RESEARCH ARTICLE

Central Arctic weather forecasting: Confronting the ECMWF IFS with observations from the Arctic Ocean 2018 expedition

Michael Tjernström[1] | Gunilla Svensson[1] | Linus Magnusson[2] |
Ian M. Brooks[3] | John Prytherch[1] | Jutta Vüllers[3] | Gillian Young[3]

[1]Department of Meteorology and the
Bolin Centre of Climate Research,
Stockholm University, Stockholm, Sweden

[2]European Centre for Medium-Range
Weather Forecasts, Reading, UK

[3]School of Earth and Environment,
University of Leeds, Leeds, UK

Correspondence
M. Tjernström, Department of
Meteorology, Stockholm University, 106
91 Stockholm, Sweden.
Email: michaelt@misu.su.se

Abstract

Forecasts with the European Centre for Medium-Range Weather Forecasts' numerical weather prediction model are evaluated using an extensive set of observations from the Arctic Ocean 2018 expedition on the Swedish icebreaker *Oden*. The atmospheric model (Cy45r1) is similar to that used for the ERA5 reanalysis (Cy41r2). The evaluation covers 1 month, with the icebreaker moored to drifting sea ice near the North Pole; a total of 125 forecasts issued four times per day were used. Standard surface observations and 6-hourly soundings were assimilated to ensure that the initial model error is small. Model errors can be divided into two groups. First, variables related to dynamics feature errors that grow with forecast length; error spread also grows with time. Initial errors are small, facilitating a robust evaluation of the second group; thermodynamic variables. These feature fast error growth for 6–12 hr, after which errors saturates; error spread is roughly constant. Both surface and near-surface air temperatures are too warm in the model. During the summer both are typically above zero in spite of the ongoing melt; however, the warm bias increases as the surface freezes. The warm bias is due to a too warm atmosphere; errors in surface sensible heat flux transfer additional heat from the atmosphere to the surface. The lower troposphere temperature error has a distinct vertical structure: a substantial warm bias in the lowest few 100 m and a large cold bias around 1 km; this structure features a significant diurnal cycle and is tightly coupled to errors in the modelled clouds. Clouds appear too often and in a too deep layer of the lower atmosphere; the lowest clouds essentially never break up. The largest error in cloud presence is aligned with the largest cold bias at around 1 km.

KEYWORDS

Arctic boundary layer, Arctic climate, Arctic clouds, Arctic reanalysis, Arctic weather prediction, model error, model evaluation, surface energy budget

1 | INTRODUCTION

Weather forecasting for the Arctic Ocean is becoming increasingly important (Jung *et al.*, 2016). Arctic warming is at least twice as large as the global average warming (Hartfield *et al.*, 2018; IPCC, 2019; Meredith *et al.*, 2019); this is labelled "Arctic Amplification" (Holland and Bitz, 2003; Serreze and Francis, 2009; Serreze and Barry, 2011). The most obvious manifestation is the rapid reduction in sea ice extent (Onarheim *et al.*, 2018), thickness, and age (Ricker *et al.*, 2017; Kwok, 2018). This opens up the Arctic Ocean for increased shipping (Smith and Stephenson, 2013), creating opportunities for resource extraction and tourism, and hence economic growth as well as environmental risks. It changes living conditions for indigenous peoples, who may no longer be able to trust traditional knowledge, with extreme weather occurring more often as sea ice becomes more vulnerable (Holland and Stroeve, 2011).

To effectively predict and manage these opportunities and risks, skilful prediction systems tailored to the special conditions of the Arctic are needed (e.g. Jung *et al.*, 2016). Numerical weather prediction (NWP) is also the basis for reanalysis, a powerful and sophisticated by-product from NWP that has become a major source of scientific understanding on Arctic climate. Reanalysis is a series of analyses based on short-term weather predictions, repeatedly constrained by observations in a consistent data assimilation cycle (Parker, 2016). Their quality is limited by both that of the numerical models used to progress information forward in time, and by the quality and availability of the observations constraining the analyses.

A proper evaluation of how faithfully models reproduce processes unique to the Arctic Ocean environment requires detailed observations from the Arctic Ocean. The only way these can be obtained in sufficient amounts is from icebreaker-borne field campaigns. This form of evaluation has a long tradition for climate models (e.g. Tjernström *et al.*, 2005; 2008; Wyser *et al.*, 2008; Birch *et al.*, 2009; de Boer *et al.*, 2014; Wesslén *et al.*, 2014; Sotiropoulou *et al.*, 2016a; Sedlar *et al.*, 2020), while evaluation of global NWP models' evaluations often focus on comparing larger-scale model output against the modelling system's own analysis (e.g. Bauer *et al.*, 2016; Jung and Matsueda, 2016). Especially challenging is evaluating model vertical structure. Although there is a wealth of satellite observations in polar regions, where polar-satellite orbital tracks converge, this does not fully compensate for the lack of *in situ* observations over the Arctic Ocean, creating a problem for NWP (e.g. Naakka *et al.*, 2019).

In this article we use detailed observations from the Arctic Ocean 2018 (AO2018) expedition (Vüllers *et al.*, 2020) to evaluate operational forecasts by the
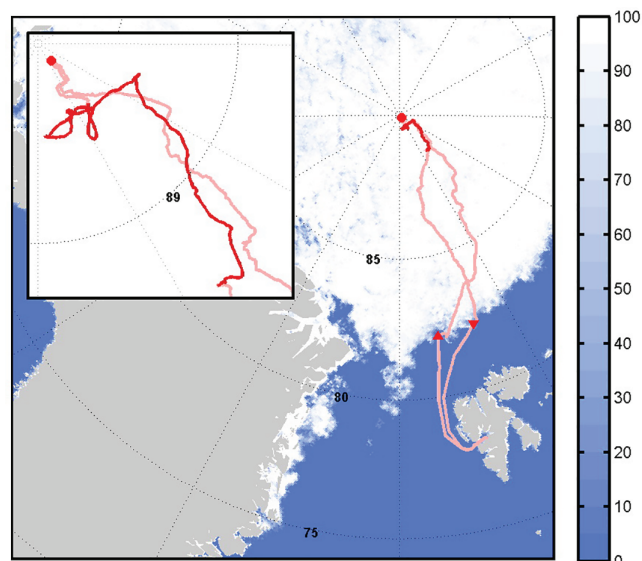

FIGURE 1   Map of the ship's track for the Arctic Ocean 2018 expedition with the whole expedition (light red) and the ice drift (dark red, also enlarged in the insert). Colour shading shows ice concentration (%) for 1 September 2018, from the University of Bremen satellite sea ice product (Spreen *et al.*, 2008) and red triangles indicate position for day-long research stations in the marginal ice zone

European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS). The IFS version evaluated is very similar to the model that powers the most recent ECMWF reanalysis: ERA5 (Hersbach *et al.*, 2020). We therefore argue that weaknesses and strengths identified from evaluating operational forecasts will also appear in ERA5. This evaluation thus provides useful information for further development of the IFS and of new reanalyses, for users of the operational IFS forecasts, and on potential systematic errors in ERA5.

2 | OBSERVATIONS AND MODEL

2.1 | The Arctic Ocean 2018 (AO2018) observations

The AO2018 expedition took place on the Swedish icebreaker *Oden*, leaving from and returning to Longyearbyen on 1 August and 21 September 2018, respectively (Figure 1). We focus on the period when *Oden* was moored to, and drifted with, the sea ice: 13 August through to 14 September. The sea-ice fraction was >90%, dominated by kilometre-sized or smaller ice floes with a melt-pond fraction of ~30% upon arrival. Most ponds were small and shallow and later froze over and eventually became covered by snow. A comprehensive set of atmospheric observations were performed onboard *Oden* throughout AO2018, with

additional instruments deployed on the sea ice during the drift. A complete description of the instruments, and the meteorological conditions during the expeditions, can be found in Vüllers *et al.* (2020). Here we use a subset of the instruments, mostly deployed onboard.

The 6-hourly radiosoundings, launched at 0000, 0600, 1200, 1800 UTC, are a core component for the model evaluation. Data from these were shared globally in near-real time over the Global Telecommunication System (GTS), along with the routine SHIP observations performed by the ship's crew, and were hence assimilated in the forecast system. For winds and surface fluxes we use measurements from an eddy-covariance turbulence flux system installed on a foremast at the bow of *Oden*, with instruments located at ∼20 m above the sea surface. Wind measurements are corrected for platform orientation and motion and for flow distortion around the ship (Prytherch *et al.*, 2015; 2017). Eddy-covariance fluxes were estimated using 30 min averages. We also use observations from a weather station located on the seventh deck ∼25 m above the sea surface for basic meteorology (pressure, temperature, relative humidity, wind speed and direction), as well as broadband downwelling solar and infrared radiation and ice-surface skin temperature.

For cloud observations we use a suite of remote-sensing measurements, either in isolation or combined using the Cloudnet algorithm (Illingworth *et al.*, 2007). Co-located with the weather station was a ceilometer measuring cloud-base heights and a so-called present-weather sensor for visibility and precipitation observations. A scanning Doppler Ka-band cloud radar was located on the roof of a container on *Oden's* foredeck, while a scanning micro-pulsed Doppler lidar was deployed on top of another container located on the foredeck laboratory roof; a scanning microwave radiometer was installed alongside the lidar.

Periods with flow from the aft of the ship are problematic for onboard *in situ* observations located around the front of the ship, because of turbulence and flow distortion from the ship's superstructure. Fortunately, *Oden* re-oriented into the wind during the ice drift to maintain clean sampling for aerosol measurements. *In situ* observations were excluded for the ∼1% of the evaluation period when the relative wind direction was adverse, with one exception: the soundings. These had to be done from the helipad often in the lee of the superstructure, affecting some results below ∼50 m.

2.2 | The NWP model

We use operational forecasts issued at the time from the high-resolution deterministic (HRES) version of the ECMWF IFS, Cy45r1. The atmospheric model has a horizontal resolution of ∼9 km and 137 vertical levels, the lowest at 10 m with 8 levels below ∼200 m and 20 below 1 km, and is coupled to a 0.25° resolution ocean and sea-ice model. A detailed description of IFS can be found at https://www.ecmwf.int/en/publications/ifs-documentation. Vertical profiles of state variables and clouds were extracted from native model levels every 6 hr, while near-surface and some integrated bulk variables were extracted at hourly resolution.

Although ERA5 uses an older version of IFS (C41r2), and also has a lower horizontal resolution (∼31 km), the model physics is very similar (see https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model). The largest systemic difference is that the operational model (C45r1) is coupled to ocean and sea-ice models, while sea-ice cover and sea-surface temperature (SST) are prescribed from observations in ERA5; however, the sea-ice temperature is calculated in the atmospheric model in both. Given the similarities, it is reasonable to assume that ERA5 errors have similar characteristics to the operational model.

2.3 | Error analysis

Given the stochastic nature of the atmosphere, an inherent limitation is the length of the evaluation period and the number of forecasts evaluated. We use the first 3 days of forecasts initialized four times per day (0000, 0600, 1200 and 1800 UTC) from 0000 UTC 12 August to 0000 UTC 12 September; a total of 125 forecasts for the period when *Oden* was almost stationary. While the ice drift started almost 2 days later, *Oden* remained stationary at the North Pole on 12 August and the ice drift started nearby (within ∼30 km). The last forecast extends to the end of the ice drift late on 14 September. This period spans the end of the melt and the initial freeze of the surface (Vüllers *et al.*, 2020).

To denote time, we use decimal Day of the Year (DoY), defined with DoY = 1.0 at 0000 UTC on 1 January. We used data from the model grid point closest to *Oden's* location at the forecast start, ignoring the drift of the icebreaker with the ice, which may move it into neighbouring grid cells during a forecast. The median velocity during the ice drift was ∼0.1 m·s^{-1}; *Oden* moved on average ∼25 km during a forecast. Errors are defined as model minus observations, and unless otherwise stated we use the median error.

Time-series observations are averaged over 10 min centred on the model times, except for the turbulent fluxes, already averaged by definition; we use the 30 min averaged flux closest to the model time. For cloud-base heights we use the Vaisala ceilometer's "sky-condition" algorithm, incorporating data from 30 min with larger weight on the

last 10 min (e.g. Šálek and Szabó-Takács, 2019). For soundings we use actual launch time, disregarding the time it takes for the balloon to ascend through the troposphere and its horizontal displacement.

To make use of the higher temporal model resolution for near-surface temperature and wind, observations are first interpolated to the respective 2 and 10 m standard heights using surface-layer theory (e.g. Foken, 2006). The opposite approach, interpolating model profile data to instrument heights, does not change the main results. Turbulent fluxes and wind direction are assumed height invariant in the surface layer. For vertical profiles, it is not obvious how parametrizations and numerical methods affect the effective model resolution and therefore unclear how to average observations to fit a model grid. Once averaged, it becomes impossible to explore more detail. Therefore, we interpolate model variables linearly to a high-resolution observation grid using height above the surface, rather than averaging observations over model grid boxes in pressure or hybrid levels. Since the observations have higher vertical resolution than the model, error profiles will feature many details that come only from the observations. With this approach, however, error profiles can still be averaged to any scale needed and linear interpolation does not add any variability.

The evaluation of modelled clouds and cloud-layer characteristics is complicated for many reasons. Therefore, we do not compare cloud details between model and observations directly; this is only done for cloud presence and vertically integrated properties. The multi-sensor Cloudnet algorithm is vertically limited by the lowest radar range gate, 157 m. However, cloud bases below this were detected by the ceilometer about half the time when clouds were indicated; fog (visibility <1 km) was indicated ∼25% of all time. Hence we chose not to use the Cloudnet for cloud geometry. Vüllers *et al.* (2020) used the less frequent radar range-height indicator (RHI) scans to cover the lowest layer, but we instead adopted a simpler method.

Cloud-radar reflectivity was used to indicate cloud layers, except for the lowest layer where we used the ceilometer and visibility for cloud-base height, to bridge the gap between the surface and the lowest radar height; this also avoids misinterpreting precipitation as low clouds. If visibility was below 1 km, the lowest cloud-base height was set to zero. If the lowest cloud-base height was below 157 m and the first cloud-radar height indicated cloud presence, clouds were assumed to extend to the lowest cloud-top height from radar, else lowest cloud-top height was set to 150 m. We then proceeded to search the cloud-radar reflectivity profile upward for the next cloud layers, continuing until reaching the highest radar range gate. Since the radar cannot distinguish precipitation from clouds, multilayered clouds may become underrepresented; precipitation falling between two cloud layers will appear as one cloud.

