 2. On the one hand, this could indicate that the radiative technique better captures partly cloudy periods than the single point ceilometer measurements. Or, on the other hand, the radiative technique may identify cloud even when none are present, owing to very moist but cloudless vertical profiles. Regardless, the radiative method allows for much deeper analysis with the available model output and will be used henceforth.

Daily cloud fraction (defined as the proportion of the day during which clouds were present; 1 meaning clouds were present for the entire 24 h) values demonstrate agreement between the model and observations (median differences of less than 10%), with the greatest error in the median cloud fraction of up to 15% occurring in the winter (Fig. 11). These errors also have a seasonal component, likely connected to the average cloudiness conditions throughout the year. This daily cloud cover analysis indicates that RAPv4 possesses some accuracy when forecasting cloud occurrences; however, calculated accuracy metrics highlight the seasonal nature of errors in RAPv4's forecasting of clouds (lower CSI values and higher frequency bias values), which are particularly large during the winter and spring (Table 2). Because these two seasons seem to be characterized by more sporadic cloudiness overall (Fig. 11), this result is not too surprising. For instance, if the RAPv4 correctly forecasted partly cloudy conditions throughout a spring day, but did not model the clouds at the correct time of day, the skill metrics in Table 2 would be lowered, despite a close match in the daily cloud fractions shown in Fig. 11. In the summer and fall, when cloudiness is most persistent, skill metrics report better model performance (Table 2). In short, it

TABLE 2. Seasonal accuracy metrics between three cloud detection mechanisms at Oliktok Point (OLI): RAP cloud detection at a forecast range of 6 h (using the longwave radiation method described in the text), observed cloud detection using the longwave radiation method (OBS-RAD), and observed cloud detection using a ceilometer (OBS-CEIL). The cells of the contingency tables used to calculate these metrics are filled using a simple cloud present/cloud absent dichotomy for each detection mechanism. Metrics include critical success index [CSI; positives/(positives + false negatives + false positives)] and frequency bias [FB; (positives + false positives)/(positives + false negatives)]. Positives include cases where a cloud was present in both methods, while false positives and negatives include cases where a cloud was present in one method, but not the other, and vice versa.

	Statistic	Spring	Summer	Fall	Winter	All
RAP/OBS-RAD	CSI	0.673	0.846	0.939	0.711	0.767
	FB	0.987	1.054	1.021	0.809	0.992
RAP/OBS-CEIL	CSI	0.463	0.748	0.743	0.561	0.592
	FB	1.511	1.231	1.299	1.059	1.313
OBS-RAD/OBS-CEIL	CSI	0.594	0.847	0.770	0.747	0.699
	FB	1.531	1.167	1.273	1.308	1.365

