

An NWP Model Intercomparison of Surface Weather Parameters in the European Arctic during the Year of Polar Prediction Special Observing Period Northern Hemisphere 1

MORTEN KØLTZOW

Norwegian Meteorological Institute, Oslo, Norway

BARBARA CASATI

Environment and Climate Change Canada, Dorval, Quebec, Canada

ERIC BAZILE

Météo France, Toulouse, France

THOMAS HAIDEN

ECMWF, Reading, United Kingdom


TERESA VALKONEN

Norwegian Meteorological Institute, Oslo, Norway

(Manuscript received 11 January 2019, in final form 24 May 2019)

ABSTRACT

Increased human activity in the Arctic calls for accurate and reliable weather predictions. This study presents an intercomparison of operational and/or high-resolution models in an attempt to establish a baseline for present-day Arctic short-range forecast capabilities for near-surface weather (pressure, wind speed, temperature, precipitation, and total cloud cover) during winter. One global model [the high-resolution version of the ECMWF Integrated Forecasting System (IFS-HRES)], and three high-resolution, limited-area models [Applications of Research to Operations at Mesoscale (AROME)-Arctic, Canadian Arctic Prediction System (CAPS), and AROME with Météo-France setup (MF-AROME)] are evaluated. As part of the model intercomparison, several aspects of the impact of observation errors and representativeness on the verification are discussed. The results show how the forecasts differ in their spatial details and how forecast accuracy varies with region, parameter, lead time, weather, and forecast system, and they confirm many findings from mid- or lower latitudes. While some weaknesses are unique or more pronounced in some of the systems, several common model deficiencies are found, such as forecasting temperature during cloud-free, calm weather; a cold bias in windy conditions; the distinction between freezing and melting conditions; underestimation of solid precipitation; less skillful wind speed forecasts over land than over ocean; and difficulties with small-scale spatial variability. The added value of high-resolution limited area models is most pronounced for wind speed and temperature in regions with complex terrain and coastlines. However, forecast errors grow faster in the high-resolution models. This study also shows that observation errors and representativeness can account for a substantial part of the difference between forecast and observations in standard verification.

 Denotes content that is immediately available upon publication as open access.

Corresponding author: Morten Køltzow, famo@met.no



This article is licensed under a [Creative Commons Attribution 4.0 license](http://creativecommons.org/licenses/by/4.0/) (<http://creativecommons.org/licenses/by/4.0/>).

DOI: 10.1175/WAF-D-19-0003.1

© 2019 American Meteorological Society

Unauthenticated | Downloaded 12/16/20 11:55 AM UTC

1. Introduction

The Arctic is experiencing rapid changes in its harsh climate and environment, for example, the observed annual averaged near-surface temperatures at Svalbard are now increasing at between 1.04° and $1.76^{\circ}\text{C decade}^{-1}$ (Hanssen-Bauer et al. 2019). Anticipated increases in ship traffic, resource exploitation, tourism, and other activities (WMO 2017) call for accurate and reliable weather predictions for safe and efficient operations. Despite improved Arctic forecast skill in recent decades (Bauer et al. 2016; Jung and Leutbecher 2007), Jung et al. (2016) argues that existing numerical weather prediction (NWP) systems do not meet existing user requirements. Furthermore, forecast errors in the Arctic are larger than at lower latitudes (e.g., Nordeng et al. 2007; Bauer et al. 2016; Gascard et al. 2017). Nordeng et al. (2007) argue that the main reasons for this are the sparse conventional observational network and the small spatial scales of many (high impact) Arctic-specific weather phenomena. NWP systems are also often developed and tuned with a focus on mid and lower-latitude weather.

Arctic verification studies of global model systems often use model analyses as truth, given the relative sparseness of observations. However, this introduces uncertainty in the interpretation because a higher spread between analyses compared to lower latitudes is found since the analyses are less constrained by observations and are closer to their inherent model climatology (Jung and Matsueda 2016; Bauer et al. 2016). Bauer et al. (2016) found that verifying near-surface temperatures in the Arctic against observations gave substantially larger errors compared to verifying against model analyses. To establish the state of the art on Arctic forecast capabilities, more verification of near-surface parameters, including snow and sea ice characteristics, are needed (Jung et al. 2016).

The use of regional models can, compared with global models, improve forecast accuracy by the use of optimized physics for the targeted area and finer horizontal and vertical resolution (Jung et al. 2016). However, operational convection permitting resolution models have just recently started to appear for the Arctic domain. Müller et al. (2017) and Yang et al. (2018) describe added value from operational high-resolution HIRLAM–ALADIN Research on Mesoscale Operational NWP in Euromed (HARMONIE)–Applications of Research to Operations at Mesoscale (AROME) runs in the Arctic compared to coarser resolution systems. Furthermore, specific Arctic weather phenomena, often connected to high-impact weather, have been studied in both global and regional high-resolution models. Models have been compared with

field observations and used as a tool to better understand the investigated phenomena. For example, polar lows have received substantial attention (e.g., Kristjánsson et al. 2011), but remain a challenge in operational forecasting because of their rapid growth and mesoscale nature (e.g., Spengler et al. 2017). Arctic cyclones (e.g., Yamagami et al. 2018), and sudden stratospheric warming events (e.g., Jung and Leutbecher 2007; Karpechko 2018) have also been the subject of recent Arctic forecast skill evaluations. Other examples of high-impact weather that have been studied are severe precipitation events at Svalbard (Hansen et al. 2014; Serreze et al. 2015) and maritime icing on vessels (Sultana et al. 2018; Samuelsen 2018).

The difference between a forecast value (grid box average from an NWP system) and a point observation can be decomposed into model, observation, interpolation, and representativeness errors (Kanamitsu and DeHaan 2011). The latter three components are nonnegligible for verification studies, in particular in the Arctic environment characterized by spatiotemporal sparseness and uncertainty in the observations (Casati et al. 2017). The observation uncertainty has been neglected in verification practices for several decades. As forecast capabilities improve, however, a larger part of the forecast–observation difference is due to observational uncertainty and representativeness mismatch rather than to model errors alone, in particular for short-range forecasts.

The Year of Polar Prediction (YOPP), with extra availability of observations and model simulations (Jung et al. 2016), is a great opportunity to improve our understanding of forecast capabilities in the Arctic. In this study we compare three high-resolution regional NWP systems and one global NWP system during the YOPP Special Observing Period Northern Hemisphere 1 (SOP-NH1, 1 February–31 March 2018) with focus on surface weather parameters. In addition, issues related to observational uncertainty are discussed to improve our interpretation of the verification results.

The NWP systems are briefly described in section 2, together with the observations and weather during YOPP SOP-NH1. The models are compared in terms of objective verification scores in section 3, including a discussion on some aspects of observation errors and representativeness issues. In section 4, two cases of high-impact weather are discussed before we summarize main findings in section 5.