In the model data we used non-zero specific cloud-water content to indicate clouds, similar to cloud-radar echo's. However, the model quite often has very small cloud-water content, especially at low altitude; hence, we threshold the model data. After subjective inspection of the modelled cloud-water statistics (not shown) we somewhat arbitrarily consider a grid point cloudy when cloud water exceeded 0.001 (0.0001) g·kg^{-1} below 1 km (above 4 km), with linear interpolation in between.

3 | RESULTS

The development of some state variables are illustrated in Figures 2 and 3. The model data were constructed merging the second day of all forecasts initiated at 0000 UTC. The model reproduces the variability, timing and magnitude of high wind-speed events realistically (Figure 2a,b) and captures the gradual cooling from late summer to early autumn (Figure 2c,d), manifested by the lowering of isotherms. The model also captures some higher-frequency temperature variations, for example cooling events around DoY ∼230 and 245 and the warming around DoY ∼256. Similarly, deep high-humidity events associated with weather systems and deep frontal clouds also agree well with observations (Figure 2e,f); see for example around DoY 233 and 240, and several systems that appear during DoY 245–253.

But there are also differences, especially in humidity. The model's moist layer closest to the surface, with RH$_i$ >90%, is 1–2 km thick in between passing weather systems, while the observations show several extended periods when this layer is considerably thinner, ∼500 m or less. The model's deep moist columns appear smeared in time and some are consolidated into longer periods of high humidity. Variability in the observations is not unexpectedly larger; compare for example the period DoY 243–252. This is also seen in temperature, for example the warm pulse around DoY ∼256 is almost a day longer in IFS than in the observations. The strongest winds are somewhat too weak in the model, especially for wind-speed maxima below 4 km; see for example the episodes DoY 242–246 and around DoY 256.

The thicker modelled moist layer has consequences for low clouds. The cloud-radar reflectivity and ceilometer cloud base (Figure 3a) display periods with only thin low clouds or even cloud-free conditions. One such period appears early in the ice drift, before DoY 230; the cloud radar and ceilometer indicate only intermittent thin low clouds with brief clear periods, while the model has
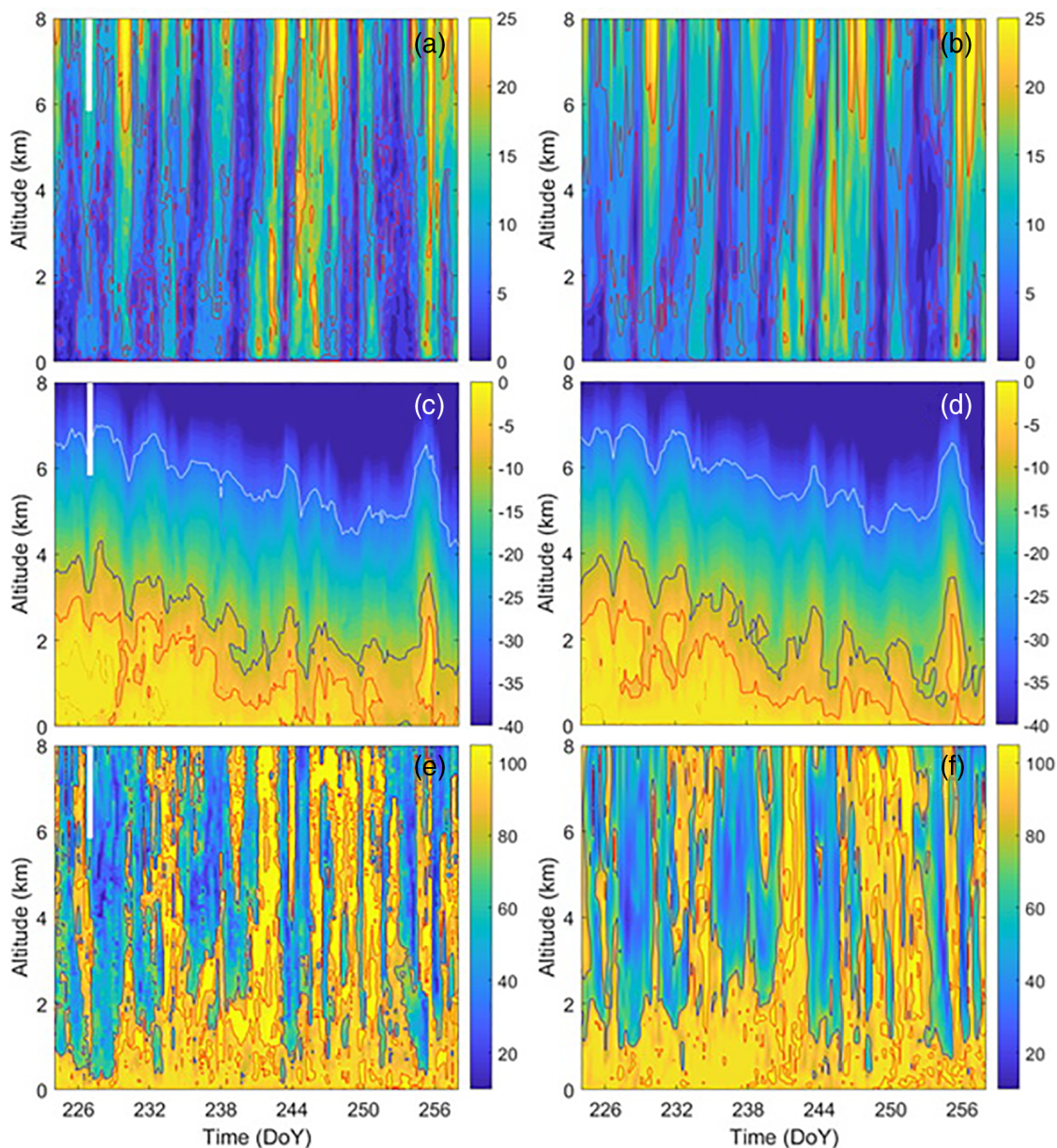
FIGURE 2 Time–height cross-sections of (a,b) scalar wind speed (m·s⁻¹), (c,d) temperature (°C), and (e,f) relative humidity w.r.t. ice (%) for the AO2018 ice drift period, with (a,c,e) from observations (radiosoundings) and (b,d,f) from IFS (see the text for a discussion). Contours are also shown: for wind speed at 5, 10 and 20 m·s⁻¹; for temperature at 0, −5, −10 and −30 °C; and for relative humidity at 80 and 100%

solid cloud cover through the lowest 1–2 km. This happens again around DoY 244 and for a period starting at DoY ∼253. The timing of deeper clouds associated with synoptic-scale weather systems are captured by the model; however, the amounts of the higher clouds appear somewhat underestimated.

Several interesting things can be seen in Figure 4, showing observed air temperature at ∼20 m above the surface and overlapping 2 m temperature forecasts from all 125 three-day forecasts. Prior to the freeze onset, estimated to around DoY 240 (Vüllers *et al.*, 2020), the surface is melting and the surface temperature cannot respond to the

surplus in the energy budget due to the latent heat transfer from the phase change and cannot exceed the melting point of fresh water, 0 °C; hence the air temperature is also constrained. However, the model's air temperature is persistently but unphysically larger than zero, by ∼0.5 °C.

This late in the melt season, periods when observed air temperature suddenly drops are frequently observed (e.g. Tjernström, 2005; Tjernström *et al.*, 2012). This happens when the low clouds become tenuous or dissipate, as the loss of net long-wave radiation overwhelms the gain of net solar radiation in the surface energy budget and the surface temperature falls (e.g. Sedlar *et al.*, 2011). One such period
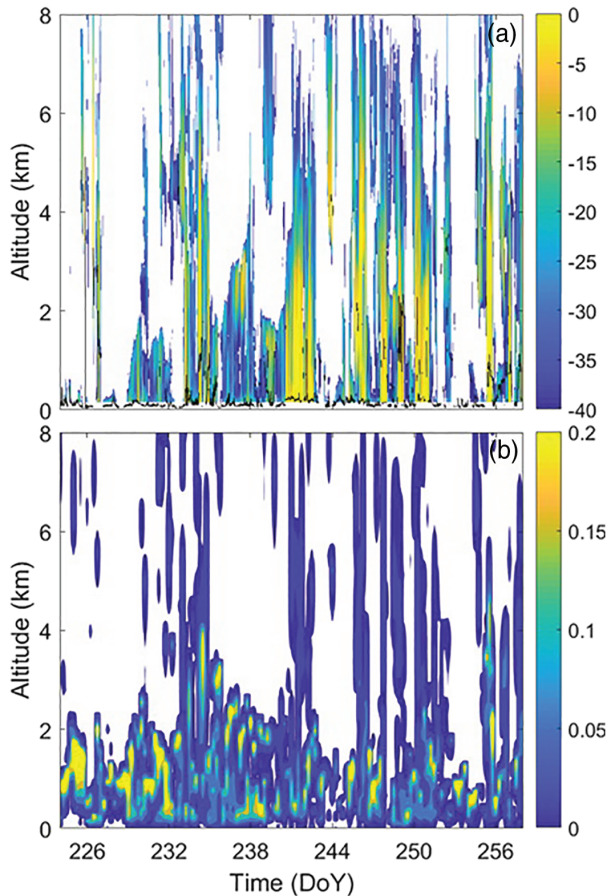
FIGURE 3 Time–height cross-sections of clouds showing (a) shading of 1 hr averaged cloud-radar reflectivity (dBz) with the lowest ceilometer cloud-base height as black markers, and (b) model cloud water mixing ratio (liquid plus ice, $g \cdot kg^{-1}$) from IFS. The model data were constructed merging the second day of all forecasts initiated at 0000 UTC

FIGURE 4 Time series of near-surface temperature (°C) as measured (blue solid) on *Oden* at ~20 m above the surface and (red dashed) from the IFS model. The model data are from overlapping three-day forecasts initiated every 6 hr. The dashed black vertical line indicates the estimated onset of the freeze-up from Vüllers *et al.* (2020)

occurs during DoY 227–230; in the model low clouds persist and hence it misses this entirely. Interestingly, however, the initial state of each forecast here is closer to reality, likely because of the assimilated observations from *Oden*, but in less than a day it reverts to the model's unphysical >0 °C state, from which it hardly deviates. As the melt ends the structure changes and the model now follows the observed trends and variability quite faithfully, although with an even larger warm bias.

A picture emerges from Figures 2–4: the model has a problem with moisture, clouds and temperatures, especially in the lower atmosphere. In the following we will explore this in detail.

3.1 | Near-surface variables

Here we explore forecast errors for a selection of near-surface variables as a function of forecast length
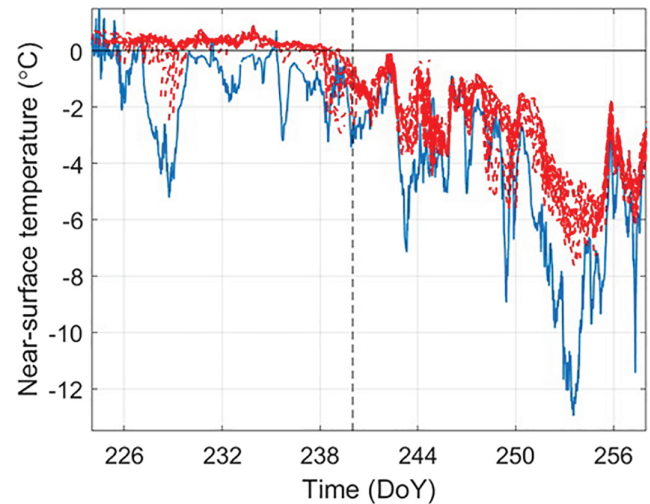
(Figure 5), showing the full error distribution (colour shading) along with the median and the mean errors; the latter is enclosed by the ±0.1% significance interval from a two-sided Student's *t*-test. If the error is normally distributed and this interval does *not* enclose zero, the null hypothesis, that the error is not different from zero, is rejected; there is <0.1% likelihood that the error is due to insufficient sampling.

The near-surface temperature warm bias is obvious (Figure 5a). The median error in the 2 m temperature is slightly larger than 1 °C and changes very little with forecast length. The error is strongly skewed; the peak of the error distribution appears at slightly below 1 °C, while the mean error is about 0.5 °C larger than the median. Although the skewness renders the *t*-test inapplicable, it is clear this is a statistically significant systematic error. The skin-surface temperature error displays a similar pattern (Figure 5f). The error distribution appears tighter but more skewed, and the median error is slightly smaller. The near-surface specific moisture is almost entirely controlled by the surface temperature; unsurprisingly it is also biased positive in the model (Figure 5b), and while also skewed it is closer to a normal distribution than the temperature errors. For temperature and moisture, the error spread is roughly constant in time through the forecast, except for during the first 6–12 hr.