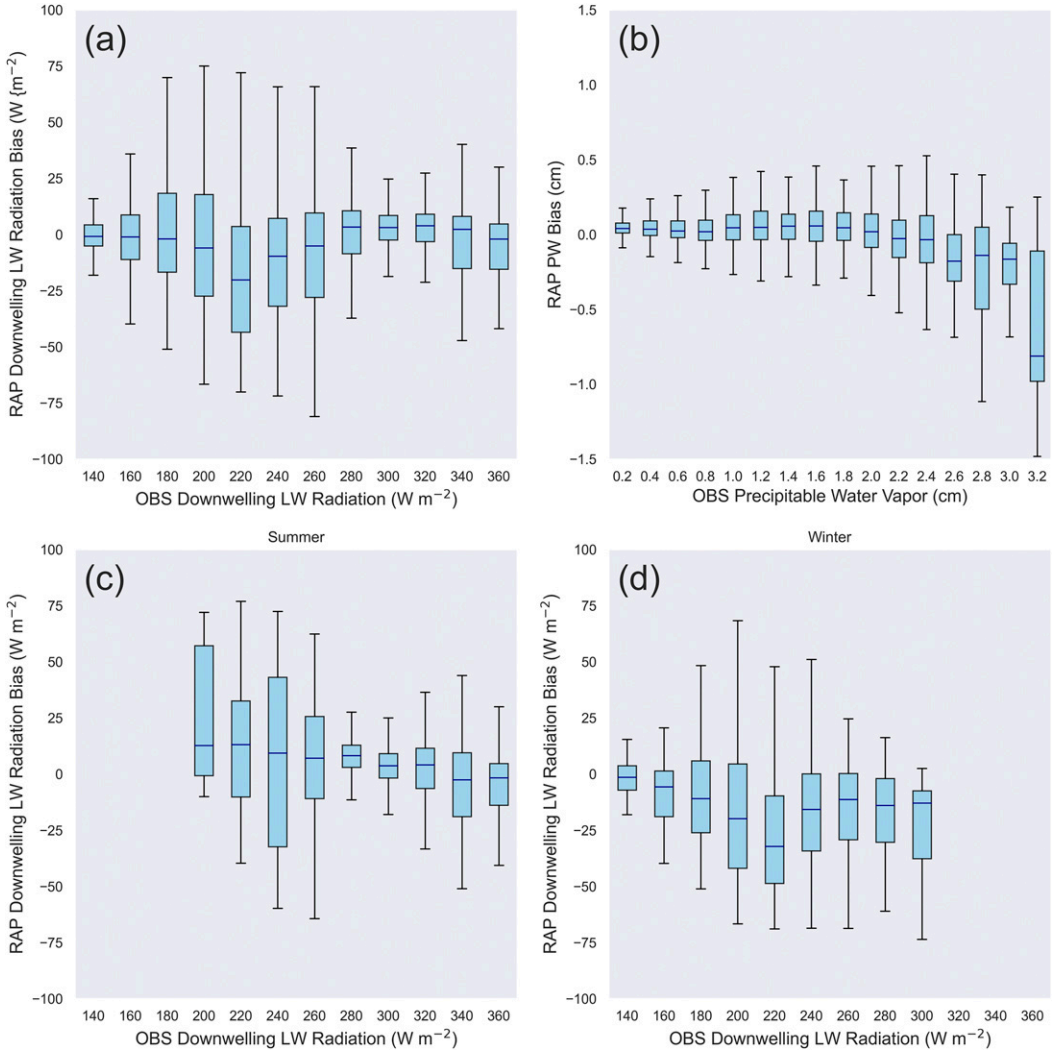

FIG. 12. (a) Box-and-whisker plots of downwelling longwave radiation bias (model minus observed; W m^{-2}) binned by the observed downwelling longwave radiation (W m^{-2}) for positive cloud cases at forecast hour 6 at Oliktok Point (OLI). (b) As in (a), but with precipitable water vapor biases (cm) by the observed precipitable water values. (c) As in (a), but for only summer cases. (d) As in (a), but for only winter cases.

appears that the RAPv4 more accurately forecasts periods of long-lasting clouds than periods of broken cloud cover, leading to a seasonal trend in model performance.

To understand how well RAPv4 simulates the radiative and thermodynamic properties of clouds, as opposed to the simple occurrence of a cloud, we analyze ''positive'' cases—that is, cases where both the RAP and the observations indicate that a cloud is present (using the longwave radiation method). Leveraging observed downwelling longwave radiation and precipitable water vapor to understand the properties of modeled and observed clouds, biases in downwelling longwave radiation are shown to be on the order of 10 W m^{-2}, which while small in comparison to the total downwelling flux, could be significant in a stable boundary layer underneath a decoupled cloud (Fig. 12a). These biases are especially small (less than 10 W m^{-2}) for cases with high optical depth (large downwelling longwave radiation values) and low optical depth (small downwelling longwave radiation values), which are indicative of thick, low clouds or high, wispy clouds, respectively (Fig. 12a). The largest errors emerge in cases where the observed downwelling longwave radiation is neither particularly high nor low (Fig. 12a). While downwelling longwave radiative flux is sensitive to the precipitable water vapor (PWV), the RAP does a good job in predicting the PWV at all lead times (Fig. 12b shows the results at forecast hour 6). However, the downwelling LW flux is also strongly impacted by clouds, and thus biases in this flux could illustrate problems with the simulated macro and microphysical properties of clouds.

In the cases with the largest errors in downwelling longwave radiation, the model tends to exhibit a negative bias, which would imply that the modeled clouds are either not optically thick enough, have insufficient spatial coverage, or are at the