2. NWP systems, observations, and weather

a. NWP systems

The NWP systems included in the comparison are the high-resolution version of the global ECMWF

Integrated Forecasting System (IFS-HRES) with 9-km grid spacing (Buizza et al. 2017) and the three regional convection permitting NWP systems: AROME-Arctic with 2.5-km grid spacing (Müller et al. 2017; Bengtsson et al. 2017), the Canadian Arctic Prediction System (CAPS) with 3-km grid spacing (G. C. Smith et al. 2019, unpublished manuscript), and AROME with Météo-France setup (MF-AROME) with 2.5-km grid spacing (Seity et al. 2011). Apart from spatial resolution, the four forecast systems differ in their model formulations, initialization methods, and in lateral and surface forcing (details in Table 1). IFS-HRES and AROME-Arctic forecasts are taken from daily operational runs and include data assimilation. CAPS and MF-AROME have been set up as a dedicated effort during YOPP and are initialized from global models without direct assimilation of observations. Furthermore, AROME-Arctic and MF-AROME are both configurations of the same model system but use different parameterizations in the turbulence representation and for shallow convection, and in addition a sea ice model is used in AROME-Arctic. Despite their differences, they all provide short-range forecasts for a common domain covering northern Scandinavia, the Barents Sea, and Svalbard (Fig. 1) during YOPP SOP-NH1.

b. Observations

In this study, we use quality controlled observations from the Norwegian Meteorological Institute (MET Norway; eklima.met.no). The quality control system consists of both automatic and human quality control routines to flag or remove suspicious or erroneous observations (Kielland 2005). In this study we only use observations flagged as high-quality observations. Pressure and temperature observations are from instantaneous measurements, while 10-m wind speed is the mean wind over the last 10 min. Total cloud cover is visually observed, which has some implications for the verification process, for example, the observations represent a larger spatial area, are taken less frequently, and have different uncertainty characteristics than automatic cloud cover observations (Mittermaier 2012). Furthermore, most of the precipitation gauges have single-Alter shields (or are less shielded) implying an undercatch of solid precipitation (Rasmussen et al. 2012).

To stratify the verification we divide the observation sites into six regions (Fig. 1); Svalbard (14 stations), islands (3 stations), coast (40 stations), fjords (39 stations), inland (25 stations), and mountains (9 stations). The assignment of each station to a region is done subjectively by operational forecasters at MET Norway based on their knowledge about individual stations.

Over the open ocean, we utilize near-real-time data from the global Advanced Scatterometer (ASCAT) coastal wind product on a 12.5-km grid (Verhoef et al. 2012). The ASCAT wind products, provided by the EUMETSAT Ocean and Sea Ice Satellite Application Facility (OSI SAF), include a thorough quality control. We utilized only the data with the highest-quality flags. For the model comparison, the ASCAT data were reprojected on the intercomparison domain, and NWP model data were regridded onto a 12.5-km grid corresponding to the grid spacing of the ASCAT data.

c. Weather during YOPP SOP-NH1

February 2018 was dominated by high pressure systems over Scandinavia and low pressure activity in the Iceland–Greenland Sea, which led to a negative temperature anomaly over northern Scandinavia and warm anomalies over the ocean and at Svalbard (ECMWF Copernicus Climate Change Service, <https://climate.copernicus.eu/>). In March, the pressure patterns were less consistent, but on average a high pressure system was present north of Svalbard with a low pressure system northeast of Scandinavia organizing the advection of cold air southward over the Barents Sea. In March a positive temperature anomaly was only present in the northwestern part of the intercomparison domain. The sea ice concentration anomaly was negative for the entire domain and period.

The North Atlantic Oscillation (NAO) is the dominant mode of variability in the North Atlantic region from synoptic to interannual and decadal time scales (Woollings et al. 2015). It indicates that February was an unusual month. An NAO index of 1.58 is the fifth-highest value for all February months from 1950 to 2018 (NOAA Climate Prediction Center, <https://www.cpc.ncep.noaa.gov/>). An NAO index of -0.93 in March indicates a clear difference in weather during the two months. However, March (ranked as the fifteenth-lowest value of all March months) was not as extreme in terms of NAO as February.

3. Model intercomparison

In the following we first present a general overview of verification results before we focus on individual parameters. At the end of the section some aspects of observation, interpolation and representativeness errors are discussed.

a. General verification

Standard deviation of the error (SDE) and bias, averaged over all stations, for mean sea level pressure (MSLP), 2-m air temperature (T2), and 10-m wind speed (WS10) are presented in Fig. 2 together with

TABLE 1. Summary of model systems for IFS-HRES, AROME-Arctic, CAPS, and MF-AROME during YOPP SOP NHI.

	IFS-HRES	AROME-Arctic	CAPS	MF-AROME
	Operational model ECMWF	Operational model MET Norway	Operational (experimental) model ECCO	Dedicated APPLICATE/YOPP version of Météo-France AROME
Model version	Cy43r3	Cy40h1.1	GEM 4.9.2	Cy42_op2
Upper air assimilation/initialization	4D-Var cutoff: 3 h	3D-Var (conventional obs, scatterometer data, satellite microwave and infrared radiances) cutoff: ~2 h 15 min	Downscaled from GDPS (EnVar, Buehner et al. 2015)	Dynamical adaptation from the operational global model Arpege production. The cutoff time is 2 h 15 min for the 0000 UTC and 1 h 50 min for the 1200 UTC analysis.
Surface assimilation/initialization	Optimal interpolation (t2m, rh2m, snow depth and temperature, soil temperature and moisture)	Optimal interpolation (t2m, rh2m, snow depth)	Downscaled from GDPS (optimal interpolation, t2m, rh2m, snow depth, soil temperature and moisture)	Dynamical adaptation from operational global model ARPEGE
Model runs per day	0000, 1200 UTC + 240 h; 0600, 1800 UTC +96 h	0000, 0600, 1200, 1800 UTC + 66 h; 0300, 0900, 1500, 2100 UTC + 3 h	0000, 1200 UTC + 48 h	0000, 1200 UTC + 48 h
Resolution	10 km, 137 L	2.5 km, 65 L (~14 levels below 500 m)	3 km, 62 L	2.5 km, 65 L (~14 levels below 500 m)
Sea ice and sea surface temperature	Sea ice and SST prescribed from OSTIA; prognostic sea ice temperature	Simple sea ice scheme (Batra et al. 2018); prognostic temperature and SST from IFS-HRES	Persisted surface conditions: 3D-Var ice analysis (Buehner et al. 2016) and CMC ice analysis (Brasnett 2008)	Sea ice and SST prescribed from OSTIA (constant during integration)
Surface model and snow treatment	H-TESSEL with 1-layer snow scheme; prognostic water equivalent, snow density, and surface albedo (Balsamo et al. 2009; Dutra et al. 2010)	SURFEX with 1-layer snow scheme; prognostic water equivalent, snow density, and surface albedo (Douville et al. 1995)	ISBA with 1-layer snow scheme; prognostic water equivalent, snow density, and surface albedo (Noilhan and Planton 1989; Béclair et al. 2003)	SURFEX with 1-layer snow scheme; prognostic water equivalent, snow density, and surface albedo (Douville et al. 1995). Soil scheme ISBA (Noilhan and Planton 1989).
Boundary conditions	None	LBC; IFS-HRES from 6-h old IFS-HRES run (hourly coupling)	LBC from GDPS 0000, 1200 UTC forecast runs; upper-level boundary nesting (McTaggart-Cowan et al. 2011)	LBC; ARPEGE from same cycle (hourly coupling)
References	Buizza et al. (2017)	Müller et al. (2017); Bengtsson et al. (2017)	G. C. Smith et al. (2019, unpublished manuscript) http://dd.alpha.meteo.gc.ca/yopp/model_caps/doc/README_CAPS.txt	Seity et al. (2011)
Model data	ECMWF's MARS archive	http://thredds.met.no	http://dd.alpha.meteo.gc.ca/yopp/model_caps/	https://yopp.met.no/ (YOPP data portal)