Simulating winds accurately is difficult (e.g. Sedlar *et al.*, 2020) and wind forecasts are often considered more uncertain than those for temperature (e.g. Haiden *et al.*, 2019), however, the 10 m scalar wind-speed
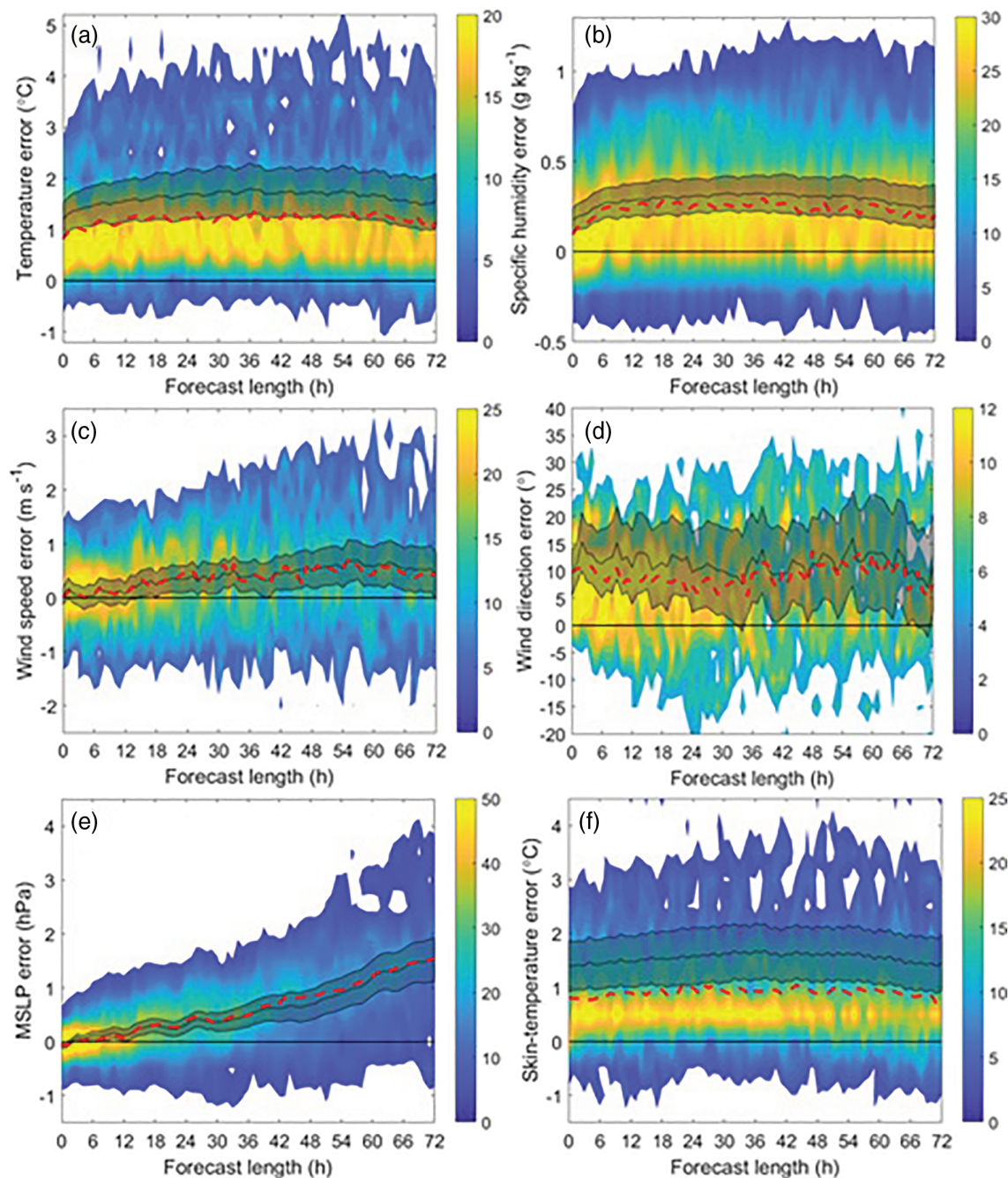
FIGURE 5 Distribution of model forecast error, defined as model minus observation, as a function of forecast length for the near-surface variables: (a) temperature (°C), (b) specific water vapour (g·kg^{-1}), (c) scalar wind speed (m·s^{-1}), (d) wind direction (°), (e) mean-sea-level pressure (hPa), and (f) skin-surface temperature (°C). Each panel shows the relative distribution of the forecast error (colour shading), the median error (red dashed), and the mean error (thin solid black), along with the Student's t-test ± 1-percentile confidence intervals in grey shading

(Figure 5c) reveals only a modest error, growing with forecast length to ∼0.5 m·s^{-1} on the third day; the spread of the wind-speed error increases with forecast length. This error is close to a normal distribution and is statistically significant only after the first forecast day. The wind direction (Figure 5d) has a significant error of about 10°, consistent with a too small boundary-layer wind turning (Lindvall

and Svensson, 2019). The spread of this error is $\pm 20°$, growing marginally with forecast length. The high quality of the wind forecasts is consistent with anecdotal experience (cf. e.g. Tjernström *et al.*, 2019).

The median sea-level pressure error (Figure 5e) is somewhat surprising: zero at forecast start, as expected due to data assimilation, then growing almost linearly
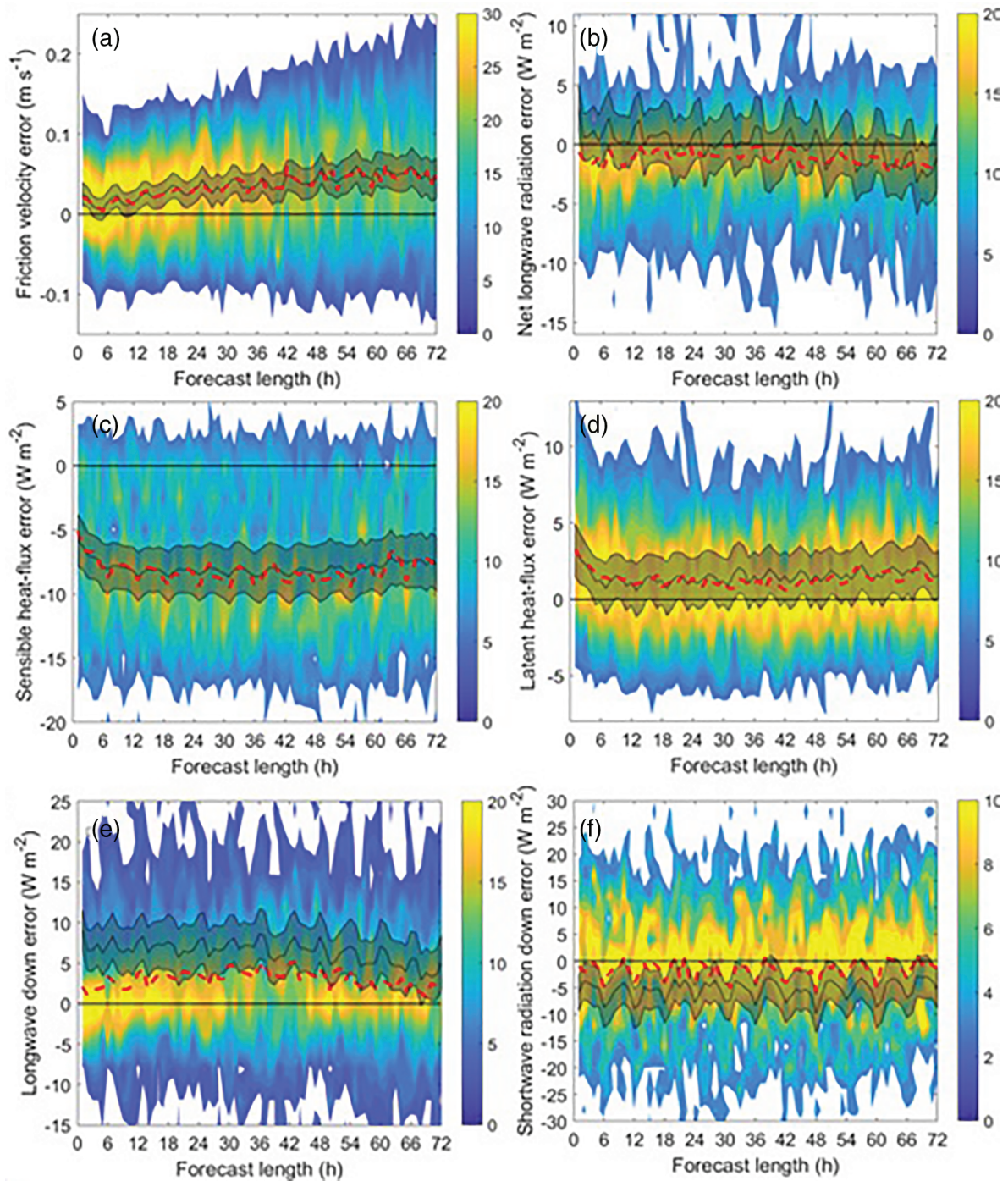
FIGURE 6 As Figure 5 but for some surface energy fluxes, showing (a) friction velocity (m·s^{-1}), (b) net long-wave radiation (W·m^{-2}), turbulent (c) sensible and (d) latent heat flux (W·m^{-2}) and downwelling (e) long-wave radiation and (f) short-wave radiation (W·m^{-2})

in time becoming significant after ~12 hr, and reaching ~1.5 hPa at the end of the third day. Since the ice moves with the wind, which is related to the pressure gradient, the ice drift may contribute systematically to the pressure error. However, even moving perpendicular for 3 days with a speed of 0.1 m·s^{-1} toward lower pressure across a gradient corresponding to a wind of 10 m·s^{-1} would only cause an observed pressure drop by less than 0.5 hPa.

An analysis of average pan-Arctic IFS surface pressure errors (not shown) reveals a previously unknown positive summer bias, with interannually varying magnitude and spatial structure. For late summer 2018 it has a widespread maximum near the Pole, larger than for a few years before and after. The reason is not understood; however, Renfrew *et al.* (2019) suggested that surface pressure is sensitive to surface momentum exchange and found that changing the

surface drag parametrization to more realistically depend on sea-ice concentration can cause pressure differences of this magnitude.

Figure 6 similarly explores model performance for some of the surface-energy budget terms. The first hourly value is always missing; parametrized variables are only calculated during the forecast and are undefined in the analysis. Unlike the previously inspected near-surface state variables, these errors seem to all have a distinct 6-hourly periodicity of unknown origin. Some of this is an artefact from a combination of rare but large and consistent errors and the evaluation procedure; their origins are not understood at present. Appearing consistently at the same actual time in all consecutive forecasts initiated 6 hr apart, one single event reappears 12 times in the error analysis: once per day for three forecasted days and four initializations per day. When evaluating errors as a function of forecast length, the same occasion will appear as a 6-hourly periodicity. However, even after removing these few very large errors, some 6 hr periodicity remains; this will not be explored further in this study.

Consistent with the wind speed error, friction velocity (Figure 6a) is slightly too large initially and grows with forecast length; the error spread also increases with forecast length. This is very likely due to the wind-speed error; roughness-length may also play a role, here coming from the sea-ice model. Turbulent sensible heat fluxes are too low on average by about 8–9 W·m^{-2} (Figure 6c). Both the error and its spread are nearly constant. This is due to a combination of too small upward and too large downward fluxes; hence, too much energy is transferred from the atmosphere to the surface. The turbulent latent heat flux is slightly too large; this error is not significant (Figure 6d).

The error in net surface long-wave radiation is not significantly different from zero (Figure 6b); however, incoming long-wave radiation from the too warm model atmosphere is, as expected, too large by 5 W·m^{-2} (Figure 6e). Incoming solar radiation is too small by ~5 W·m^{-2} (Figure 6f); both these errors are significant. Unfortunately, it is impossible to observe net surface solar radiation from a ship. Although there were albedo observations from the ice, these are for a shorter time period and are local, not representing the areal-average albedo a model needs.

From Figure 5 it is clear that both atmosphere and surface skin temperatures are too warm. Closer inspection shows that the median temperature bias in the atmosphere is larger than that of the surface by ~0.2 K during days two and three of the forecast; however, it starts out opposite but changes sign during the first ~12 hr. Hence the net long-wave forcing error on average remains close to zero while the deficit in incoming short-wave solar radiation is smaller than the excess energy transferred to the

surface by the turbulent heat flux. This indicates that the surface is too warm because of energy transferred to it from the atmosphere, and not the other way around. The warm-biased air temperatures therefore do not appear to rest with the surface energy budget, but with some other problem causing the lower atmosphere to be too warm. In this perspective, the unphysically warm (>0 °C) surface temperature during the melt season is a noticeable but less important problem, while the too warm lower atmosphere is concerning.

3.2 | Exploring the vertical structure of errors

In this section we use the 6-hourly soundings to explore the vertical structure of some of the model errors discussed in the previous section. Figure 7a shows a distinct vertical structure in the median temperature error as a function of forecast time. Below 200 m the model is up to ~1 °C too warm, but around 300–400 m the bias changes sign and between 500 m and 2 km the model is too cold, at ~1 km by >1 °C. Initially, the model is close to observations, again likely due to assimilating the local soundings, but the low-level warm bias establishes within <6 hr and variations after this appear random. The cold bias centred around 1 km develops more slowly; however, most of the error is established after 24 hr, although the depth of the cold-bias layer continues to increase. Above 3 km the error again changes sign and the model is slightly too warm up to ~10 km; this error grows more slowly, reaching ~0.5 °C at 4–5 km on day three. In the uppermost troposphere/lower stratosphere there is an increasing cold bias. The specific humidity error (Figure 7b) is consistent with the temperature error: too moist in the lowest layer and too dry between 500 m and 3 km; aloft, specific humidity becomes so low that defining an absolute bias becomes pointless.