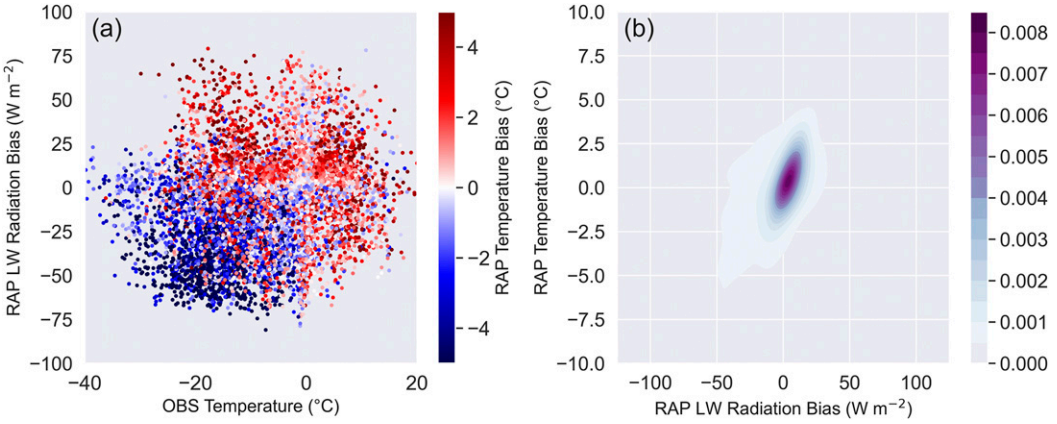

FIG. 13. (a) Scatterplot of the RAP's downwelling longwave radiation bias (W m^{-2}) and observed near-surface air temperature (°C) with each point shaded by the model's near-surface air temperature bias for all positive cloud cases at forecast hour 6 at Oliktok Point (OLI). (b) A probability density distribution of model's near-surface air temperature bias (°C) and downwelling longwave radiation bias (W m^{-2}) for all positive cloud cases at forecast hour 6 at Oliktok Point (OLI).

wrong altitude. Looking at this seasonally, this negative bias is not present in the summer, instead being replaced by a positive error of around 10 W m^{-2} (Fig. 12c). It is important to note, however, that the sample size of summer cases in this part of the longwave radiation spectrum (180–230 W m^{-2}) is quite low (50 total cases), so this positive bias may not be significant. On the other hand, the winter cases exhibit a negative bias on the order of −25 W m^{-2} in the middle of the longwave radiation spectrum (Fig. 12d).

Due to their radiative effect on the atmosphere below, clouds and lower atmospheric temperatures are closely linked (e.g., Miller et al. 2013). RAPv4 cloud property errors show a strong connection with surface temperature errors, potentially helping to explain some of the lower atmospheric biases discussed in section 3b. Cases with positive downwelling longwave radiation biases by and large also exhibit positive surface temperature biases and vice versa (Fig. 13a). Furthermore, these errors are moderately correlated (with a correlation coefficient of 0.47), so that larger errors in downwelling longwave radiation correspond to larger temperature errors (Fig. 13b). These biases likely stem from RAPv4 cloud property errors, such as incorrect hydrometeor phase, cloud height, or cloud depth, though it is difficult to deduce the exact cause without more information on subgrid-scale cloud properties in the RAP.

Breaking down cloud property biases by season and focusing on observed and simulated cloud radiative effect (CRE), which is defined as the difference between the actual downwelling longwave radiation value and the hypothetical clear sky downwelling longwave radiation value, winter clouds emerge as the primary driver of the downwelling LW radiation errors. During summer, simulated and observed clouds exhibit very similar magnitudes of CRE, with 90% of cases falling between 40 and 100 W m^{-2} to the surface radiation budget (Fig. 14a). In the winter, on the other hand, while the observed CREs tend to fall uniformly within the 40–80 W m^{-2} range, around 80% of simulated CREs are much lower—between 0 and 50 W m^{-2} (Fig. 14b).

Further stratifying cloud longwave radiation biases by approximate cloud base temperature, a new pattern appears (Fig. 14c). For clouds with a base temperature of less than −10°C (which overwhelmingly occur during the winter), model biases in downwelling longwave radiation increase dramatically, and remain elevated for cloud temperatures down to approximately −30°C (Fig. 14c). These cold cloud biases fall between −75 and −25 W m^{-2}, while clouds warmer than −10°C (which mostly occur in the summer) produce bias values very close to 0 W m^{-2}. Supercooled liquid water clouds exist frequently at temperatures between −30° and −10°C in the Arctic, and one possible explanation is that the RAP is treating clouds in this temperature range as ice only (i.e., they do not contain liquid water), and thus the model is underestimating the total optical depth of the cloud and hence underpredicting the downwelling LW flux.

4. Discussion and summary

The seasonal near-surface thermodynamic biases presented in section 3b appear to reveal the model's treatment of sea ice as a primary driver for biases observed in wintertime northerly wind cases. Although a weak cold, dry bias is present in winter cases with southerly winds, the magnitude of the error is much larger with northerly winds (Figs. 5 and 7). Further, because the winter bias is relatively small at model initialization (i.e., forecast hour 0) and grows rapidly during the model run, it would seem that a model physics problem, rather than a data assimilation error, is the most likely cause (Fig. 5b). As the ocean north of OLI would be mostly ice-covered during the wintertime, sea ice treatment appears to be the most likely source for the bias. Further evidence for this hypothesis is provided by a marked decrease of this bias in the newer RAPv5 (Fig. 7), which includes support for fractional sea ice coverage.