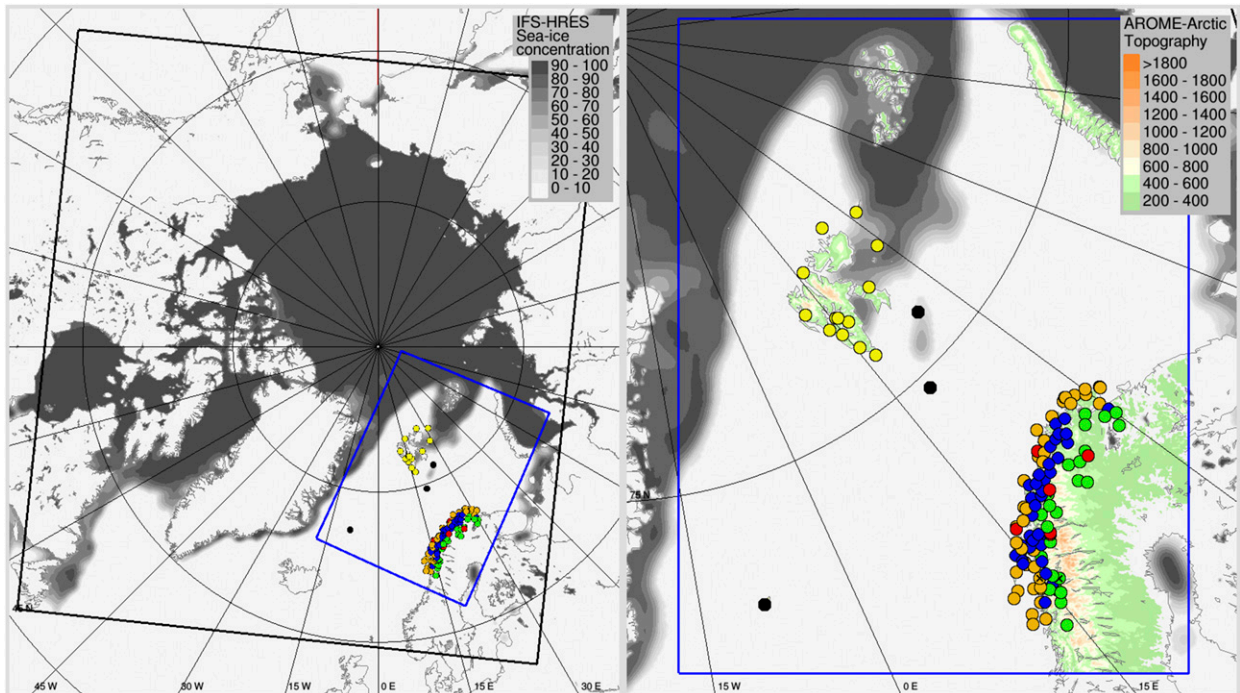


FIG. 1. Model integration domains: CAPS is employed inside the black frame, AROME-Arctic and MF-AROME are inside the blue frame, and IFS-HRES has global coverage. The model intercomparison area is inside the blue domain. Norwegian SYNOP observation used for verification are plotted as black (3 island stations), yellow (14 Svalbard stations), orange (40 coast stations), blue (39 fjord stations), green (25 inland stations), and red (9 mountain stations) circles. Not all stations observe all parameters. Shown in gray colors is the sea ice concentration from IFS-HRES 0000 UTC 1 Mar and in green/brown colors the model topography from AROME-Arctic.

information about statistical significance. It is important to note that errors in Fig. 2 are not weighted and therefore do not represent the model domain average, but the average errors over the irregularly distributed observational network shown in Fig. 1. The Initial MSLP SDEs are small for all forecast systems, but increase rapidly with lead time. For shorter lead times than +10 h AROME-Arctic has significant smaller SDEs than IFS-HRES, while after +30 h the opposite is true. CAPS has significantly larger errors than the other models after +12 h, which is however not found in the driving Canadian Global Deterministic Prediction System, and which is under investigation. While AROME-Arctic and IFS-HRES show a negligible bias, CAPS develops a positive bias after a few hours. Only forecasts initialized at 0000 UTC are included in the statistics, and the results indicate a small common diurnal cycle in SDE with a maximum error in the morning (+6 and +30h).

The T2 forecasts show a large SDE already in the analysis (3°–4°C), and the increase with lead time is more moderate than for MSLP. Furthermore, a diurnal cycle in SDE is present with higher accuracy during daytime, in the presence of solar radiation and higher temperatures, while larger errors are found during

nighttime (similar to MSLP). While AROME-Arctic and MF-AROME only have minor biases, CAPS and IFS-HRES show a diurnal cycle with a cold bias during daytime. Most of the differences seen between model performances are significant.

AROME-Arctic and MF-AROME show slightly lower SDE for WS10 than CAPS and IFS-HRES (only significant for the shortest lead times). The largest difference between the models is found in the biases, which for most lead times are statistically significant. AROME-Arctic and CAPS have negligible biases, while IFS-HRES and MF-AROME on average underestimate WS10 by $\sim 1 \text{ m s}^{-1}$. Only a weak diurnal cycle is seen in WS10 biases (maximum underestimation during daytime). A short spinup time of CAPS WS10 from the initial conditions is seen.

For all three parameters, errors grow more slowly in IFS-HRES than in the three high-resolution models (i.e., the added value of high-resolution models are dependent on lead time). In the case of MSLP, which is a surface field but represents a vertically integrated quantity, this reflects the leading role of the IFS in terms of synoptic-scale dynamics (Haider et al. 2018a). In the case of T2 and WS10, the apparently slower error growth actually results from larger errors already at

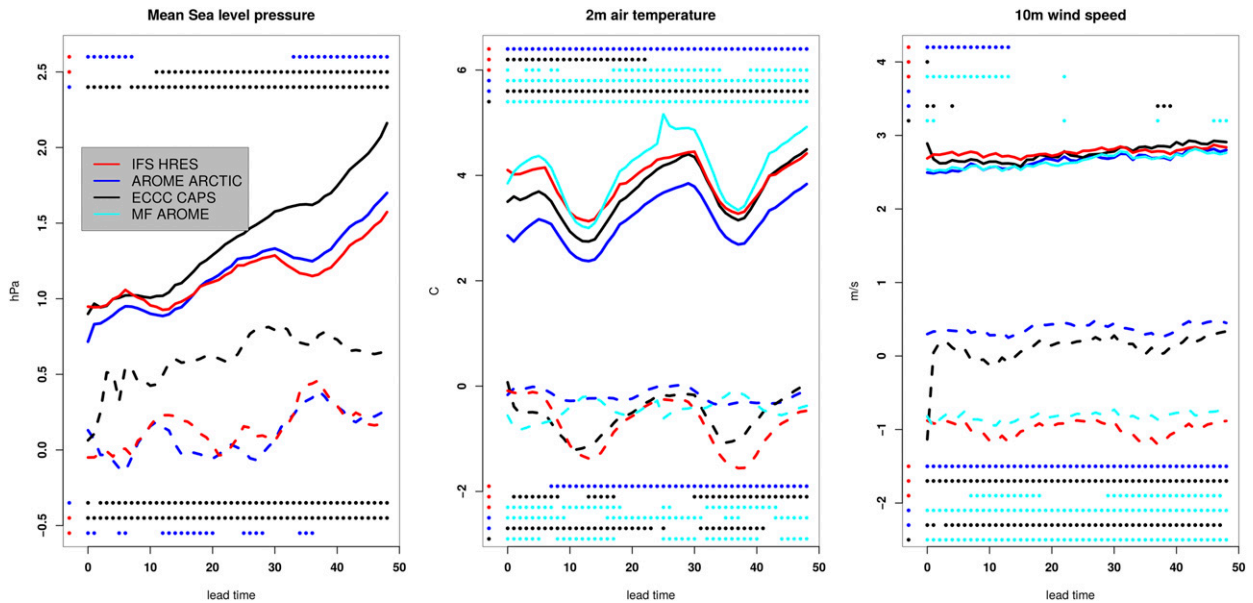


FIG. 2. Standard deviation of error (solid lines) and bias (dashed lines) as function of lead time. Models are IFS-HRES (red), AROME-Arctic (blue), CAPS (black), and MF-AROME (cyan; MSLP not available from MF-AROME). Verified parameters are MSLP, T2, and WS10. Verification period is YOPP SOP-NH1, and all forecasts are initialized at 0000 UTC. The 95% confidence interval is calculated by bootstrapping. The dots on top (bottom) indicate significance for SDE (bias). Colors indicate which models are compared (model one to the left, lead time < 0), and model two is shown for all individual lead times if confidence intervals are not overlapping.

initialization time in the IFS compared to the higher-resolution models, as can be seen in Fig. 2.