The wind-speed error (Figure 7c) below 100 m is much larger (>2 m·s^{-1}) than the near-surface wind-speed error in Figure 5c. A large fraction of this is very likely due to sondes being caught in the lower-speed wake behind the superstructure of the ship. Above 50–100 m and up to ~2 km the wind speed bias is close to zero initially, increasing with forecast time reaching ~0.5 m·s^{-1}, consistent with Figure 5c. Above ~3 km the bias is again positive, reaching 1–1.5 m·s^{-1} around 5 and 9 km in separate maxima. Wind direction errors are small below 5 km during the first 1.5 days (Figure 7d), smaller than the 10° bias in Figure 5d; then they increase but stay at ~10°. Note that the winds from the soundings away from the surface are independent from measurements on board, affected by the ship's flow distortion only in the lowest 50–100 m.
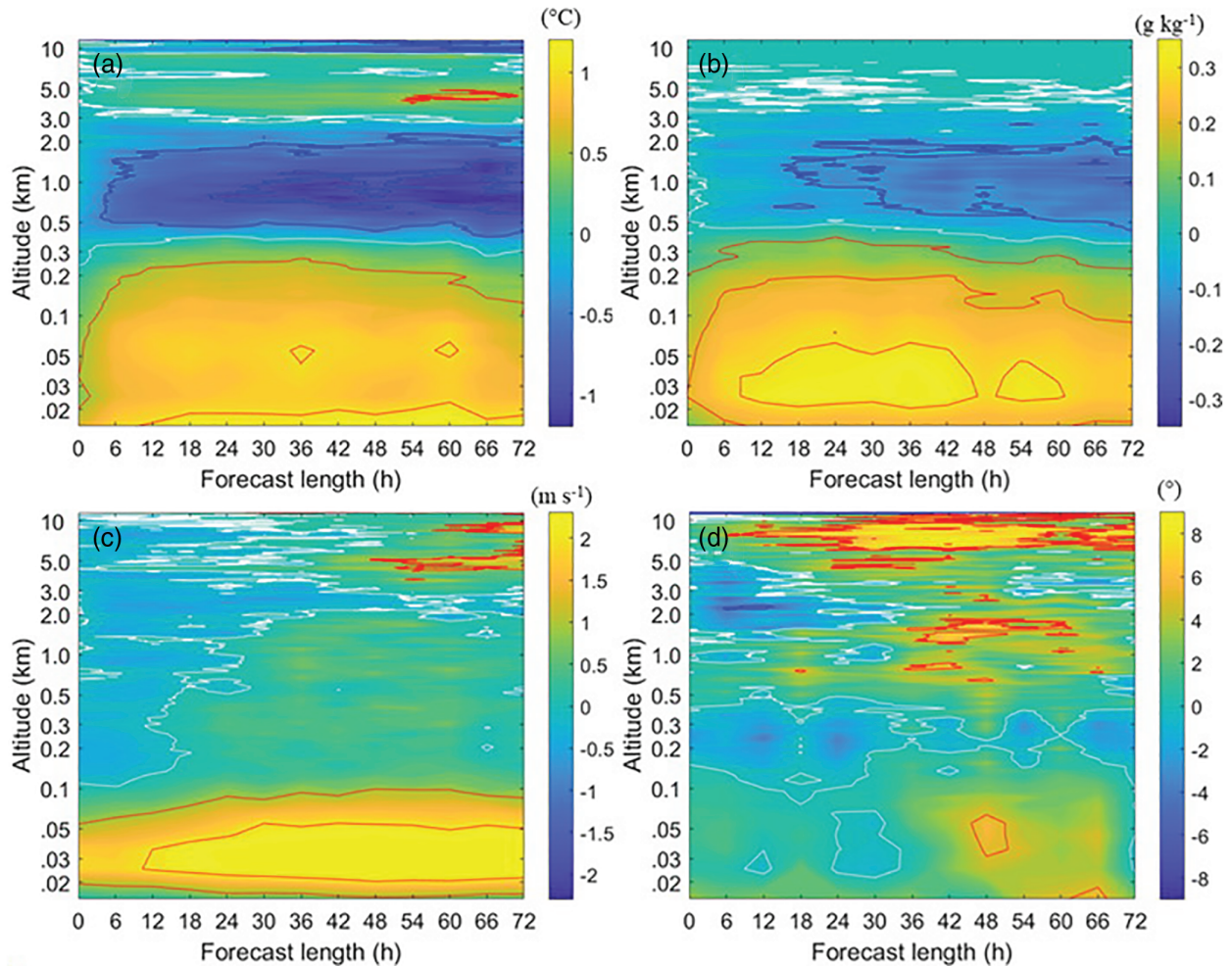
FIGURE 7 Time–height (hr–km) contour plots of median forecast errors using all forecasts during the AO2018 ice drift: (a) temperature (°C), (b) specific water vapour (g·kg⁻¹), (c) scalar wind speed (m·s⁻¹), (d) wind direction (°). Red, white and blue isolines outline positive, negative and zero errors, respectively, in (a) at 0.5 °C, in (b) at 0.1 g·kg⁻¹, in (c) at 1 m·s⁻¹ and in (d) at 5° intervals. Note the logarithmic vertical scale

Combining moisture and temperature, the lowest-layer bias in equivalent potential temperature (Figure 8a) reaches ~2 °C, while the lower free troposphere cold bias around 1 km reaches −2 °C at day three. This significantly changes the moist-static stability in the lower half of the atmosphere and should have consequences for parametrized convection, especially mid-level convection not initiated at the surface. For relative humidity (Figure 8b), temperature and moisture errors compensate for a <±3% error, positive below 1 km and negative around 3–5 km; these errors are within the measurement accuracy.

3.3 | Errors over different time-scales

Figure 4 indicates that the near-surface temperature error is larger after the onset of the freeze-up. Figure 9 shows the vertical structure of the temperature error, similar to

Figure 7a, separating forecasts into before (Figure 9a) and after (Figure 9b) DoY 240, spanning the seasonal change with 47 and 53% of the data before and after, respectively. The vertical structure is similar but with larger variability due to the smaller sample. However, the magnitude of the error is substantially larger after the freeze-up in all three layers; the lowest-troposphere error goes from <~1 to >1.5 °C, while the thickness of the too warm layer shrinks slightly from ~500 to 300 m. The lower free troposphere cold bias goes from −1 to −1.5 °C, while the mid-troposphere error changes less. Hence, the warm bias in the lowest layer is weaker but deeper during the end of the melt season and stronger but more shallow as the freeze-up has started.

Analysing errors as a function of forecast length using forecasts initiated at different times during each day effectively averages over local time and hides any potential diurnal signal. In Figure 10 this is circumvented by
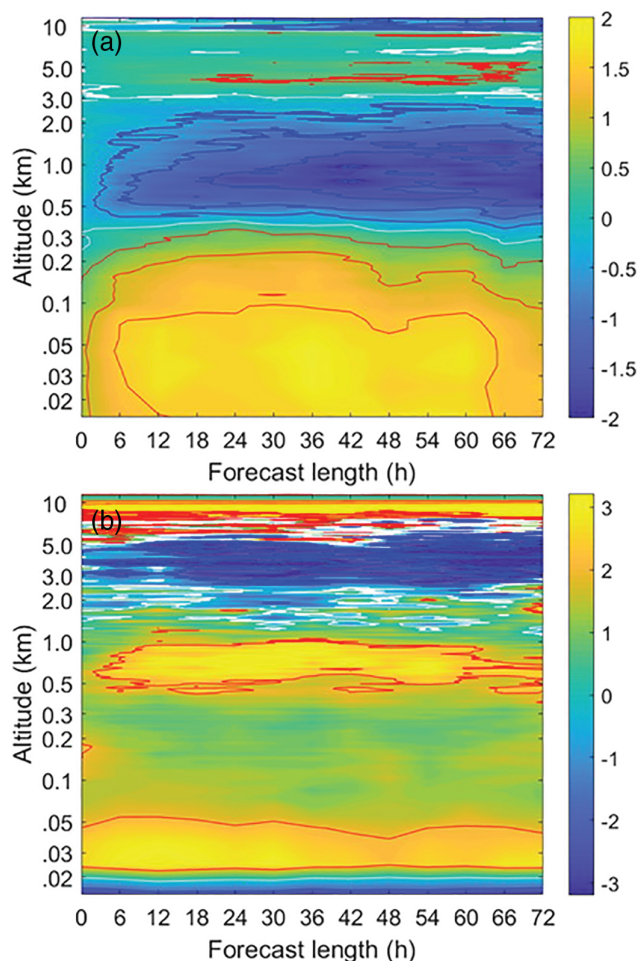
FIGURE 8 Same as Figure 7 but for (a) equivalent potential temperature (°C) and (b) relative humidity (%). Red, white and blue isolines outline positive, negative and zero errors, respectively, in (a) at 0.5 °C and in (b) at 2% intervals

FIGURE 9 Same as Figure 7 but for temperature (°C) (a) before and (b) after DoY 240 when the freeze-up period is assumed to have begun

evaluating forecasts initiated at different times of the day separately; each evaluation now uses only a quarter of the forecasts. A clear near-diurnal cycle now becomes evident in the temperature error. In forecasts initiated at 0000 UTC (Figure 10a), the errors are at maximum at 24, 48 and 72 hr. In forecasts initiated at 1200 UTC (Figure 10c) the same pattern emerges, only shifted 12 hrs earlier in the forecast; at 12, 36 and 60 hr into the forecast, the same local time as for the 0000 UTC forecasts. The same pattern appears in the forecasts initiated at 0600 and 1800 UTC, shifted 6 and 18 hr compared to the 0000 UTC forecast. There also seems to be a corresponding cycle aloft, in the 0.5-to-2 km cold-bias layer. The error magnitudes are in phase but the signs are out of phase, also affecting lower troposphere static stability.

The timing of the diurnal peaks in Figure 10 are somewhat smeared and the cycle does not always appear exactly at 24 hr. However, first, early in the forecast any diurnal cycle is muted by the initial error growth. Second, there
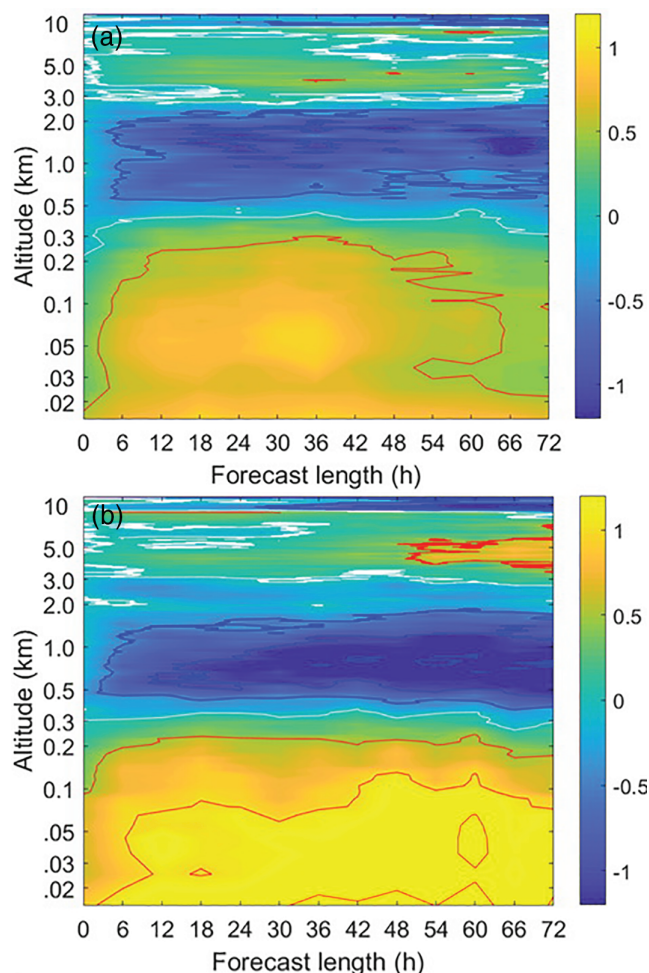
is a local timing difference relative to UTC across the ice drift. While *Oden* was drifting, the longitude shifted and therefore the *true* local time (LT) in *each* forecast deviated from the *mean* LT over *all* forecasts by as much as ±2 hr; on average LT was about 2 hr ahead of UTC. Finally, with a 6 hr time resolution in both model output and soundings, smearing of a diurnal cycle is expected.

Combining both time aspects discussed above, Figure 11 show the same as Figure 9 but only for the forecast initialized at 0000 UTC. Note that the number of forecasts in each evaluation is now down to only about 15. There appears to be a weak diurnal cycle of the error during the end of the melt season (Figure 11a), but a much more pronounced cycle appears after the freeze. While the median amplitude of the temperature error cycle is ±~0.3 °C for the whole ice drift, it doubles to ±~0.6 °C during the freeze-up. To explore this further, we use near-surface temperature forecasts but calculate the median error separately for forecasts initiated at different
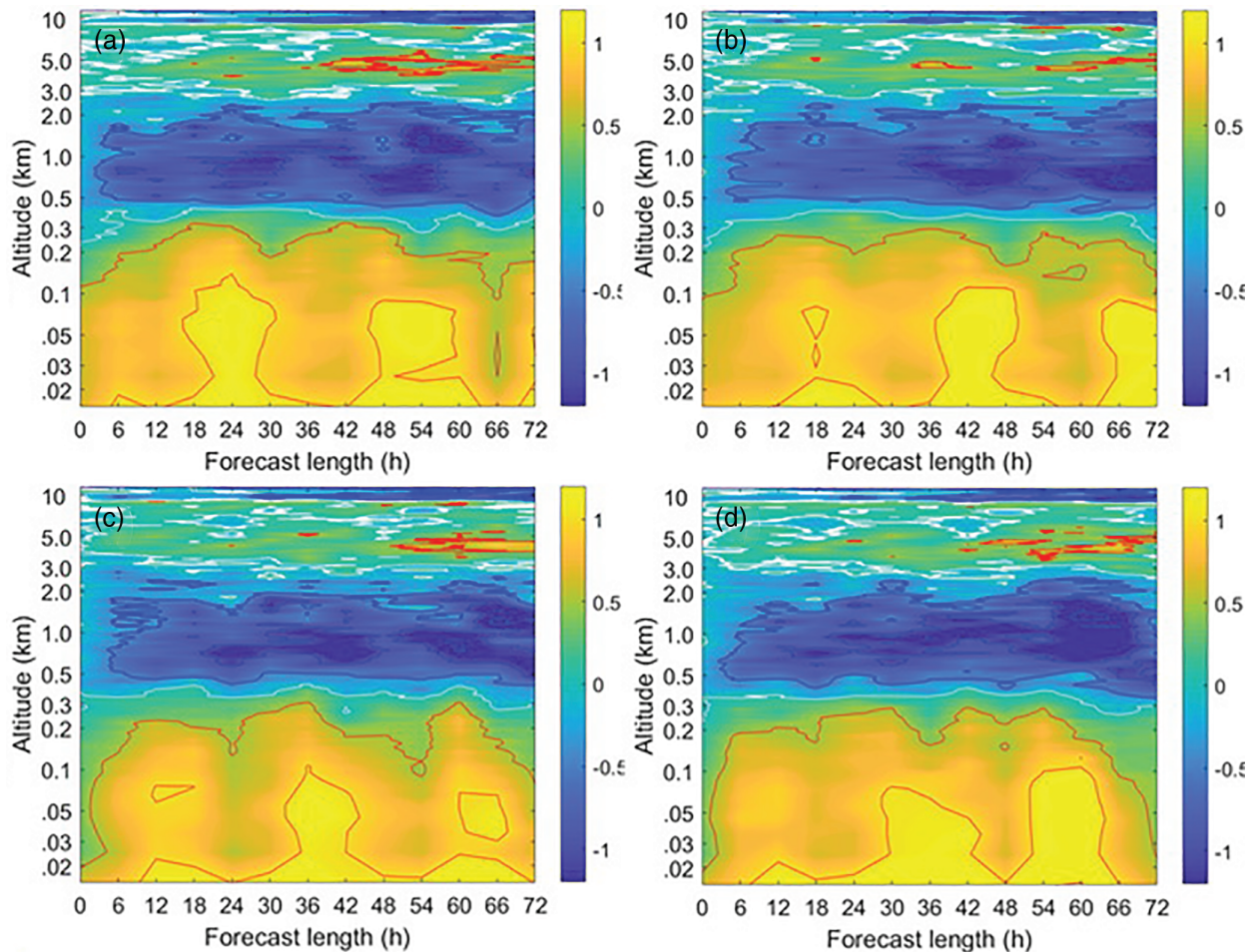
FIGURE 10 Same as Figure 7a but separately for forecasts initialized at (a) 0000, (b) 0600, (c) 1200, and (d) 1800 UTC

times. We then plot the results shifted in time by 6 hr, to correspond to the same LT regardless of when the forecast was started. Figure 12a shows the end of the melt and Figure 12b the freeze-up. During the melt, the forecasts errors do have distinct peaks at 24 hr intervals: at 12, 36, 60 and 84 hr, corresponding to early afternoon. However, variations are not sinusoidal; instead they have a strange spike-like appearance, exceeding a baseline by ∼0.4–0.5 °C every 24 hr. During the freeze-up, errors grow larger and noisier. It appears, however, that the diurnal cycle in the temperature errors has grown to an amplitude of ±∼0.8 °C, consistent with the results from the soundings, but the timing of the maxima is different from during the melt, at 24, 48 and 72 hr after 0000 UTC, which would be in the early morning.