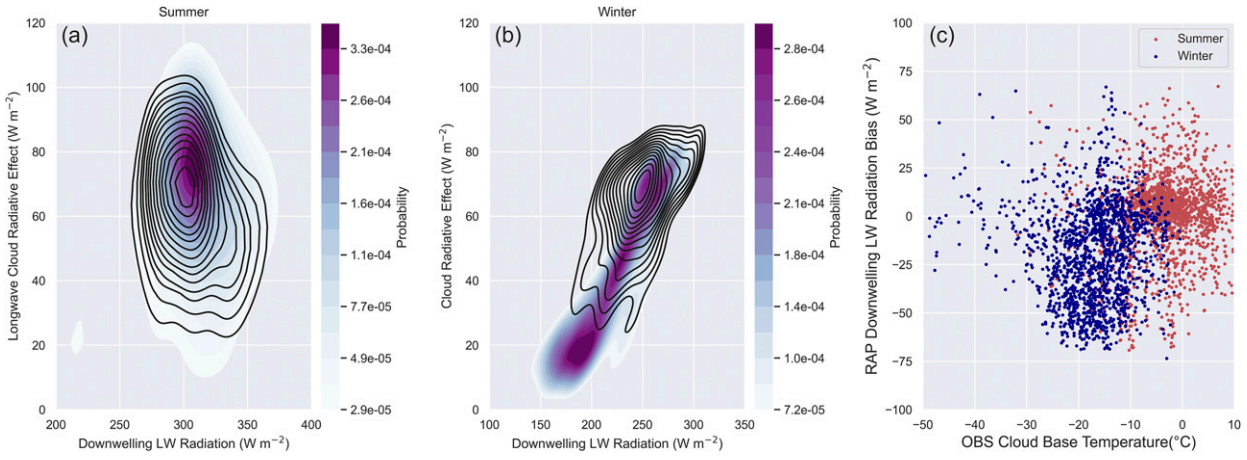

FIG. 14. (a) Cloud radiative effect (CRE; observed downwelling longwave radiation minus clear-sky downwelling longwave radiation) vs observed downwelling longwave radiation for positive cloud cases at forecast hour 6 during the summer at Oliktok Point (OLI). Observed values are in black contours, while RAP modeled values are in color fill. (b) As in (a), but for winter cases. (c) Bias in downwelling longwave radiation (RAP minus observations) vs observed cloud base temperature for positive cloud cases at Oliktok Point (OLI) at forecast hour 6. Cloud-base temperature is calculated using observed soundings interpolated to the ceilometer indicated cloud base height. Red points indicate summer cases, while blue points indicate winter cases.

Nevertheless, to some degree these biases remain even in RAPv5, and so other parameterizations, including those governing cloud properties and simulated turbulence and mixing, likely contribute as well.

Explanations for the remaining lower atmospheric biases, including the warm bias present under both wind directions in the summer (Figs. 5 and 6), are less clear. Persistence of a summer warm bias across all forecast hours and wind directions implies the strong likelihood of a systematic error that is connected to the initial analysis, rather than a model physics issue. Interestingly, a persistent warm bias is also present for both model versions for southerly wind cases in February 2019 (Figs. 8a,b). This warm bias, along with a low-level wind bias, is reduced slightly in the newer RAP version, possibly owing to a more refined land use system in the RAPv5; however, there are relatively fewer cases in February 2019 period relative to the larger multimonth period and thus the statistical comparison between v4 and v5 have more uncertainty.

Any of these thermodynamic biases (in both the summer and the winter) may also be attributable to other physical processes. Cloud errors, including sky cover fraction, height, and composition, likely play a large role in driving these boundary layer biases; indeed, we have already shown a link between longwave radiation errors and surface temperature errors (Fig. 13). Errors in the modeling of turbulent and radiative fluxes from the surface or from other couplings with the model surface may also play a role, though a more detailed investigation would be needed to confirm this hypothesis. Such errors in turbulent fluxes may also serve to connect thermodynamic and dynamic errors in the model. Away from the lowest levels of the atmosphere which directly interact with the surface, it is more difficult to pinpoint exact causes for model biases; however, it is possible that insufficient (or overaggressive) mixing driven by errors in surface inversions, cloud cover, or other near-surface thermodynamic processes may impact how biases are propagated vertically in the model.