Statistics averaged over all stations, as presented in Fig. 2, may hide important information. Figure 3 shows verification of MSLP, T2m, WS10, daily precipitation (precip24), and total cloud cover (TCC; no observations available in mountain areas) forecasts for different regions (see section 2b for details). To give information about statistical significant differences between regions and forecast systems, 95% confidence intervals are calculated by bootstrapping (not shown). For MSLP, T2, WS10, and TCC these confidence intervals are 0.1 hPa, 0.14°C, 0.13 ms⁻¹, and 3.3% respectively. Differences seen for these parameters are therefore mostly significant. For daily precipitation, the uncertainty is much higher due to fewer observations, and the differences are not all significant.

The first feature to notice is the huge spread in forecast accuracy across regions, parameters, and models. Furthermore, no model is superior for all parameters and regions. IFS-HRES verifies consistently better for MSLP than AROME-Arctic and CAPS across regions. The inaccurate treatment of lateral boundary forcing in regional models is discussed, for example, by Warner et al. (1997) and Davies (2014) and may explain part of this behavior. Other possible explanations are better assimilation of large-scale weather in global models, tuning of global models with focus on synoptic cyclones

(e.g., Sandu et al. 2013), more small-scale noise in higher-resolution systems, and for AROME-Arctic the use of 6-h older LBC from IFS-HRES. Furthermore, all models have a pronounced positive MSLP bias in mountain regions (and inland and in Svalbard) most likely to be attributable to the uncertainty in reduction of observations and/or forecasted pressure to MSLP (Pauley 1998).

The largest T2 errors are found inland, in mountains (IFS-HRES and MF-AROME), and at Svalbard (CAPS). The bias varies from -4°C (CAPS at Svalbard) to +1°C (MF-AROME at islands), while SDE varies from ~1°C (IFS-HRES at islands) to ~6°C (MF-AROME inland). The CAPS bias at Svalbard (stations at the coast and in fjords) is related to an unrealistic sea ice cover around Svalbard (not shown). In general, small forecast errors are found where sea surface temperatures, which most of the time are reasonably well described in the models, have a substantial influence (i.e., coasts, fjords, and islands). Also, WS10 biases vary substantially across regions and models from -3 ms⁻¹ in mountain regions (IFS-HRES and MF-AROME) to +1.5 ms⁻¹ at islands (AROME-Arctic and CAPS). In general, SDE can be expected to scale with wind speed itself and is therefore higher in windier regions. However, forecast accuracy of WS10 is not fully evaluated by SDE and bias, and other aspects will be discussed below.

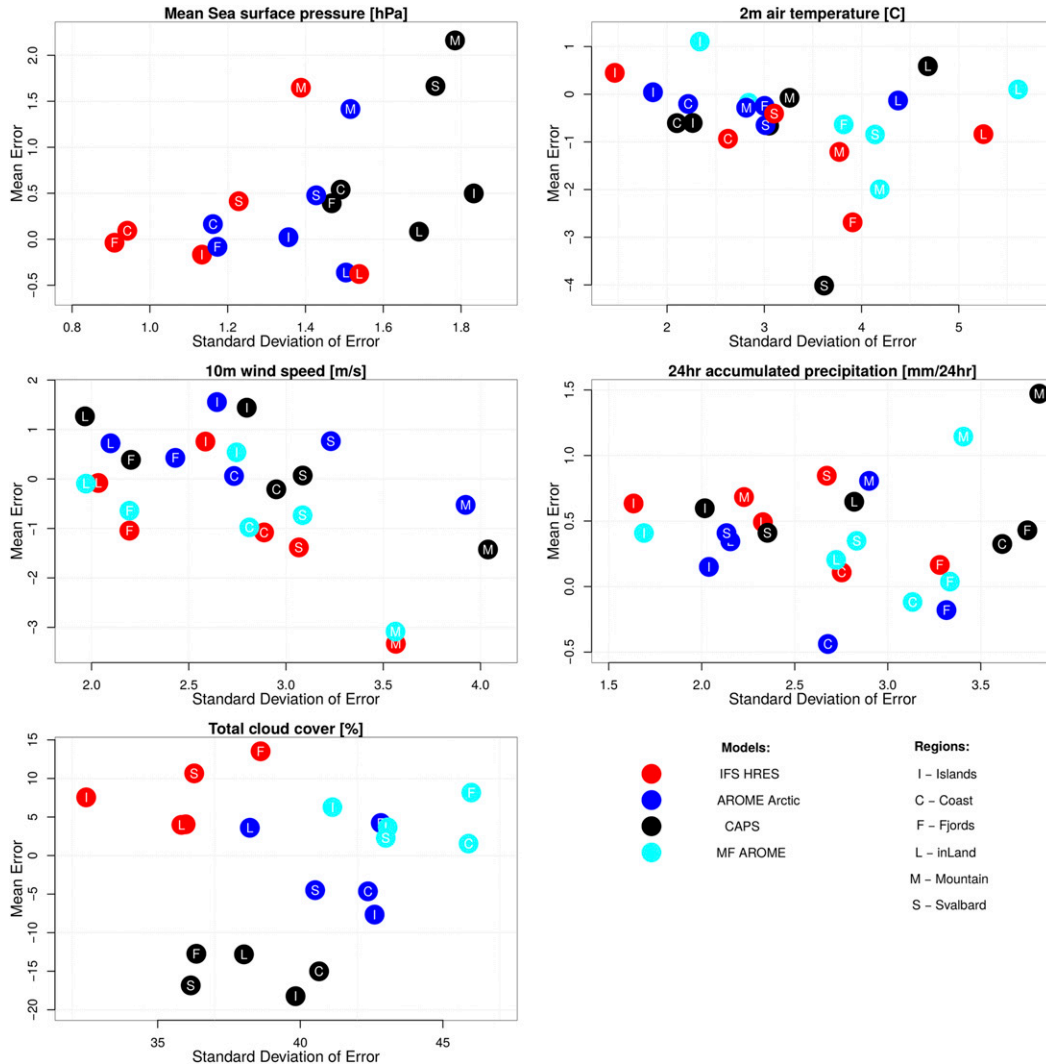


FIG. 3. Mean error (bias) and SDE for MSLP, T2m, WS10, daily precipitation (precipitation), and TCC during YOPP SOP-NH1. Each circle represents one region and one model. Models are given by color: IFS-HRES (red), AROME-Arctic (blue), CAPS (black), and MF-AROME (cyan). Regions are indicated by letter (see Fig. 1): islands (I), coast (C), fjords (F), inland (L), mountain (M), and Svalbard (S). Lead times included are +25, +26, . . . , +48 h for all parameters, with the exception of accumulated precipitation where lead times +42 h minus +18 h are used. Forecasts used are initialized at 0000 and 1200 UTC.

While precip24 scores vary across regions and models, some common significant features are low (high) SDE at islands (mountains and fjords) and a positive bias in mountain regions. In addition, AROME-Arctic and MF-AROME forecast less precipitation than CAPS and IFS-HRES (significant at the coast, fjords and inland). Undercatch of solid precipitation in observations (Rasmussen et al. 2012) is a severe problem for precipitation verification at high latitudes and/or altitudes. This is not taken into consideration in Fig. 3 (but discussed below) hence we suspect that the positive bias in the mountains is actually smaller, that the small positive bias at the coast and in the fjords for IFS-HRES and

CAPS most likely will change to a negative bias, and that the underestimations of AROME-Arctic and MF-AROME are actually even more pronounced.