The diurnal cycle in the solar forcing in the summer Arctic is weak and therefore we expect only a weak diurnal cycle in near-surface temperature from observations (Tjernström, 2007). However, exploring this in the presence of larger synoptic or sub-synoptic variability is difficult. A diurnal cycle in the *error* can be due to the presence of one in the model, absent or very weak in reality – or the opposite. There can also be a diurnal cycle in both, but too strong or out of phase in the model. We analysed the temperature cycle in the observations by first adjusting observations from UTC to LT, then resampling and high-pass filtering to retain variability corresponding to diurnal and shorter fluctuations. When averaging these according to time of the day (not shown) we do find a median near-surface temperature diurnal anomaly for the whole ice drift with a weak, <±0.05 °C amplitude, consistent with Tjernström (2007). When exploring the melt and freeze periods separately, the cycles are out of phase and stronger in both, still only <±0.1 °C. The maximum temperature during the melt (freeze) was at 0200 LT (1700–2000 LT) while the minimum was at 1300 LT (0100 and 1100 LT). Hence, the diurnal cycle in the observations is much smaller and out of phase with that in the model error.

The presence of a diurnal cycle in the error for other variables is more difficult to detect. An obvious source of a diurnal cycle is the radiative forcing. Although noisy, there seems to be a signal in the error of incoming short-wave radiation (Figure 13a), peaking at −8 W·m^{-2} around 9, 33, 57 and 81 hr, relaxing to near-zero at a 12 hr lag. This
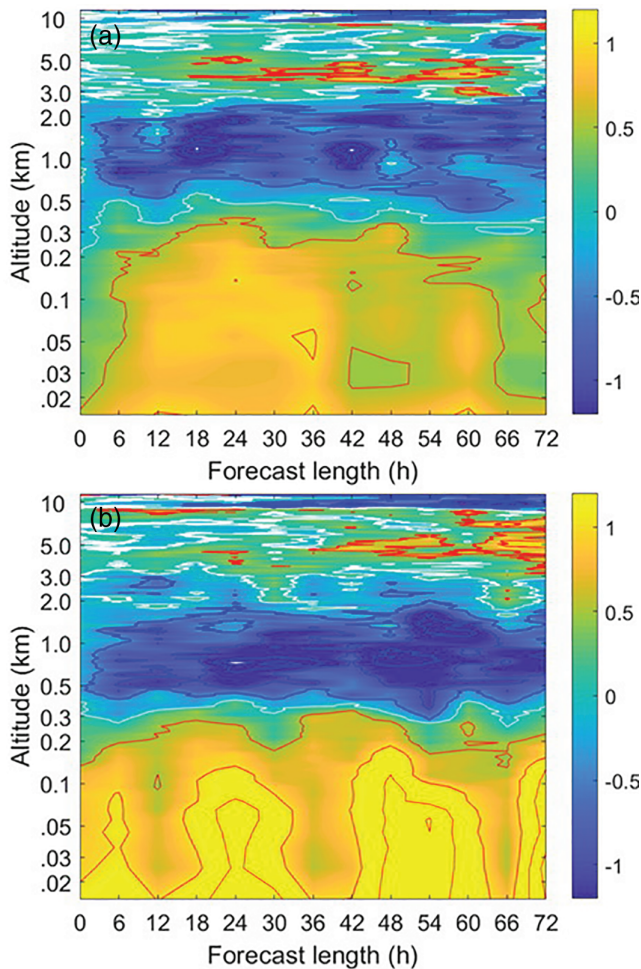
FIGURE 11 Same as Figure 7a but only for the forecast initialized at 0000 UTC and split between (a) before and (b) after DoY 240, when the freeze-up period is assumed to have begun

FIGURE 12 Median near-surface air-temperature error (°C) calculated separately for forecasts initialized at different times of the day (hr UTC), plotted against forecast time (hr) relative to 0000 UTC for the first day, split between (a) before and (b) after DoY 240, when the freeze-up period is assumed to have begun. The dashed line is the mean error of all forecasts over the overlapping period

is likely caused by an error in the cloud attenuation. The error in net long-wave radiation (Figure 13b) is also noisy but also features a diurnal cycle, in sync with that for incoming short-wave radiation. Assuming that any error in surface albedo has no diurnal component, there is hence a diurnal variability in surface radiative forcing error with maximum forcing around 1900 LT and minimum around 0700 LT. Using the median longitude of the ice drift, this is consistent with a maximum lower-layer temperature error 6 hr after the forcing maximum, close to the observed periodicity in Figure 10.

3.4 | Errors in the representation of clouds

The vertically integrated cloud condensate, the cloud liquid and ice water paths (LWP and IWP), were observed using microwave radiometry and cloud radar reflectivity, respectively; the uncertainties are typically $\pm 0.02 \, \text{kg} \cdot \text{m}^{-2}$
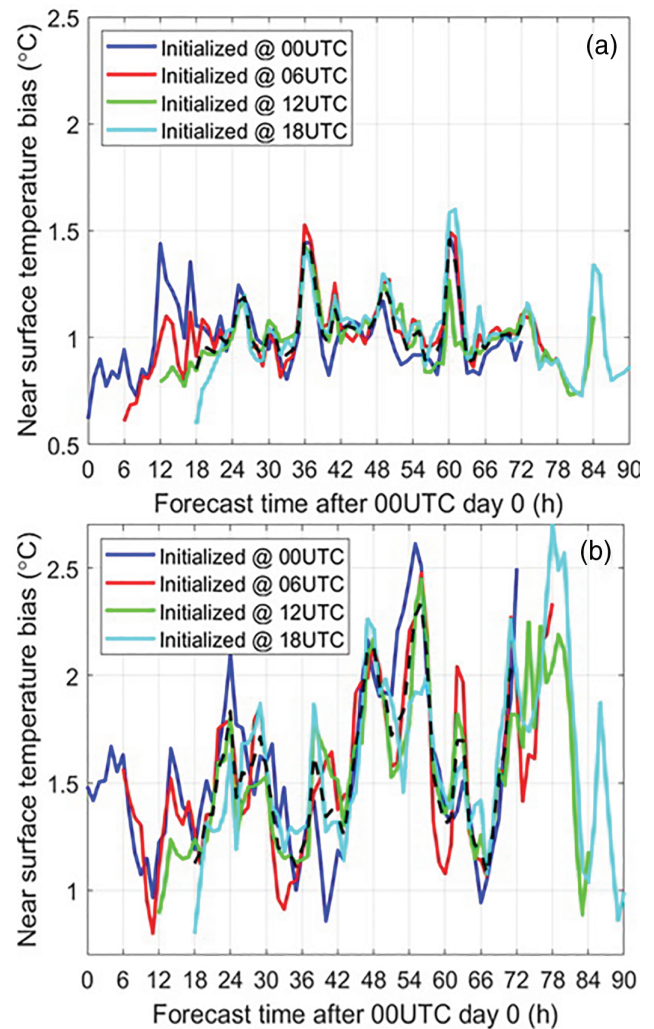
for LWP (Westwater *et al.*, 2001) and a factor of two for IWP (Shupe *et al.*, 2005). Displaying the error distribution for these two bulk cloud variables (Figure 14) suggests that the model has too much cloud water, both liquid and ice. The overestimation is substantial; the median errors are ~0.04 and ~0.008 $\text{kg} \cdot \text{m}^{-2}$ in LWP and IWP, respectively; with median values at 0.11 and 0.015 $\text{kg} \cdot \text{m}^{-2}$ the overestimation is 36 and 53%, respectively.

To evaluate the vertical structure of cloud errors we defined a "cloudiness error" parameter. Using cloud radar reflectivity to indicate cloud presence, we set this parameter to zero everywhere the model and the radar agree on absence or presence of clouds. If the model indicates clouds and the radar does not it is set to unity, while if there is a cloud in the radar but not in the model, it is set to minus
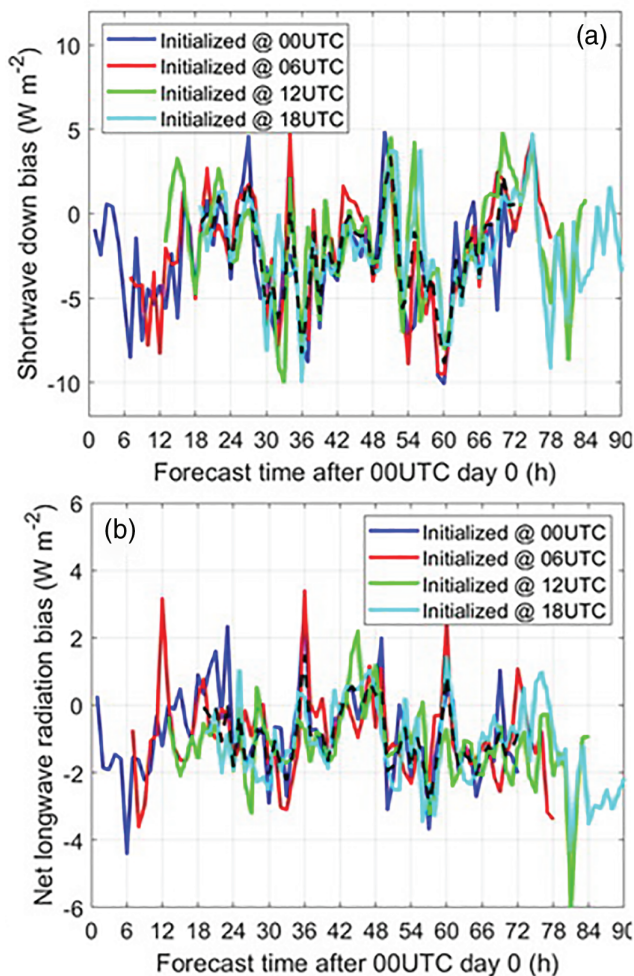
FIGURE 13 Same as Figure 12 but for forecast errors for the whole ice drift of (a) downwelling surface short-wave and (b) net long-wave radiation (W·m^{-2})

unity. This method does not indicate any magnitude of the error; however, when averaged over time and over many forecasts it still contains some quantitative information; larger than zero averages can be interpreted as too much cloud and vice versa. Figure 15 indicates a substantial overestimation of clouds below 3 km, with a maximum in the 400–800 m layer, larger initially but gradually improving somewhat with forecast time. This is consistent with the discussion of Figures 2f and 3, where IFS often had a too thick and persistent cloud layer. In a shallow layer of growing depth close to the surface, conditions gradually improve, while in the 1–3 km layer, the mean error remains essentially unchanged through forecasts. Farther aloft, the results indicate a lack of clouds initially that grows worse with forecast time.

Figure 16 shows profiles of cloud liquid- and ice-water content averaged across the whole ice drift excluding all clear values; the observations use the Cloudnet retrieval (Achtert *et al.*, 2020; Vüllers *et al.*, 2020). The magnitudes

of cloud water in the model when clouds appear are reasonable, compared with the retrieved values, with a slight overestimation of cloud ice below ∼2 km and an even smaller underestimation around 4–7 km. The highest values of liquid water in the lower troposphere around 0.1 g·kg^{-1} are similar in the IFS and the retrieval. The main difference is an overestimation of cloud liquid in the 1–3 km layer: the upper fraction of the thick layer where clouds appear too often in Figure 15. The cloud phase partitioning in IFS is also in agreement with the observations, although with much more layering in the observations. Hence clouds in IFS are at least on average realistic; they just appear too often below 3 km, consistent with the too large LWP.

We also analyse the cloud layering in Figures 17 and 18. Directly comparing the geometry of individual cloud layers is very difficult because of differences in how they are defined in observations and in the model; what is two cloud layers in one could in the other be a single layer without there being any real difference in the physics. Therefore, we compare the separate cloud-layer statistics for observations and the model. The number of cloud layers are quite similar between the observations and the model (Figure 17). Single-layer clouds dominate in both, at 36 and 39%, respectively, with decreasing occurrence for multiple layers, for two-layer systems 25 and 35%, respectively. There are more cases with >3 cloud layers in the observations than in the model, expected from the higher cloud-radar vertical resolution. An important difference is for clear conditions. This happens only 3% of the time in the observations, but the model has no cases without at least one cloud layer, consistent with Figure 3b. The lack of cloud-free cases in the model forecast is also consistent with anecdotal experience (e.g. Tjernström *et al.*, 2019).