Although the dynamic errors in this study were generally quite small (on the order of $1\ \mathrm{m\,s^{-1}}$ throughout the boundary layer), several notable wind bias patterns appeared, as discussed in section 3b. These errors may be the result of misrepresentations of surface drag (due to boundary layer parameterizations or land surface issues) or larger-scale orographic drag from the complex terrain in the region (especially noting that the bias patterns were dependent on wind direction). On the other hand, these wind errors may be more process driven, owing to misrepresentations of physical phenomena like low-level jets induced by baroclinicity or inertial oscillations or to misrepresentations of synoptic-scale weather systems.

Focusing on lower atmospheric temperature profiles, RAPv4 generally performs well across all seasons. Statistically, even the most extreme lower atmospheric lapse rates are captured by the model, as evidenced by the matching tails of the distributions in Fig. 10. Still, some biases are evident, including an apparent overrepresentation of inversions in the winter. Such errors could be explained through any number of processes, including errors in cloud properties. For example, overrepresentation of winter inversions might occur if the RAP does not simulate the properties of inversion-eroding clouds correctly, which will be discussed more below. In general, though, the RAP appears to handle radiative processes well in clear conditions (e.g., strong inversions and well-mixed layers), in contrast to many previous studies of radiation models in the Arctic (e.g., Kleczek et al. 2014; Hogan et al. 2017).

Looking at the statistics of simulated surface inversions across all seasons, there appears to be a clear preference for deeper inversions (by up to a couple hundred meters) in RAPv4 relative to observations, despite the simulated inversions

having a strength similar to what was observed. Most likely, this is an issue of vertical resolution; the model may simply not include enough vertical levels in the lowest 500 m of the atmosphere to accurately represent the extreme inversions found in the Arctic. Additionally, these biases may be a result of filtering errors in the RAP's PBL scheme. PBL schemes have been shown to introduce vertical smoothing to strong gradients in previous studies (e.g., Skamarock et al. 2019), and it is possible that this contributes to the inversion errors.

Based on our analysis, the RAPv4 seems to forecast cloud occurrence fairly accurately across all seasons. Focusing on cloud properties, we find pronounced biases in simulated downwelling longwave radiation at the surface during cloudy time periods (Fig. 12). This bias is predominantly negative, except in summer, and peaks in intensity during winter and spring. Indeed, RAPv4 winter clouds are generally associated with a much smaller cloud radiative effect than their observed counterparts (Fig. 14b). Thus, RAPv4 clouds tend to produce too little downwelling radiation, especially in colder months, which could be due to errors in cloud height, thickness, or thermodynamic phase.

Breaking down these longwave radiation biases by cloud temperature, a clear stratification occurs at cloud temperatures of around −10°C (Fig. 14c). Clouds warmer than −10°C, which mostly occur in the summer, tend to have small radiation errors, while those between −30° and −10°C, which occur mostly in the winter, exhibit more negative biases. Although this temperature link could point toward cloud height errors as a source of bias, the abrupt change at −10°C is more indicative of a mixed phase cloud issue, including misrepresentation of microphysical processes like the Wegener–Bergeron–Findeisen process between supercooled water and ice crystals. The subgrid-scale cloud scheme in RAPv4 treats any cloud hydrometeors as all ice below −19°C and all liquid above −4°C, with a linear transition for temperatures in between these limits (J. Kenyon 2019, personal communication). As a result, clouds containing supercooled water, which occur frequently in the Arctic down to temperatures as low as −30°C or colder (Shupe 2011), are underrepresented by RAPv4. The treatment of supercooled liquid clouds in the Arctic has been a problem for many modeling systems, and remains a challenge to this day (e.g., Klein et al. 2009; Komurcu et al. 2014). This hypothesis is also congruent with the observed negative bias in downwelling longwave radiation, as ice-only clouds would tend to have smaller total optical depth, and thus contribute a smaller amount to the downwelling LW radiative flux, than their mixed-phase counterparts.

These results are being used to guide the development of future versions of the RAP, and its ultimate replacement, the Rapid Refresh Forecast System (RRFS) which is part of NOAA's move toward a Unified Forecast System (UFS). In particular, while changes to the microphysical properties of the subgrid-scale clouds could not be made to the RAPv5 due to its release cycle, RRFS model developers are evaluating the impact of changing the cloud phase of the subgrid-scale clouds from all-ice below −35°C, all liquid above −4°C, and linearly interpolated between those two thresholds (J. Kenyon 2019, personal communication). We anticipate that this change will not only improve the agreement in downwelling LW radiative flux, but that the distribution of the surface-based inversions will improve also due to cloud radiative effects. Furthermore, RRFS developers are continually working to refine the stability functions and how their impact on turbulent mixing in the planetary boundary layer (PBL) scheme, and the extremely stable cases observed at NSA and OLI provide excellent cases to use in this evaluation, which may lead to improvements in the wind and thermodynamic structure in the lower 500 m of the Arctic PBL.