For TCC, forecast characteristics are more dependent on the forecast model than on the region. IFS-HRES has a smaller SDE than the other forecast systems, which at least partly can be attributed to manual observations representing a larger area, and a more binary cloud cover field in the high-resolution models. IFS-HRES has a positive bias and AROME-Arctic and MF-AROME have smaller biases, while CAPS has a negative bias partly related to a long spinup of cloud properties which currently is under investigation.

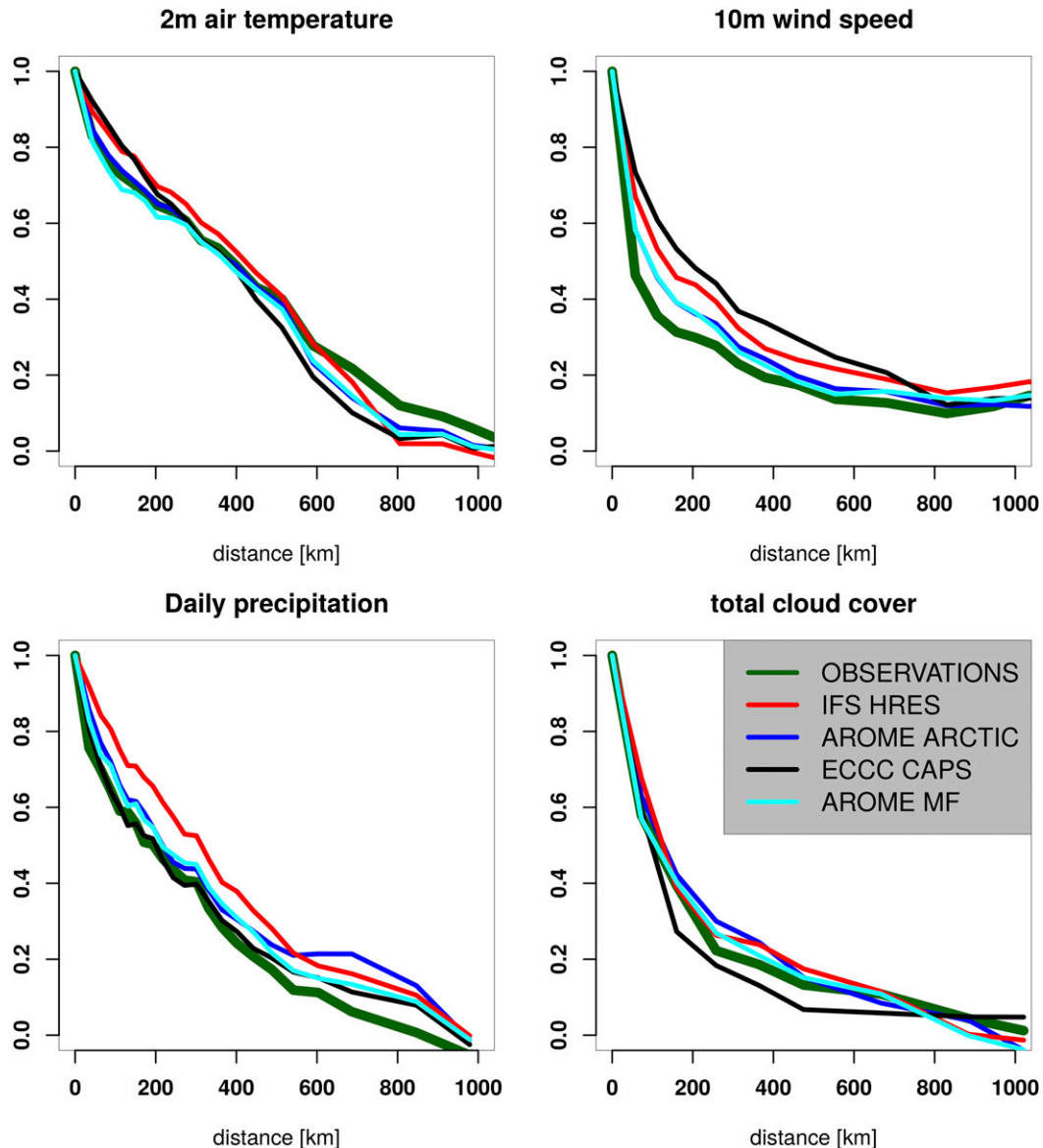


FIG. 4. Variograms showing spatial correlation between sites in observations and forecasts. Correlation as a function of distance between SYNOP sites is calculated, and the average over stations with similar distances are plotted. Observations are in green, IFS-HRES in red, AROME-Arctic in blue, CAPS in black, and MF-AROME in cyan. Parameters are 2-m air temperature, 10-m wind speed, daily precipitation, and total cloud cover.

Ideally, (gridded) high-resolution observation datasets are needed to evaluate spatial patterns in the forecasts. However, in this study we use point observations for verification. We therefore calculate the correlation between all observation sites for T2, WS10, TCC, and precip24. Correlations are then averaged in bins by the distance between the stations and plotted as variograms in Fig. 4 (Marzban et al. 2009). A rapid decorrelation with distance indicates stronger dominance of small-scale features. The observations of WS10 show a steep drop of correlation (approximately 0.35 after 100 km),

followed by TCC (approximately 0.55 after 100 km), precip24 (approximately 0.6 after 100 km), and T2 (approximately 0.7 after 100 km). It should be noted that WS10, T2, and TCC are hourly data while precipitation are daily totals due to the limited availability of hourly precipitation data. A shorter accumulation period for precipitation would most likely reduce the spatial correlation. In general, IFS-HRES has a higher spatial correlation than the other models, which is expected due to the coarser horizontal resolution. Furthermore, none of the models is able to reproduce the very steep