Yet another perspective on this is offered in Figure 18, showing the statistics for the lowest cloud-base height and the thickness for the three lowest cloud layers, counting from below. All results are scaled to the total cloud occurrence, hence, cumulative probability for a lowest cloud layer tends to 100%, while for the second and third layers it is proportional to the occurrence of more cloud layers. In the observations, the lowest cloud base is most often below 200 m (Figure 18a); finding a lowest cloud base >400 m is an order of magnitude, and >2 km two orders of magnitude, less likely. This is in line with previous late summer/early autumn central Arctic studies (e.g. Tjernström, 2005; 2007; Tjernström *et al.*, 2012; Sotiropoulou *et al.*, 2016b). In the model (Figure 18b), a lowest cloud base below 200 m dominates even more than in the observations; there are almost no cases with a lowest cloud base above 300 m. When a second cloud layer is observed, this often appears around ∼200 m indicating that boundary layer clouds were often thin and multi-layered; however, a
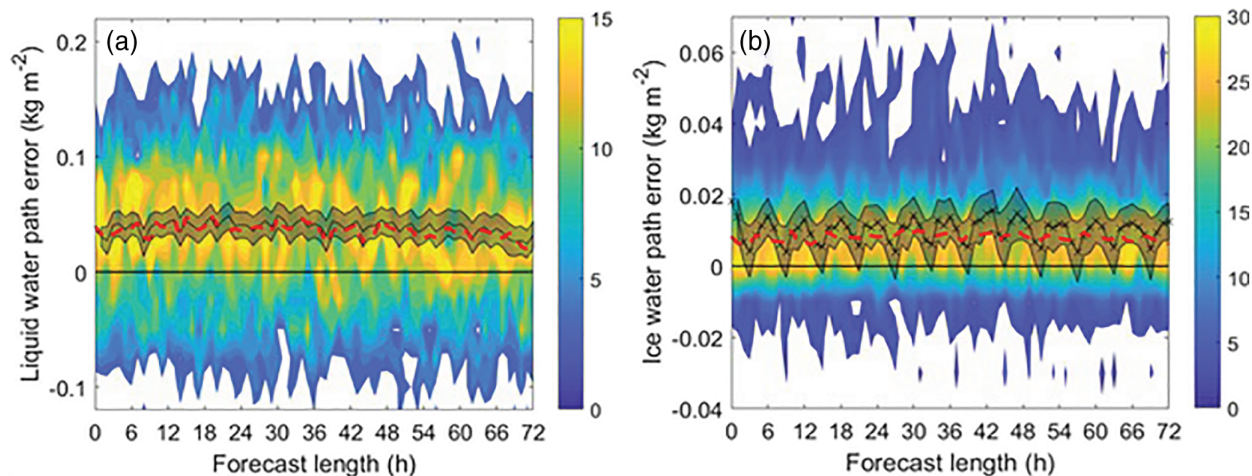
FIGURE 14 Same as Figure 5 but for (a) cloud liquid water path and (b) cloud ice path, both in kg·m^{-2}
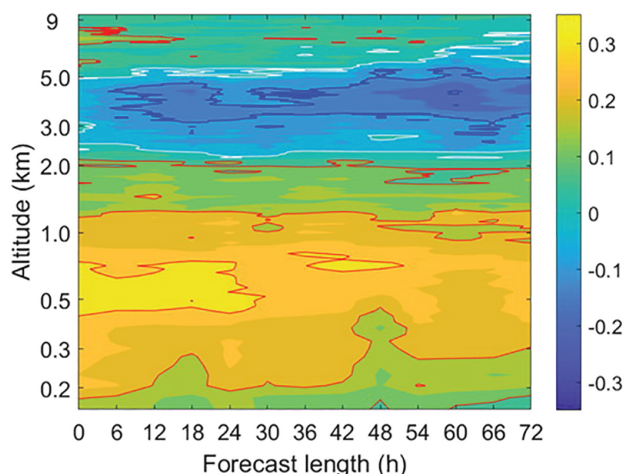

FIGURE 15 Time–height cross-section of cloud occurrence agreement as a function of forecast length (hr) and altitude (km), see the text for a discussion. The red (blue) solid lines are for every 0.1 (−0.1) while the white line indicates zero error

higher second layer is about equally likely to appear over a broad range 1–8 km. If a third layer is present, it almost always appears above 1 km. Second cloud layers in the model most often happens around 1 km, although there is also some probability for a second cloud base at ~100 m, not present in the observations; third layers are too few for stable statistics.

Cloud thickness probability distribution in observations (Figure 18c) is very similar for all the first three cloud layers, peaking at <300 m; deep clouds, thicker than 4 km, are associated with single-layer frontal clouds and hence appear less often but mostly as the first cloud layer. While the model has a lowest absolute cloud-thickness peak at ~100 m, it has a broad secondary but dominating peak at 0.5–2 km (Figure 18d); the first layer is hence quite often much thicker in the model than observed. Also,

second cloud layers are often thicker than in observations, ~0.2–1 km. In general, clouds in the model are thicker than in observations.

4 │ DISCUSSION

Many of the errors diagnosed in this study are systematic, with a very rapid growth over the first 6–12 hr, then becoming quasi-constant; the error spread is also roughly constant. It thus appears that, at least for the thermodynamics, the inherent model climate is different from observations. However, for dynamic variables, such as wind or pressure, assimilation keeps the model true at initialization. It is important to realize that the well-behaved dynamics allows the closer study of the thermodynamics. Hence, assimilation of observations later used for model evaluation is an advantage. Without this, the initial state of the model might have been sufficiently far off from reality that it would become difficult to separate random and systematic errors.

Both the reanalysis and the forecast versions of IFS have previously been found to have a near-surface warm bias (e.g. Jakobson *et al.*, 2012; de Boer *et al.*, 2014; Sotiropoulou *et al.*, 2016a); in particular, the above melting-point near-surface temperatures during summer melt was pointed out by Sotiropoulou *et al.* (2016a) and Wesslén *et al.* (2014). This is also where this study started (Figure 4), assuming that this is due to errors in either the coupling to the sea ice or in the surface energy budget. That also the surface skin temperature is above the melting point, not discussed by previous studies, seems to indicate a problem with the energy budget. Our results, however, indicate that even with too little solar radiation reaching the surface there is still an excessive turbulent sensible heat flux from the atmosphere to the surface.
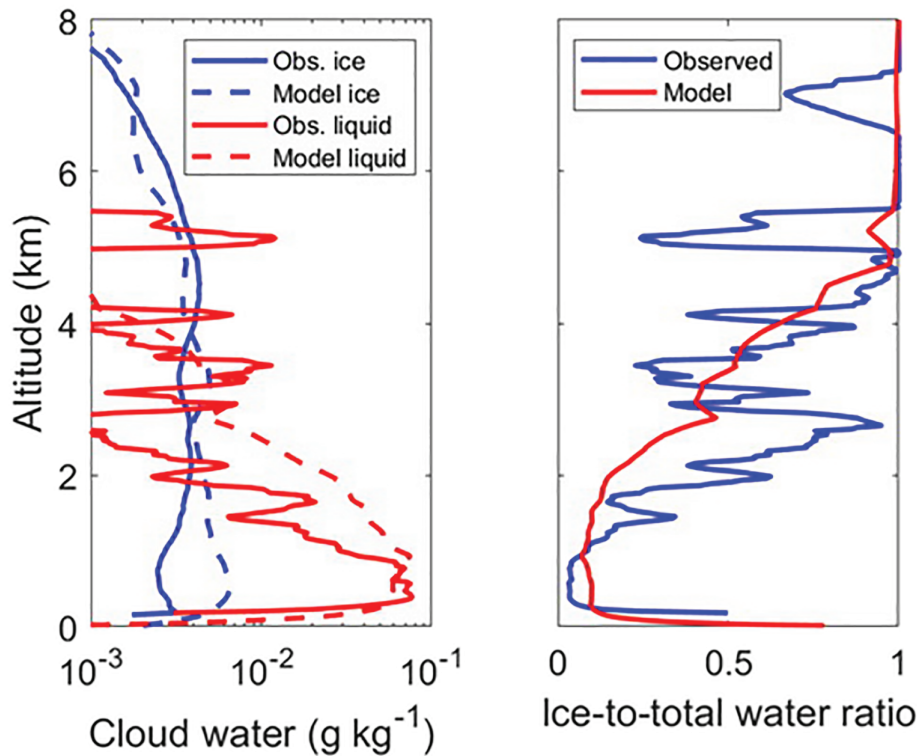
FIGURE 16 Profiles of mean modelled and observed (a) liquid and ice cloud water (g·kg^{-1}) and (b) ice to total cloud water ratio for the AO2018 ice drift period
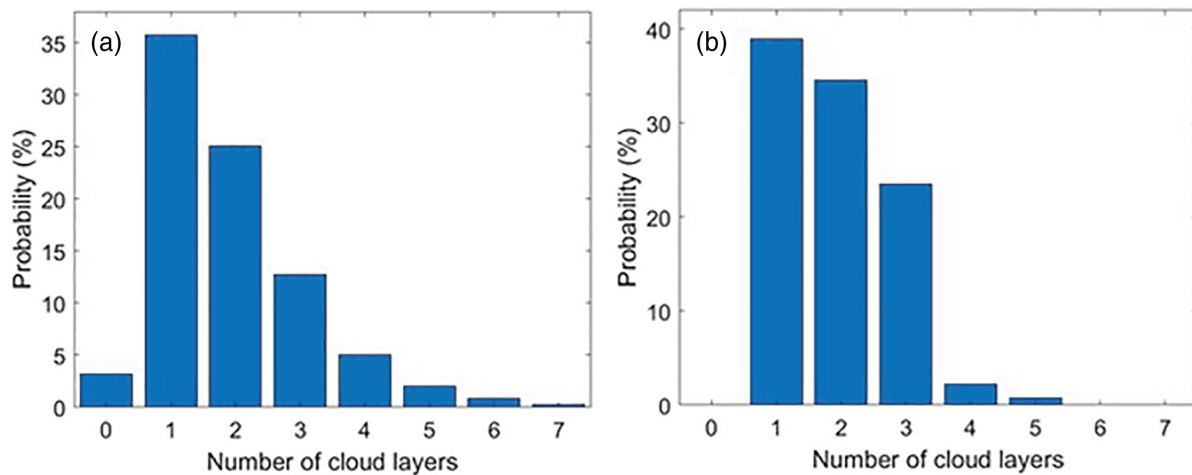

FIGURE 17 Histogram showing the frequency of occurrence for the number of cloud layers in (a) observations and (b) model

We therefore conclude that, while the unphysical >0 °C surface temperature during melt may be a sign of a coupling problem, the only way this error can be sustained over time is if it has its roots in a too warm model atmosphere, incompletely held back by the surface during the melt; this is consistent with larger errors when the surface temperature is well below freezing.

Both Jakobson *et al.* (2012) and Wesslén *et al.* (2014) pointed out that the warm bias was not confined to near the surface but occurred over a layer. We find that the two most striking errors, in temperature and cloudiness, have consistent and partly coherent vertical structure. The

average error profiles of cloudiness (Figure 15) and temperature (Figure 7a) are almost mirror images (Figure 19). The largest cold bias appears between 0.6 and 1.4 km, while the largest overestimation of cloudiness appears in the 0.4 to 1.2 km layer. Further exploring this relationship, Figure 20 shows the temperature error distribution sampled according to cloud liquid water content; cloud water typically increases toward the cloud top. For low cloud liquid water content temperature errors are distributed roughly around zero, but for increasing cloud liquid water the error gradually shifts toward negative values. For cloud liquid water content >∼0.2 g·kg^{-1}, the median
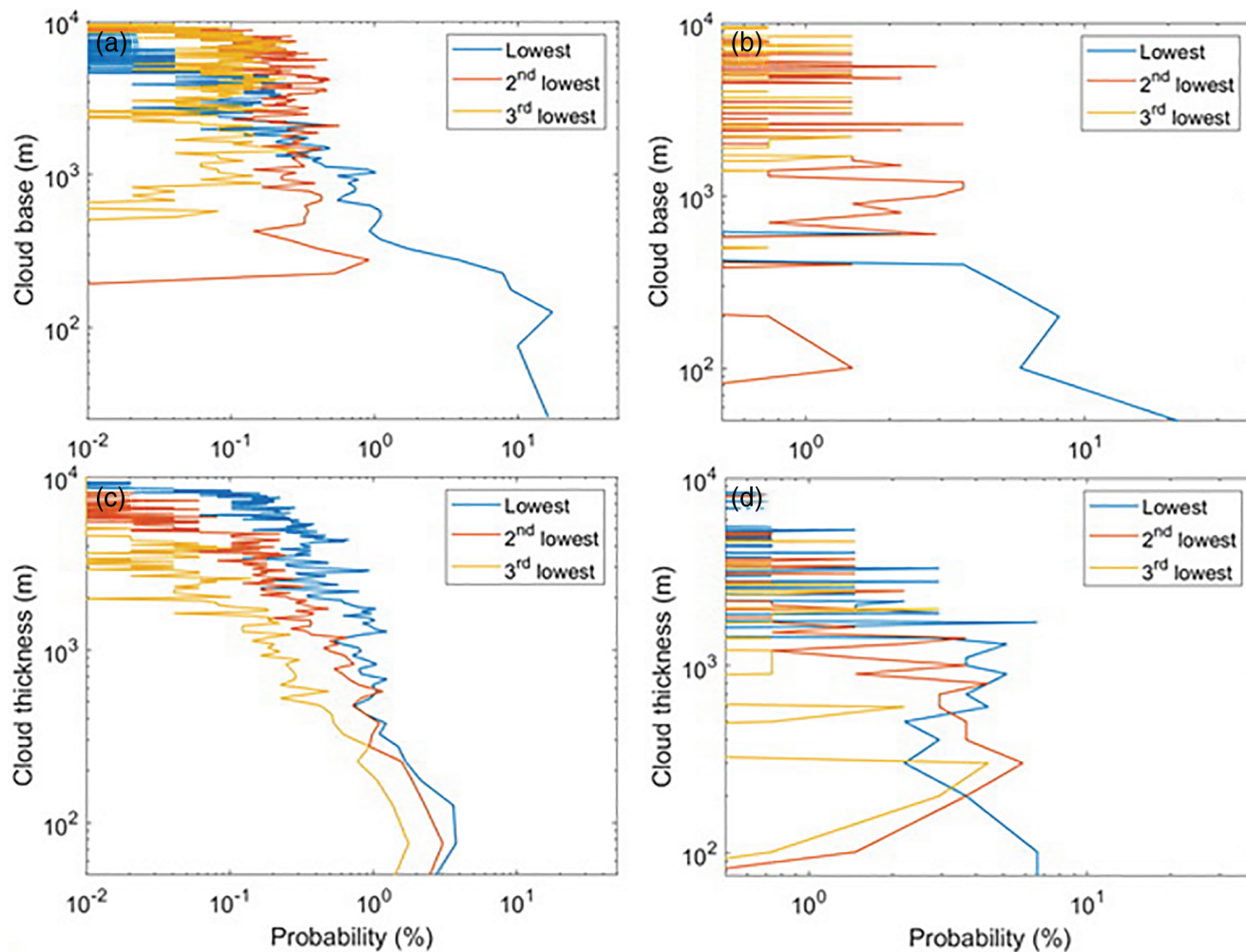
FIGURE 18 Statistics of cloud boundaries showing the relative probability (%) of (a,b) the cloud base and (c,d) thickness of the three lowest cloud layers; (a,c) are observations, (b,d) are from the model. Note the logarithmic scales

temperature error is $< -3\,°C$. For the largest cloud liquid water values, the distribution becomes increasingly complicated, partly because of sampling, but at ∼0.4 g·kg^{-1} the median error is around $-5\,°C$. A link between cloudiness errors and the lower troposphere cold bias maximum seems obvious. Attributing one error to the other is more difficult; are clouds overestimated because of the cold bias or is the cold bias due to cloudiness errors?