In general, the present study has focused on an evaluation of the RAP's performance for small-scale Arctic phenomena (e.g., seasonal sea ice, surface inversions, and mixed-phase clouds). Future work to further this goal could include utilizing additional datasets from field campaigns to more deeply explore the cloud property and boundary layer errors introduced above. Moreover, future studies could examine the RAP's performance for additional phenomena commonly found in the Arctic, such as nocturnal low-level jets, and more deeply assess errors in surface radiation and turbulent fluxes under different regimes.

At the same time, it is necessary to recognize that because this study focuses on biases at small spatial scales, more work is needed to investigate the RAP's performance with respect to larger-scale features in the Arctic, like Arctic cyclones. Further, it is important to note that this analysis is limited predominately to two sites with high-resolution observations, and focuses on only one modeling system. So, caution should be taken in translating these results to other operation models or to other regions of the Arctic. Future work studying the RAP and other operational models in the Arctic is needed to close these gaps. This work could include the incorporation of radiosonde and surface measurements from across the Arctic, or focus on an evaluation against a reanalysis dataset for thorough coverage of the data-sparse Arctic.

Acknowledgments. The authors would like to recognize the NOAA Hollings Program, which supported the lead author during his time at the NOAA Global Systems and Physical Sciences Laboratories. Additional support for this analysis was provided by the NOAA Physical Sciences Laboratory, the NOAA Global Systems Laboratory, and the U.S. Department of Energy Atmospheric Systems Research (ASR) program under Award DE-SC0013306 (de Boer, PI). We would like to recognize valuable discussions held on initial results of this analysis with numerous people, including Amy Solomon, Maximilian Maahn, Elise Koskelo, and Nathan Hadland. We thank Joe Olson for providing comments on an earlier version of this paper, and Jaymes Kenyon and Joe Olson for discussions on how these results might impact future model development within the RRFS team.

Data availability statement. Observations from ARM can be acquired online at the ARM data archive (https://www.archive.arm.gov/discovery/) or via the following DOIs: balloon-borne soundings (doi: 10.5439/1021460), surface meteorological data (doi: 10.5439/1025220), quality-controlled surface radiative flux data (doi: 10.5439/1027372), and ceilometer

data (doi: 10.5439/1181954). Archived RAP column data at Barrow and Oliktok is available via ftp; please contact the authors for instructions on how to access these data.

REFERENCES

Beesley, J. A., C. S. Bretherton, C. Jakob, E. L Andreas, J. M. Intrieri, and T. A. Uttal, 2000: A comparison of cloud and boundary layer variables in the ECMWF forecast model with observations at Surface Heat Budget of the Arctic Ocean (SHEBA) ice camp. *J. Geophys. Res.*, **105**, 12 337–12 349, https://doi.org/10.1029/2000JD900079.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

de Boer, G., W. Chapman, J. E. Kay, B. Medeiros, M. D. Shupe, S. Vavrus, and J. Walsh, 2012: A characterization of the present-day Arctic atmosphere in CCSM4. *J. Climate*, **25**, 2676–2695, https://doi.org/10.1175/JCLI-D-11-00228.1.

——, M. D. Shupe, P. M. Caldwell, S. E. Bauer, O. Persson, J. S. Boyle, S. A. Klein, and M. Tjernström, 2014: Near-surface meteorology during the Arctic Summer Cloud Ocean Study (ASCOS): Evaluation of reanalysis and global climate models. *Atmos. Chem. Phys.*, **14**, 427–445, https://doi.org/10.5194/acp-14-427-2014.

——, and Coauthors, 2018: A bird's-eye view: Development of an operational ARM unmanned aerial capability for atmospheric research in Arctic Alaska. *Bull. Amer. Meteor. Soc.*, **99**, 1197–1212, https://doi.org/10.1175/BAMS-D-17-0156.1.

——, and Coauthors, 2019: Atmospheric observations made at Oliktok point, Alaska, as part of the Profiling at Oliktok Point to Enhance YOPP Experiments (POPEYE) campaign. *Earth Syst. Sci. Data*, **11**, 1349–1362, https://doi.org/10.5194/essd-11-1349-2019.