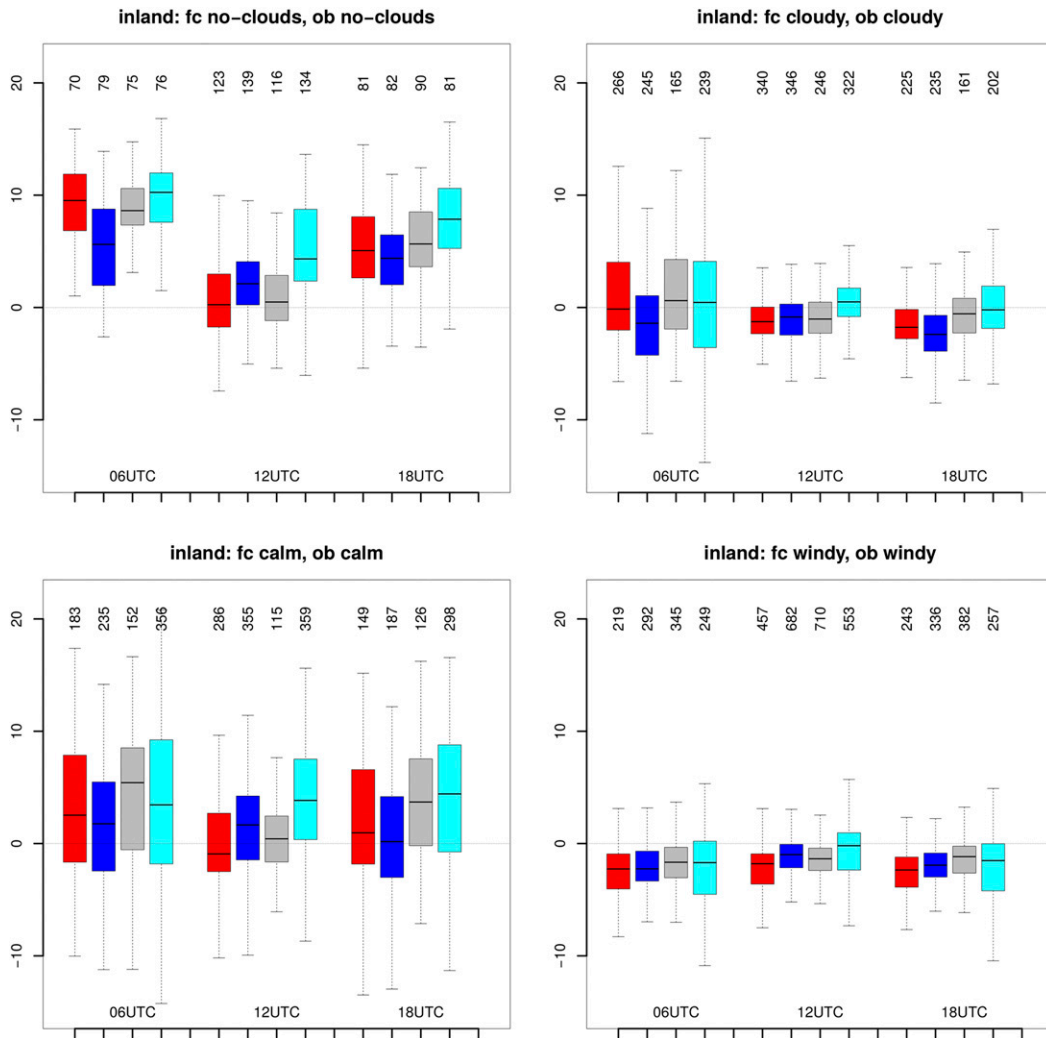


FIG. 5. Conditional verification of T2 for inland stations. Box-and-whiskers plot of T2 errors (forecasted minus observed) conditioned by (top) TCC and (bottom) wind. Cloud-free is defined as TCC less than 30% and cloudy as TCC larger than 70%. Calm conditions are defined as WS10 less than 1.5 m s^{-1} and windy conditions as WS10 larger than 3 m s^{-1} . Each box is divided into models (IFS-HRES in red, AROME-Arctic in blue, CAPS in black, and MF-AROME in cyan) and time of day. Number of cases is plotted at top, and outliers are omitted to increase readability in plots.

observed spatial decorrelation of WS10. For T2 and WS10 AROME-Arctic and MF-AROME are closest to the observed decorrelation, while CAPS matches best the observed curve for precip24 up to about 400 km. The lower spatial resolution of IFS-HRES shows up most clearly for precipitation. That the models find the small scales difficult to simulate is not surprisingly all the time the effective resolution of the models are even larger than the model grid spacing (Skamarock 2004). For distances beyond $\sim 650 \text{ km}$ the correlations are mainly between the Norwegian mainland and Svalbard where the forecasts underestimate (overestimate) the correlation of observed temperature (precipitation).

b. Temperature

All forecast systems struggle with T2 forecasts inland (Fig. 3). To better understand this problem we verify T2 stratified by TCC and WS10 (Fig. 5) and separate results into 0600, 1200, and 1800 UTC (when reliable cloud observations are available). T2 forecast errors increase in cloud-free conditions, while an increase in TCC reduces forecast errors. During calm conditions, a large spread in errors is seen as well as a positive bias, while errors are reduced in windy conditions, but with a small negative bias for all models. The fact that this negative bias is present throughout the day points to turbulent

TABLE 2. T2 errors inland for MF-AROME, for MF-AROME initialized by AROME-Arctic surface conditions, and AROME-Arctic. Errors are averaged over lead times (+1, +2, +3, . . . , +48 h) for 4 runs initialized at 0000 UTC 19 Feb and 12, 15, and 22 Mar 2018.

	MF-AROME	MF-AROME initialized by AROME-Arctic	AROME- Arctic
Mean absolute error	4.0°C	3.3°C	3.1°C
SDE	4.8°C	4.0°C	3.8°C
Mean error	1.7°C	1.6°C	1.3°C

mixing rather than cloud effects, however only IFS-HRES and MF-AROME show an underestimation of wind speed in cases where both observations and forecasts are $>3 \text{ m s}^{-1}$. Both stratified by TCC and WS10 the forecast errors decrease in the presence of solar radiation and higher temperatures (i.e., smaller errors at 1200 UTC). The conditional verification has some uncertainties for a number of reasons: 1) the need to set specific thresholds for cloud-free, cloudy, calm, and windy situations; 2) there is no one-to-one relation between TCC and cloud radiative effect; 3) dependence on the weather development prior to the verification time; and 4) limited sample size in terms of station number and total number of pairs of observations. Nevertheless, the increase in T2 forecast error during calm, cloud-free conditions without the presence of solar radiation points toward issues in the representation of the stable boundary layer as a common problem for all forecast systems. Haiden et al. (2018b) have recently investigated this problem for the IFS. They found that in areas with persistent snow cover the nighttime drop of T2 in the model is underestimated due to the use of a single-layer snowpack representation. It distributes the surface cooling over the entire depth of the snow, thereby underestimating the speed and magnitude of the near-surface drop in snow temperature, which adversely affects T2 evolution. Furthermore, nighttime wind speeds near the surface tend to be too high in low-wind conditions, which contribute to a positive bias, as well as a higher RMSE, in T2.

When the near-surface energy budget is determined by local processes, the representation of surface conditions becomes critical. We choose two days with cold temperatures and two days with more mixed conditions for reruns of MF-AROME starting from AROME-Arctic initial surface conditions. In the original runs MF-AROME initial surface conditions are interpolated from the coarser-resolution global ARPEGE model, while AROME-Arctic performs its own surface analysis. The results (Table 2) show that the initial differences

TABLE 3. Mean difference (m) between model height and observation site height for different regions. AROME is both MF-AROME and AROME-Arctic, which use the same orography.

	Islands	Coast	Fjord	Inland	Mountain	Svalbard
IFS-HRES	31	69	201	265	-254	100
AROME	10	38	70	109	-144	36
CAPS	10	48	125	174	-234	69

in the analysis explain almost the entire difference found in T2 errors (approximately a difference in SDE of 1 K in Fig. 3) between AROME-Arctic and MF-AROME.

Table 3 shows that the height differences are substantial between model and actual station elevation for some regions and models. This contributes to the T2 difference between models and observations. However, no height correction between model and station height is applied in the verification process since this potentially can introduce errors and noise during stable conditions. Furthermore, the implementation of well-behaving height corrections during stable conditions is beyond the scope of this study, but will potentially reduce the errors (e.g., Sheridan et al. 2010).

The ability to forecast thawing conditions (T2 above 0°C) is assessed using traditional categorical scores evaluated from the contingency table: the equitable threat score (ETS) and frequency bias index (FBI). The ETS is an accuracy measure evaluated from the threat score = hits/(hits + false alarms + misses), which then is modified for hits obtained by a random forecast. The FBI assesses the bias in the forecasted frequency of an event [see Wilks (2011, chapter 8) or Jolliffe and Stephenson (2012) for more details]. Forecast skill varies across regions (Fig. 6), from highest skill at islands, decreasing via coast to fjords to inland stations, but with slightly higher skill in the mountains (temperature more decoupled from surface) and at Svalbard (coast and fjord stations). The low skill inland is consistent with the larger T2 errors there, due to the generally higher T2 variability away from the coasts, and lower representativeness due to small-scale terrain features. In general, AROME-Arctic shows a similar or slightly better performance than the other models for all regions, which at least partly can be explained by high-resolution surface analysis and better representation of the topography.