If the maximum cold bias was located at the top of a deep well-mixed boundary layer, both it and the excess cloudiness could be a consequence of too vigorous turbulent mixing. The thermodynamic state near the surface is governed by the surface energy budget, while temperature and cloudiness at the top of a well-mixed boundary layer is given by adiabatic processes, enforced by mixing. With a too deep well-mixed layer, cloud-top cooling would provide positive feedback; more clouds lead to additional cloud-top cooling, increased negative buoyancy, cloud-overturning turbulence and more

mixing that maintains or deepens the well-mixed layer and hence leads to more cloud condensation. However, analysing surface-based mixed-layer depths in the model indicates that the median difference in altitude between the cold-bias maximum and corresponding mixed-layer tops is ∼700 m, although <200 m for ∼25% of the time (not shown). Hence, the top of the clouds is more often than not decoupled from the surface. Therefore, excessive vertical mixing cannot be the cause of the cold bias.

At this point it is useful to discuss the atmospheric boundary layer (ABL); a term that we have, until now, tried to avoid. Brooks *et al.* (2017) struggled with this in the context of stratocumulus decoupling (cf. e.g. Shupe *et al.*, 2013; Sotiropoulou *et al.*, 2014). As a compromise they use ABL for the whole layer below the main inversion, separating it into a surface-mixed layer (SML) forced by surface friction, and a cloud-mixed layer (CML) forced by cloud-top cooling. Using slightly different techniques and metrics, Brooks *et al.* (2017), Shupe *et al.* (2013) and
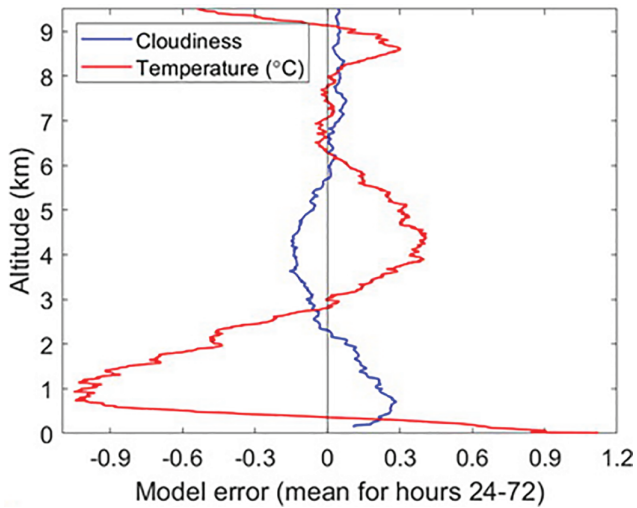
FIGURE 19 Profiles of the model "cloudiness" and the temperature errors, averaged over all forecasts and forecast times, excluding the first 24 hr
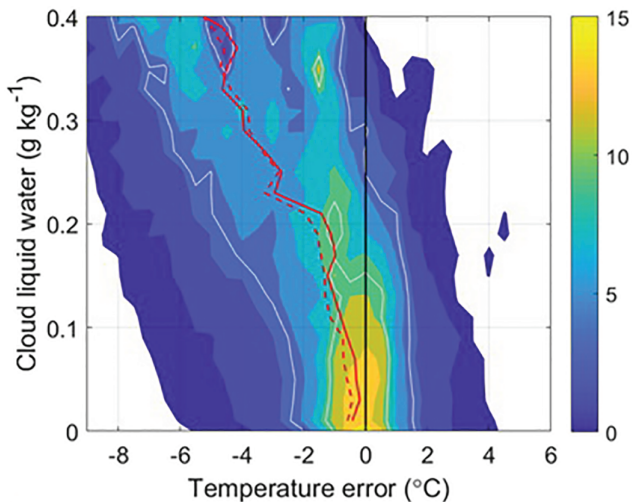

FIGURE 20 Probability (colour shading) of temperature error (°C) as a function of total cloud water (g·kg^{-1}) for the whole AO2018 ice drift period, with the median (solid red) and mean (dashed red); white contour lines show 5 and 10% probability

Sotiropoulou *et al.* (2014) suggested that these layers are decoupled about two-thirds of the time.

In this context, one may hypothesize that the lowest too-warm layer is the SML and that the low troposphere too-cold layer corresponds to the CML. The CML is too cold because of excessive cloud-top cooling and a lack of vertical mixing; with more mixing the intense cooling at the cloud top would be mixed over a deeper cloud and sub-cloud layer and SML and CML would eventually connect. Sotiropoulou *et al.* (2016a) used IFS simulations of the ASCOS expedition and concluded that IFS, with its first-order closure turbulence

scheme, is unable to respond to turbulence generated by cloud-top cooling and hence could not develop decoupled stratocumulus.

However, this does not explain why there is too much cloud in such a deep layer. To explain this, we turn to a specific parametrization in the IFS: so-called mid-level convection. This scheme triggers convection from anywhere in the troposphere, provided sufficiently low moist-static stability, relative humidity >80%, and large-scale ascending motions. The relative humidity in the Arctic ABL rarely drops below 80% (e.g. Tjernström, 2005, Tjernström *et al.*, 2014; Vüllers *et al.*, 2020; also see Figure 2e), while Figure 7a shows that the particular vertical structure of temperature and moisture errors in IFS contribute on average to a reduced moist-static stability; this reduction also has a diurnal cycle.

We hypothesize that if this stability by chance is sufficiently low at the same time as large-scale ascent is present, mid-level convection transports water vapour out of the ABL that condenses to clouds as the air is cooled, leading to excess cloud-top cooling while absorption of solar radiation excessively warms the cloud interior; the latter is consistent with too low incoming solar radiation at the surface and the diurnal cycle in the temperature error. This reinforces a vertical error structure that triggers mid-level convection too often, providing the positive feedback to sustain the vertical structure in the temperature error.

The mid-level convection is a construct to allow convection in rain bands at warm fronts and in the warm sector of extratropical cyclones (see https://www.ecmwf.int/en/elibrary/19308-part-iv-physical-processes, Chapter 6.4.3) and its design may be inappropriate in the Arctic with its very moist and shallow ABL. This is, however, very difficult to ascertain from the model results; the only way to test this hypothesis is to run the model with and without this mechanism active, which is outside the scope of this article.

5 | CONCLUSIONS

We evaluate operational ECMWF/IFS HRES forecasts using an extensive observational dataset from the Arctic Ocean 2018 expedition, deployed on the Swedish icebreaker *Oden*. The evaluation covers a month-long period when *Oden* was drifting with the ice close to the Pole, mid-August to mid-September, spanning the late-summer melt and early-autumn freeze conditions, and includes 125 three-day forecasts issued 6-hourly. Three-hourly routine surface observations and 6-hourly soundings were assimilated to provide the initial state for the forecast. The atmospheric model is essentially the same as in ERA5; hence

results provide information to both model developers and users of the IFS forecasts and of ERA5.

Most errors in the thermodynamics appear systematic and do not grow with forecast length; error spread is also constant. There is, however, a rapid growth the first 6–12 hr; we attribute this to the data assimilation and subsequent spin-up. For a few evaluated variables, related to dynamics, the error growth is more linear and then, error spread also grows with time.

A summary of our main findings is:

1. There are considerable temperature errors featuring distinct vertical and temporal structures:
   (a) IFS is too warm below ∼0.5 km (by 0.5–1 °C), too cold between ∼0.5 and ∼3 km (by 1–2 °C), too warm again around 3–5 km (by <0.5 °C) and too cold through the tropopause and lower stratosphere. The error magnitudes are substantially larger after the surface freezes permanently, especially in the lower half of the troposphere.
   (b) The surface skin temperature is also too high. During the melt it is unphysically stuck at ∼0.5 °C, when it should be zero, and errors increase substantially as the surface freezes and the temperatures drop.
   (c) Lower troposphere temperature errors have a distinct diurnal cycle, almost ±1 °C in the lowest layers after the freeze-up. The largest (smallest) error below 500 m (at 1–3 km) is from midnight to 0400 LT. Observed near-surface temperatures display a very weak diurnal cycle even after the melt, O(±0.1 °C).

2. A few variables show a gradual growth with forecast length of error and its spread. This includes the mean-sea-level pressure, from initially zero to 1.5 hPa at +72 hr, scalar wind-speed also close to zero initially, growing to 0.5 m·s⁻¹ at +72 hr. Turbulent momentum-flux errors are small initially and grow with the wind speed error. The wind-direction error, however, is constant with forecast time and <10°.

3. The surface energy budget has two larger systematic errors. There is an enhanced downward turbulent sensible heat flux and the incoming surface solar radiation is too low. This error combination suggests that the surface is on average being warmed by the atmosphere and not the opposite. The annoying unphysically warm surface during the melt season, that first caught our interest, must hence be related to the coupling to the ice but it is forced by the main problem; the too-warm boundary layer.

4. While cloud characteristics (cloud-water contents, phase partitioning, etc.) appear reasonable, IFS has too much cloud, in time and space. Cloud cover is persistent and lower troposphere non-frontal clouds are too deep. Consequently, liquid and ice water paths are systematically too large. The error in cloud occurrence has a distinct vertical structure; too much cloud below 3 km and too little between 3 and 5 km. The largest overrepresentation of clouds is around 1 km, aligned with the temperature error.

Of more technical import is an apparent but hitherto unexplained 6-hourly noise cycle in the error of most parametrized parameters, not appearing in errors in the model state variables.

We suggest that most of the errors are related to the model physics, resulting in an erroneous model climate to which the model drifts back after initialization. Many of the errors likely have their roots in the description of cloud formation in IFS, either directly or from cloud-related feedback from other model physics; we specifically point to the so-called mid-level convection as a parametrization that needs to be reviewed; Untangling these coupled relationships will require targeted experimental simulations outside the scope of this study.

ORCID

Michael Tjernström https://orcid.org/0000-0002-6908-7410
Gunilla Svensson https://orcid.org/0000-0001-9074-7623

Linus Magnusson https://orcid.org/0000-0003-4707-2231

Ian M. Brooks https://orcid.org/0000-0002-5051-1322

John Prytherch https://orcid.org/0000-0003-1209-289X

Jutta Vüllers https://orcid.org/0000-0002-6483-1159

Gillian Young https://orcid.org/0000-0002-8464-7332

REFERENCES

Achtert, P., O'Connor, E., Brooks, I.M., Sotiropoulou, G., Shupe, M.D., Persson, P.O.G., Pospichal, B., Brooks, B.J. and Tjernström, M. (2020) Properties of Arctic mixed phase clouds from ship-borne Cloudnet observations during ACSE 2014. *Atmospheric Chemistry and Physics*, 20, 14983–15002. https://doi.org/10.5194/acp-2020-56.

Bauer, P., Magnusson, L., Thépaut, J.-N. and Hamill, T.M. (2016) Aspects of ECMWF model performance in polar areas. *Quarterly Journal of the Royal Meteorological Society*, 142(695), 583–596. https://doi.org/10.1002/qj.2449.

Birch, C.E., Brooks, I.M., Tjernström, M., Milton, S.F., Earnshaw, P., Söderberg, S. and Persson, P.O.G. (2009) The performance of a global and mesoscale model over the central Arctic Ocean during late summer. *Journal of Geophysical Research*, 114(D13), D13104. https://doi.org/10.1029/2008JD010790.

Brooks, I.M., Tjernström, M., Persson, P.O.G., Shupe, M.D., Atkinson, R.A., Canut, G., Birch, C.E., Mauritsen, T., Sedlar, J. and Brooks, B.J. (2017) The turbulent structure of the Arctic summer boundary layer during ASCOS. *Journal of Geophysical Research: Atmospheres*, 122, 9685–9704. https://doi.org/10.1002/2017JD027234.

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

Foken, T. (2006) 50 years of the Monin–Obukhov similarity theory. *Boundary-Layer Meteorology*, 119, 431–447. https://doi.org/10.1007/s10546-006-9048-6.

Haiden, T., Janousek, M., Vitart, F., Ferranti, L. and Prates, F. (2019) *Evaluation of ECMWF forecasts, including the 2019 upgrade*. ECMWF Technical Memorandum 853. European Centre for Medium-Range Weather Forecasts: Shinfield Park, Reading. Available at: https://www.ecmwf.int/sites/default/files/elibrary/2019/19277-evaluation-ecmwf-forecasts-including-2019-upgrade.pdf.

Hartfield, G., Blunden, J. and Arndt, D.S. (2018) A look at 2017: takeaway points from the State of the Climate supplement. *Bulletin of the American Meteorological Society*, 99, 1527–1539. https://doi.org/10.1175/BAMS-D-18-0173.1.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Holm, E., Janiskova, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and Thépaut, J.-N. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049. https://doi.org/10.1002/qj.3803.