Grell, G. A., and S. Freitas, 2014: A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.*, **14**, 5233–5250, https://doi.org/10.5194/acp-14-5233-2014.

Hogan, R., and Coauthors, 2017: Radiation in numerical weather prediction. ECMWF Tech. Memo. 816, 51 pp., https://doi.org/10.21957/2bd5dkj8x.

Holtslag, A. A. M., and Coauthors, 2013: Stable atmospheric boundary layers and diurnal cycles: Challenges for weather and climate models. *Bull. Amer. Meteor. Soc.*, **94**, 1691–1706, https://doi.org/10.1175/BAMS-D-11-00187.1.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res.*, **113**, D13103, https://doi.org/10.1029/2008JD009944.

Keeler, E., R. Coulter, J. Kyrouac, and D. Holdridge, 2015: Balloon-Borne Sounding System (SONDEWNPN). Atmospheric Radiation Measurement (ARM) user facility, accessed 30 May 2019, https://doi.org/10.5439/1021460.

Kleczek, M. A., G. J. Steeneveld, and A. A. M. Holtslag, 2014: Evaluation of the weather research and forecasting mesoscale model for GABLS3: Impact of boundary-layer schemes, boundary conditions and spin-up. *Bound.-Layer Meteor.*, **152**, 213–243, https://doi.org/10.1007/s10546-014-9925-3.

Klein, S., and Coauthors, 2009: Intercomparison of model simulations of mixed-phase clouds observed during the ARM Mixed-Phase Arctic Cloud Experiment. I: Single-layered cloud. *Quart. J. Roy. Meteor. Soc.*, **135**, 979–1002, https://doi.org/10.1002/qj.416.

Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, W.-S. Wu, and S. Lord, 2009: Introduction of the GSI into the NCEP global data assimilation system. *Wea. Forecasting*, **24**, 1691–1705, https://doi.org/10.1175/2009WAF2222201.1.

Komurcu, M., T. Storelvmo, I. Tan, U. Lohmann, Y. Yun, J. E. Penner, and T. Takemura, 2014: Intercomparison of the cloud water phase among global climate models. *J. Geophys. Res. Atmos.*, **119**, 3372–3400, https://doi.org/10.1002/2013JD021119.

Kyrouac, J., and D. Holdridge, 2015: Surface Meteorological Instrumentation (MET). Atmospheric Radiation Measurement (ARM) user facility, accessed 30 May 2019, https://doi.org/10.5439/1025220.

Lindsay, R., M. Wensnahan, A. Schweiger, and J. Zhang, 2014: Evaluation of seven different atmospheric reanalysis products in the Arctic. *J. Climate*, **27**, 2588–2606, https://doi.org/10.1175/JCLI-D-13-00014.1.

Long, C. N. and Y. Shi, 2006: The QCRad value-added product: Surface radiation measurement quality control testing, including climatologically configurable limits. Atmospheric Radiation Measurement Tech. Rep., 69 pp.

McCorkle, T. A., J. D. Horel, A. A. Jacques, and T. Alcott, 2018: Evaluating the experimental high-resolution rapid refresh-Alaska modeling system using USArray pressure observations. *Wea. Forecasting*, **33**, 933–953, https://doi.org/10.1175/WAF-D-17-0155.1.

Miller, N. B., D. D. Turner, R. Bennartz, M. D. Shupe, M. S. Kulie, M. P. Cadeddu, and V. P. Walden, 2013: Surface-based inversions above central Greenland. *J. Geophys. Res. Atmos.*, **118**, 495–506, https://doi.org/10.1029/2012JD018867.

Mlawer, E. J., S. J. Traubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, https://doi.org/10.1029/97JD00237.

Morris, V., and B. Ermold, 2015: Ceilometer (CEIL). Atmospheric Radiation Measurement (ARM) user facility, accessed 17 June 2019, https://doi.org/10.5439/1181954.

Olson, J. B., J. S. Kenyon, W. A. Angevine, J. M. Brown, M. Pagowski, and K. Siuselj, 2019: A description of the MYNN-EDMF scheme and coupling to other components in WRF-ARW. NOAA Tech. Memo. OAR GSD-61, 42 pp., https://doi.org/10.25923/n9wm-be49.

Peckham, S. E., T. G. Smirnova, S. G. Benjamin, J. M. Brown, and J. S. Kenyon, 2016: Implementation of a digital filter initialization in the WRF Model and its application in the Rapid Refresh. *Mon. Wea. Rev.*, **144**, 99–106, https://doi.org/10.1175/MWR-D-15-0219.1.