Not all of the model differences show up in the objective verification since observations are not available for all areas. To supplement the evaluation of T2 we therefore show the average of hourly forecasts for day 2 during YOPP SOP-NH1 (Fig. 7). All models behave very similarly over the open ocean. However, over areas covered by sea ice (upper-left part of domain), CAPS and, in particular, MF-AROME have lower temperatures than IFS-HRES and AROME-Arctic. This suggests

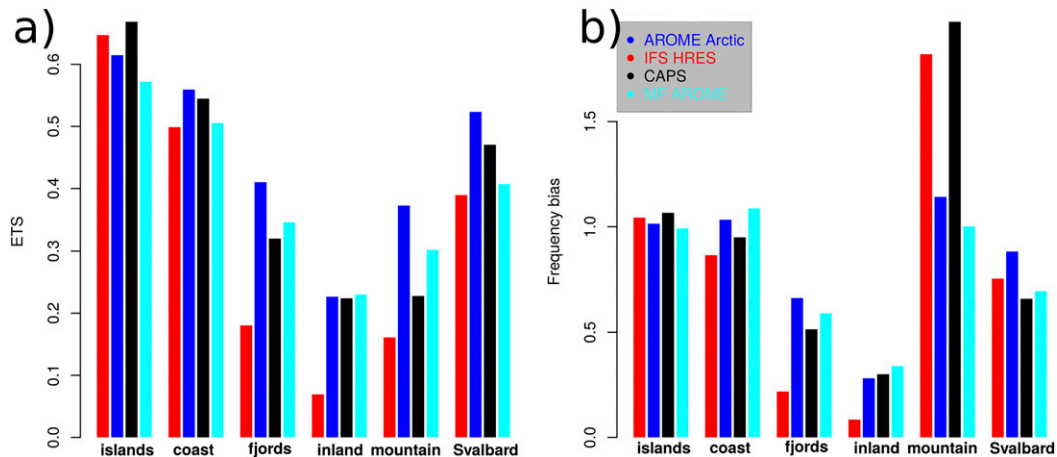


FIG. 6. (a) ETS and (b) FBI for $T_2 > 0^\circ\text{C}$ for models and regions. All statistics based on all lead times (hourly) between +25 and +48 h.

that these T_2 differences are due to differences in the representation of sea ice (see section 2a). Further inland at Svalbard and in the mountainous areas at the border between Norway and Sweden, MF-AROME is clearly colder than the other models, for example, a negative bias in the mountains is seen in Fig. 3. A similar behavior is found in the Alps for MF-AROME due to an underestimation of cloud cover (Vionnet et al. 2016). In general, the high-resolution models forecast the lowest minimum temperatures and, as expected, have more small-scale details than IFS-HRES (see also Fig. 4).

c. Wind speed

In addition to overall error metrics such as SDE, knowledge about forecast skill as a function of wind speed is of practical interest in the prediction of high-impact weather. This aspect is evaluated in Fig. 8 by the ETS and FBI obtained from a contingency table for different wind speed thresholds. The relative differences in skill between the forecasts are more pronounced for these metrics than in the SDE (Figs. 2 and 3). The frequencies of occurrence of the highest wind speeds are underestimated by CAPS, MF-AROME, and IFS-HRES, while AROME-Arctic is closer to the observed frequency. The skill (ETS) reflects the forecast climatologies, with AROME-Arctic scoring better than the other models, followed by CAPS, MF-AROME, and IFS-HRES. Large intermodel differences over land can be attributed to different representations of local processes, for example, AROME-Arctic applies a smaller surface roughness than MF-AROME. The benefit of higher spatial resolution for the prediction of high-wind events is shown by the low ETS values of IFS-HRES.

Figure 9 shows again ETS and FBI for all lead times from +25 to +48 h, but this time against scatterometer-estimated

wind speed in the Barents Sea (details in section 2). Forecasts are more similar and perform better than over land. However, when wind speeds exceed $12\text{--}13\text{ m s}^{-1}$ the models start to diverge and IFS-HRES (MF-AROME and AROME-Arctic) underestimates (overestimates) the observed frequency. For wind speeds up to $12\text{--}13\text{ m s}^{-1}$, AROME-Arctic and IFS-HRES have higher skill than MF-AROME and CAPS. Above $12\text{--}13\text{ m s}^{-1}$, the relative skill of IFS-HRES compared with AROME-Arctic is reduced at the same time as IFS-HRES starts to underestimate the observed frequencies. Since all forecast climatologies are quite similar over the ocean, we speculate that the higher skill of AROME-Arctic and IFS-HRES ($<12\text{ m s}^{-1}$) originates from more accurate initial conditions. In a case study in section 5a, this is further investigated by using initial conditions from AROME-Arctic in a MF-AROME run.

WS10 forecasts are more skillful over ocean than over land (Figs. 8 and 9) in spite of the added predictability which may be expected from topographic and coast line forcing. However, the representativeness of observations is an issue in the verification process, and especially in complex terrain. Since the scatterometer estimated wind speed represents a coarser resolution (grid size 12.5 km) over a relatively homogeneous ocean we argue that differences in observation representativeness (discussed further in section 3f) explain a large part of the difference in ETS between land and ocean.

To get a more complete overview of forecast differences in wind speed, forecast averages during YOPP SOP-NH1 are shown in Fig. 10. Wind speed forecasts over the ocean show clear similarities, but slightly less (more) wind in IFS-HRES (AROME-Arctic and MF-AROME). Also over sea ice areas the forecast systems are very similar, but MF-AROME has slightly lower