Holland, M.M. and Bitz, C.M. (2003) Polar amplification of climate change in coupled models. *Climate Dynamics*, 21, 221–232. https://doi.org/10.1007/s00382-003-0332-6.

Holland, M.M. and Stroeve, J. (2011) Changing seasonal sea ice predictor relationships in a changing Arctic climate. *Geophysical Research Letters*, 38, L18501. https://doi.org/10.1029/2011GL049303.

Illingworth, A.J., Hogan, R.J., O'Connor, E.J., Bouniol, D., Brooks, M.E., Delanoé, J., Donovan, D.P., Eastment, J.D., Gaussiat, N., Goddard, J.W.F., Haeffelin, M., Baltink, H.K., Krasnov, O.A., Pelon, J., Piriou, J.-M., Protat, A., Russchenberg, H.W.J., Seifert, A., Tompkins, A.M., van Zadelhoff, G.-J., Vinit, F., Willén, U., Wilson, D.R. and Wrench, C.L. (2007) Cloudnet. *Bulletin of the American Meteorological Society*, 88, 883–898. https://doi.org/10.1175/BAMS-88-6-883.

IPCC. (2019) Summary for policymakers. In: Pörtner, H.-O., Roberts, D.C., Masson-Delmotte, V., Zhai, P., Tignor, M., Poloczanska, E., Mintenbeck, K., Alegría, A., Nicolai, M., Okem, A., Petzold, J., Rama, B. and Weyer, N.M. (Eds.) *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*. Available at https://www.ipcc.ch/site/assets/uploads/sites/3/2019/11/03_SROCC_SPM_FINAL.pdf.

Jakobson, E., Vihma, T., Palo, T., Jakobson, L., Keernik, H. and Jaagus, J. (2012) Validation of atmospheric reanalyses over the central Arctic Ocean. *Geophysical Research Letters*, 39(10), L10802. https://doi.org/10.1029/2012GL051591.

Jung, T., Gordon, N.D., Bauer, P., Bromwich, D.H., Chevallier, M., Day, J.J., Dawson, J., Doblas-Reyes, F., Fairall, C., Goessling, H.F., Holland, M., Inoue, J., Iversen, T., Klebe, S., Lemke, P., Losch, M., Makshtas, A., Mills, B., Nurmi, P., Perovich, D., Reid, P., Renfrew, I.A., Smith, G., Svensson, G., Tolstykh, M. and Yang, Q. (2016) Advancing polar prediction capabilities on daily to seasonal time scales. *Bulletin of the American Meteorological Society*, 97, 1631–1647. https://doi.org/10.1175/BAMS-D-14-00246.

Jung, T. and 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.

Kwok, R. (2018) Arctic Sea ice thickness, volume, and multiyear ice coverage: losses and coupled variability (1958–2018). *Environmental Research Letters*, 13, 105005. https://doi.org/10.1088/1748-9326/aae3ec.

Lindvall, J. and Svensson, G. (2019) Wind turning in the atmospheric boundary layer over land. *Quarterly Journal of the Royal Meteorological Society*, 145(724), 3074–3308. https://doi.org/10.1002/qj.3605.

Meredith, M., Sommerkorn, M., Cassotta, S., Derksen, C., Ekaykin, A., Hollowed, A., Kofinas, G., Mackintosh, A., Melbourne-Thomas, J., Muelbert, M.M.C., Ottersen, G., Pritchard, H. and Schuur, E.A.G. (2019) Polar regions. In: Pörtner, H.-O., Roberts, D.C., Masson-Delmotte, V., Zhai, P., Tignor, M., Poloczanska, E., Mintenbeck, K., Alegría, A., Nicolai, M., Okem, A., Petzold, J., Rama, B. and Weyer, N.M. (Eds.) *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*. https://www.ipcc.ch/site/assets/uploads/sites/3/2019/11/07_SROCC_Ch03_FINAL.pdf.

Naakka, T., Nygård, T., Tjernström, M., Vihma, T., Pirazzini, R. and Brooks, I.M. (2019) The impact of radiosounding observations on numerical weather prediction analyses in the Arctic. *Geophysical Research Letters*, 46, 8527–8535. https://doi.org/10.1029/2019GL083332.

Onarheim, I.H., Eldevik, T., Smedsrud, L.H. and Stroeve, J.C. (2018) Seasonal and regional manifestation of Arctic Sea ice loss. *Journal of Climate*, 31, 4917–4932. https://doi.org/10.1175/JCLI-D-17-0427.1.

Parker, W.S. (2016) Reanalyses and observations: what's the difference? *Bulletin of the American Meteorological Society*, 97, 1565–1572. https://doi.org/10.1175/BAMS-D-14-00226.1.

Prytherch, J., Brooks, I.M., Crill, P., Thornton, B., Salisbury, D.J., Tjernström, M., Anderson, L., Geibel, M.C. and Humborg, C. (2017) Direct determination of the air–sea CO_2 gas transfer velocity in Arctic Sea ice regions. *Geophysical Research Letters*, 44, 3770–3778. https://doi.org/10.1002/2017GL073593.

Prytherch, J., Yelland, M.J., Brooks, I.M., Tupman, D.J., Pascal, R.W., Moat, B.I. and Norris, S.J. (2015) Motion-correlated flow distortion and wave-induced biases in air–sea flux measurements from ships. *Atmospheric Chemistry and Physics*, 15, 10619–10629. https://doi.org/10.5194/acp-15-10619-2015.

Renfrew, I.A., Elvidge, A.D. and Edwards, J.M. (2019) Atmospheric sensitivity to marginal-ice-zone drag: local and global responses. *Quarterly Journal of the Royal Meteorological Society*, 145(720), 1165–1179. https://doi.org/10.1002/qj.3486.

Ricker, R., Hendricks, S., Kaleschke, L., Tian-Kunze, X., King, J. and Haas, C. (2017) A weekly Arctic sea-ice thickness data record from merged CryoSat-2 and SMOS satellite data. *The Cryosphere*, 11, 1607–1623. https://doi.org/10.5194/tc-11-1607-2017.

Šálek, M. and Szabó-Takács, B. (2019) Comparison of SAFNWC/MSG satellite cloud type with Vaisala CL51 ceilometer-detected cloud base layer using the sky condition algorithm and Vaisala BL-view software. *Atmosphere*, 10, 316. https://doi.org/10.3390/atmos10060316.

Sedlar, J., Tjernström, M., Cassano, J., Fettweis, X., Hebestadt, I., Heinemann, G., Orr, A., Phillips, T., Rinke, A., Seefeldt, M. and Solomon, A. (2020) Confronting Arctic troposphere and surface energy budget representations in regional climate models with observations. *Journal of Geophysical Research: Atmospheres*, 125(6), e2019JD031783. https://doi.org/10.1029/2019JD031783.

Sedlar, J., Tjernström, M., Mauritsen, T., Shupe, M., Brooks, I., Persson, O., Birch, C., Leck, C., Sirevaag, A. and Nicolaus, M. (2011) A transitioning Arctic surface energy budget: the impacts of solar zenith angle, surface albedo and cloud radiative forcing. *Climate Dynamics*, 37, 1643–1660. https://doi.org/10.1007/s00382-010-0937-5.

Serreze, M.C. and Barry, R.G. (2011) Processes and impacts of Arctic amplification: a research synthesis. *Global and Planetary Change*, 77, 85–96. https://doi.org/10.1016/j.gloplacha.2011.03.004.

Serreze, M.C. and Francis, J.A. (2009) The Arctic amplification debate. *Climatic Change*, 76, 241–264. https://doi.org/10.1007/s10584-005-9017-y.

Shupe, M.D., Persson, P.O.G., Brooks, I.M., Tjernström, M., Sedlar, J., Mauritsen, T., Sjogren, S. and Leck, C. (2013) Cloud and boundary layer interactions over the Arctic sea-ice in late summer. *Atmospheric Chemistry and Physics*, 13, 9379–9400. https://doi.org/10.5194/acp-13-9379-2013.

Shupe, M.D., Uttal, T. and Matrosov, S.Y. (2005) Arctic cloud microphysics retrievals from surface-based remote sensors at SHEBA.

Journal of Applied Meteorology, 44, 1544–1562. https://doi.org/10.1175/JAM2297.1.

Smith, L.C. and Stephenson, S.R. (2013) New trans-Arctic shipping routes navigable by mid-century. *Proceedings of the National Academy of Sciences USA*, 110, E1191–E1195. https://doi.org/10.1073/pnas.1214212110.

Sotiropoulou, G., Sedlar, J., Forbes, R. and Tjernström, M. (2016a) Summer Arctic clouds in the ECMWF forecast model: an evaluation of cloud parametrization schemes. *Quarterly Journal of the Royal Meteorological Society*, 142(694), 387–400. https://doi.org/10.1002/qj.2658.

Sotiropoulou, G., Sedlar, J., Tjernström, M., Shupe, M.D., Brooks, I.M. and Persson, P.O.G. (2014) The thermodynamic structure of summer Arctic stratocumulus and the dynamic coupling to the surface. *Atmospheric Chemistry and Physics*, 14, 12573–12592. https://doi.org/10.5194/acp-14-12573-2014.

Sotiropoulou, G., Tjernström, M., Sedlar, J., Achtert, P., Brooks, B.J., Brooks, I.M., Persson, P.O.G., Prytherch, J., Salisbury, D.J., Shupe, M.D., Johnston, P.E. and Wolfe, D. (2016b) Atmospheric conditions during the Arctic Clouds in Summer Experiment (ACSE): contrasting open-water and sea-ice surfaces during melt and freeze-up seasons. *Journal of Climate*, 29, 8721–8744. https://doi.org/10.1175/JCLI-D-16-0211.1.

Spreen, G., Kaleschke, L. and Heygster, G. (2008) Sea ice remote sensing using AMSR-E 89-GHz channels. *Journal of Geophysical Research*, 113(C2), C02S03. https://doi.org/10.1029/2005JC003384.

Tjernström, M. (2005) The summer Arctic boundary layer during the Arctic Ocean Experiment 2001 (AOE-2001). *Boundary-Layer Meteorology*, 117, 5–36.

Tjernström, M. (2007) Is there a diurnal cycle in Arctic summer cloud-capped boundary layer? *Journal of the Atmospheric Sciences*, 64, 3974–3990.

Tjernström, M., Birch, C.E., Brooks, I.M., Shupe, M.D., Persson, P.O.G., Sedlar, J., Mauritsen, T., Leck, C., Paatero, J., Szczodrak, M. and Wheeler, C.R. (2012) Meteorological conditions in the central Arctic summer during the Arctic Summer Cloud Ocean Study (ASCOS). *Atmospheric Chemistry and Physics*, 12, 6863–6889. https://doi.org/10.5194/acp-12-6863-2012.

Tjernström, M., Leck, C., Birch, C.E., Bottenheim, J.W., Brooks, B.J., Brooks, I.M., Bäcklin, L., Chang, R.Y.-W., de Leeuw, G., Di Liberto, L., de la Rosa, S., Granath, E., Graus, M., Hansel, A., Heintzenberg, J., Held, A., Hind, A., Johnston, P., Knulst, J., Martin, M., Matrai, P.A., Mauritsen, T., Müller, M., Norris, S.J., Orellana, M.V., Orsini, D.A., Paatero, J., Persson, P.O.G., Gao, Q., Rauschenberg, C., Ristovski, Z., Sedlar, J., Shupe, M.D., Sierau, B., Sirevaag, A., Sjogren, S., Stetzer, O., Swietlicki, E., Szczodrak, M., Vaattovaara, P., Wahlberg, N., Westberg, M. and Wheeler, C.R. (2014) The Arctic Summer Cloud Ocean Study (ASCOS): overview and experimental design. *Atmospheric Chemistry and Physics*, 14, 2823–2869. https://doi.org/10.5194/acp-14-2823-2014.

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

Tjernström, M., Svensson, G. and Magnusson, L. (2019) Arctic weather forecasting – in the high Arctic. *ECMWF Newsletter*, 160, 29–33.

Tjernström, M., Žagar, M., Svensson, G., Cassano, J.J., Pfeifer, S., Rinke, A., Wyser, K., Dethloff, K., Jones, C., Semmler, T. and Shaw, M. (2005) Modelling the Arctic boundary layer: an evaluation of six ARCMIP regional-scale models with data from the SHEBA project. *Boundary-Layer Meteorology*, 117, 337–381. https://doi.org/10.1007/s10546-004-7954-z.

Vüllers, J., Achtert, P., Brooks, I.M., Tjernström, M., Prytherch, J. and Neely, R., III. (2021) Meteorological and cloud conditions during the Arctic Ocean 2018 expedition. *Atmospheric Chemistry and Physics*, 21, 289–314. https://doi.org/10.5194/acp-21-289-2021.

Wesslén, C., Tjernström, M., Bromwich, D.H., Bai, L.-S., de Boer, G. and Ekman, A. (2014) The Arctic summer atmosphere: an evaluation of reanalyses using ASCOS data. *Atmospheric Chemistry and Physics*, 14, 2605–2624. https://doi.org/10.5194/acp-14-2605-2014.

Westwater, R., Han, Y., Shupe, M.D. and Matrosov, S.Y. (2001) Analysis of integrated cloud liquid and precipitable water vapor retrievals from microwave radiometers during SHEBA. *Journal of Geophysical Research*, 106, 32019–32030.

Wyser, K., Jones, C.G., Du, P., Girard, E., Willén, U., Cassano, J., Christensen, J.H., Curry, J.A., Dethloff, K., Haugen, J.-E., Jacob, D., Køltzow, M., Laprise, R., Lynch, A., Pfeifer, S., Rinke, A., Serreze, M., Shaw, M.J., Tjernström, M. and Žagar, M. (2008) An evaluation of Arctic cloud and radiation processes during the SHEBA year: simulation results from eight Arctic regional climate models. *Climate Dynamics*, 30, 203–223. https://doi.org/10.1007/s00382-007-0286-1.

How to cite this article: Tjernström M, Svensson G, Magnusson L, *et al*. Central Arctic weather forecasting: Confronting the ECMWF IFS with observations from the Arctic Ocean 2018 expedition. *Q J R Meteorol Soc*. 2021;1–22. https://doi.org/10.1002/qj.3971