Pinto, J. O., D. L. Megenhardt, T. Fowler, and J. Colavito, 2020: Biases in the mesoscale prediction of ceiling and visibility in Alaska and their reduction using quantile matching. *Wea. Forecasting*, **35**, 997–1016, https://doi.org/10.1175/WAF-D-19-0230.1.

Riihimaki, L., Y. Shi, D. Zhang, and C. Long, 2019: Data Quality Assessment for ARM Radiation Data (QCRAD1LONG). Atmospheric Radiation Measurement (ARM) user facility, accessed 23 December 2020, https://doi.org/10.5439/1027372.

Shupe, M. D., 2011: Clouds at Arctic atmospheric observatories: Part II: Thermodynamic phase characteristics. *J. Appl. Meteor. Climatol.*, **50**, 645–661, https://doi.org/10.1175/2010JAMC2468.1.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note

NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

——, C. Snyder, J. B. Klemp, and S. Park, 2019: Vertical resolution requirements in atmospheric simulation. *Mon. Wea. Rev.*, **147**, 2641–2656, https://doi.org/10.1175/MWR-D-19-0043.1.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and J. S. Kenyon, 2016: Modifications to the Rapid Update Cycle land surface model (RUC LSM) available in the Weather Research and Forecasting (WRF) Model. *Mon. Wea. Rev.*, **144**, 1851–1865, https://doi.org/10.1175/MWR-D-15-0198.1.

Sotiropoulou, G., J. Sedlar, R. Forbes, and M. Tjernström, 2016: Summer Arctic clouds in the ECMWF forecast model: An evaluation of cloud parameterization schemes. *Quart. J. Roy. Meteor. Soc.*, **142**, 387–400, https://doi.org/10.1002/qj.2658.

Steeneveld, G. J., B. J. H. van de Wiel, and A. A. M. Holtslag, 2006: Modelling the Arctic stable boundary layer and its coupling to the surface. *Bound.-Layer Meteor.*, **118**, 357–378, https://doi.org/10.1007/s10546-005-7771-z.

Thompson, G., and T. Eidhammer, 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *J. Atmos. Sci.*, **71**, 3636–3658, https://doi.org/10.1175/JAS-D-13-0305.1.

Tjernström, M., J. Sedlar, and M. D. Shupe, 2008: How well do regional climate models reproduce radiation and clouds in the Arctic? An evaluation of ARCMIP simulations. *J. Appl. Meteor. Climatol.*, **47**, 2405–2422, https://doi.org/10.1175/2008JAMC1845.1.

Turner, D. D., and Coauthors, 2020: A verification approach used in developing the Rapid Refresh and other numerical weather prediction models. *J. Oper. Meteor.*, **8**, 39–53, https://doi.org/10.15191/nwajom.2020.0803.

Uttal, T., and Coauthors, 2016: International Arctic Systems for Observing the Atmosphere: An international polar year legacy consortium. *Bull. Amer. Meteor. Soc.*, **97**, 1033–1056, https://doi.org/10.1175/BAMS-D-14-00145.1.

Verlinde, J., and Coauthors, 2007: The mixed-phase Arctic cloud experiment. *Bull. Amer. Meteor. Soc.*, **88**, 205–222, https://doi.org/10.1175/BAMS-88-2-205.

——, B. D. Zak, M. D. Shupe, M. D. Ivey, and K. Stamnes, 2016: The ARM North Slope of Alaska (NSA) sites. *The Atmospheric Radiation Measurement Program: The First 20 Years*, *Meteor. Monogr.*, No. 57, Amer. Meteor. Soc., https://doi.org/10.1175/AMSMONOGRAPHS-D-15-0023.1.

Wesslén, C., M. Tjernstrom, D. H. Bromwich, G. de Boer, L.-S. Bai, and S.-H. Wang, 2014: The Arctic summer atmosphere: An evaluation of reanalyses using ASCOS data. *Atmos. Chem. Phys.*, **14**, 2605–2624, https://doi.org/10.5194/acp-14-2605-2014.

Whitaker, J. S., T. M. Hamill, X. Wei, Y. Song, and Z. Toth, 2008: Ensemble data assimilation with the NCEP global forecast system. *Mon. Wea. Rev.*, **136**, 463–482, https://doi.org/10.1175/2007MWR2018.1.

Wilson, A. B., D. H. Bromwich, and K. M. Hines, 2011: Evaluation of polar WRF forecasts on the Arctic System Reanalysis domain: Surface and upper air analysis. *J. Geophys. Res.*, **116**, D11112, https://doi.org/10.1029/2010JD015013.