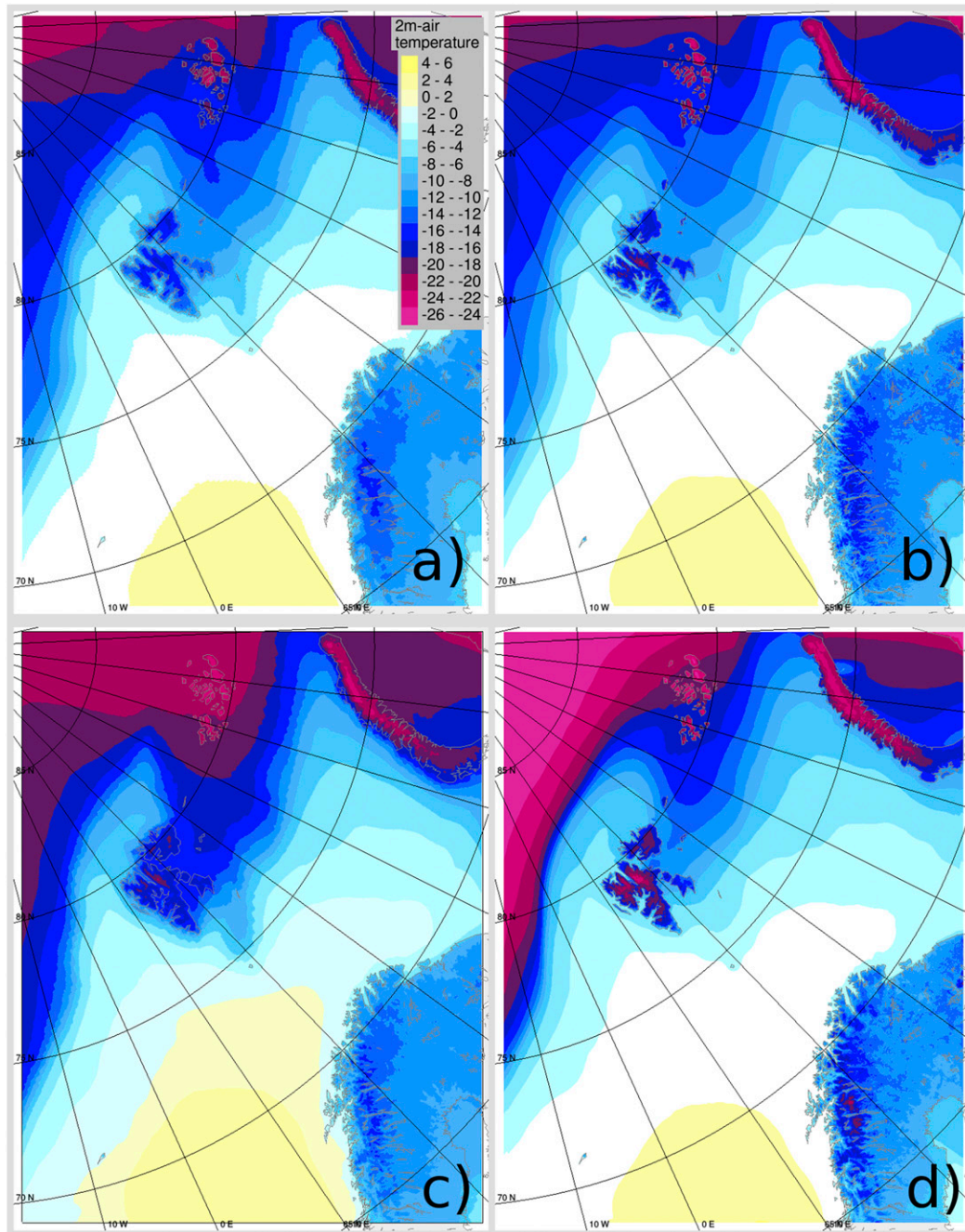


FIG. 7. Forecast average of 2-m air temperature for (a) IFS-HRES, (b) AROME-Arctic, (c) CAPS, and (d) MF-AROME. The averages are taken over YOPP SOP-NH1 for all lead times (hourly) from +25 to +48 h.

wind speed than the three other forecast systems. However, when comparing land areas we find large differences. In general, AROME-Arctic, followed by CAPS, forecast more windy conditions that are most pronounced over Svalbard and in the mountain regions, which agrees with the objective verification. As for temperature, IFS-HRES shows more smooth patterns than

the high-resolution models. A closer inspection of CAPS is also in agreement with smoother fields as indicated in Fig. 4.

d. Precipitation

To assess the forecast capabilities for precip24 further, we use ETS and FBI (Fig. 11). The MF-AROME forecasts have a similar frequency of occurrence as the

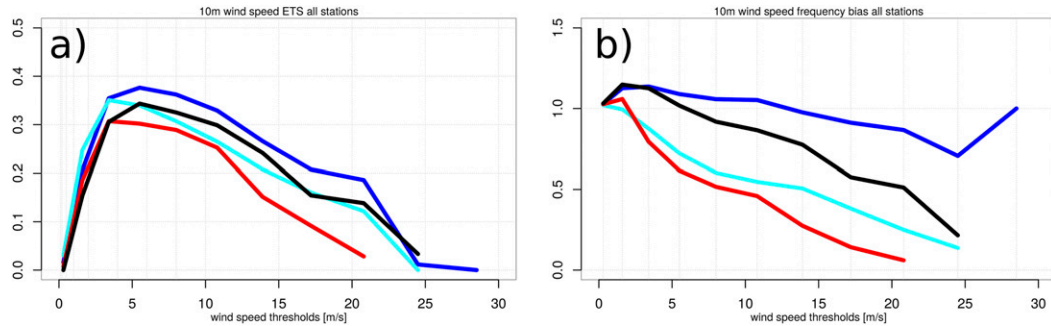


FIG. 8. ETS and FBI for wind speed over all stations used in the model-intercomparison. Models are IFS-HRES (red), AROME-Arctic (blue), CAPS (black), and MF-AROME (cyan). All hourly lead times from +25 to +48 h are used.

observations (FBI ~ 1) except for the highest precipitation amounts, while CAPS forecasts precipitation too frequently. AROME-Arctic overestimates the number of precipitation events but underestimates the frequency of events between 5 and 25 mm day⁻¹. The underestimated precipitation frequency originates mainly from coast and fjord regions (not shown). IFS-HRES produces too frequent small amounts of precipitation, which is a known problem, and underestimates the frequency of heavy precipitation. This is in part related to the coarser resolution of IFS-HRES, which means that the parameterization schemes represent the precipitation averaged over a wider area, which tends to generate a small precipitation trace and decrease intense precipitation values. The forecast skill measured by ETS reflects to some extent also the forecast climatologies. MF-AROME and IFS-HRES score better than AROME-Arctic and CAPS. In general, forecast skill decreases for high-precipitation events.

Observations of solid precipitation are associated with a high uncertainty due to wind-induced undercatch (Rasmussen et al. 2012). The undercatch varies with the type of precipitation gauge, windshield configurations, and the weather itself. In this study, most of the precipitation gauges are Geonor rain gauges with

single-Alter shields, and for 21 of them precipitation, temperature, and wind speed are measured hourly and the undercatch of solid precipitation can be estimated. We use Eq. (4) in Kochendorfer et al. (2017), Eq. (13) in Wolff et al. (2015), and Eq. (4) in Smith (2007) to adjust the observed precipitation. Figure 12 shows the accumulated precipitation from YOPP SOP-NH1, averaged over these 21 sites, from the four model systems, from the raw measurements and from the adjusted measurements. The precipitation is divided into rain, mixed precipitation, and solid precipitation by temperature thresholds. CAPS and IFS-HRES have more precipitation than the AROME models (as in Fig. 3), but all models slightly overestimate the raw measurements. Despite spread between the adjusted precipitation estimates, all models clearly underestimate the adjusted mixed phase and solid precipitation. The possible underestimation is so large that it raises a question about the adjustment of the observations. However, at Šihččajávri (68.7550°N, 23.5369°E) two estimates of accumulated precipitation during YOPP SOP-NH1 are available. One is based on a precipitation gauge and one is derived from changes in observed surface snow water equivalent, provided by the Norwegian

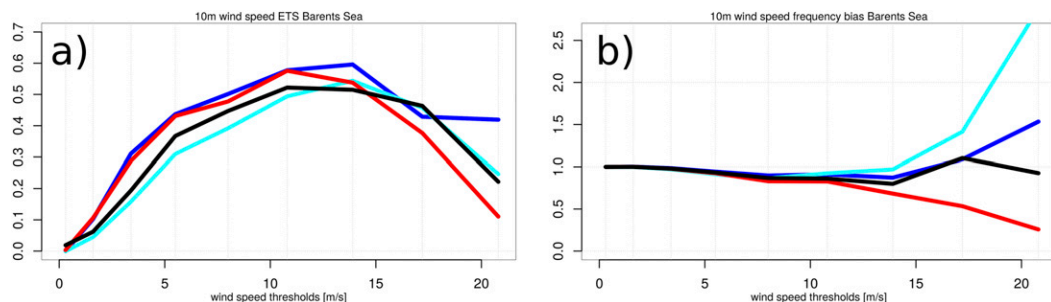


FIG. 9. As in Fig. 8, but WS10 forecasts are now compared with scatterometer-based observed wind for an area in the Barents Sea (24°–38°E and 72°–76°N). Forecasted wind from IFS-HRES (red), AROME-Arctic (blue), CAPS (black), and MFAROME (cyan). Notice that the highest threshold (20.8 m s⁻¹) includes 311 observations and 80, 477, 288, and 895 for the four models, respectively.

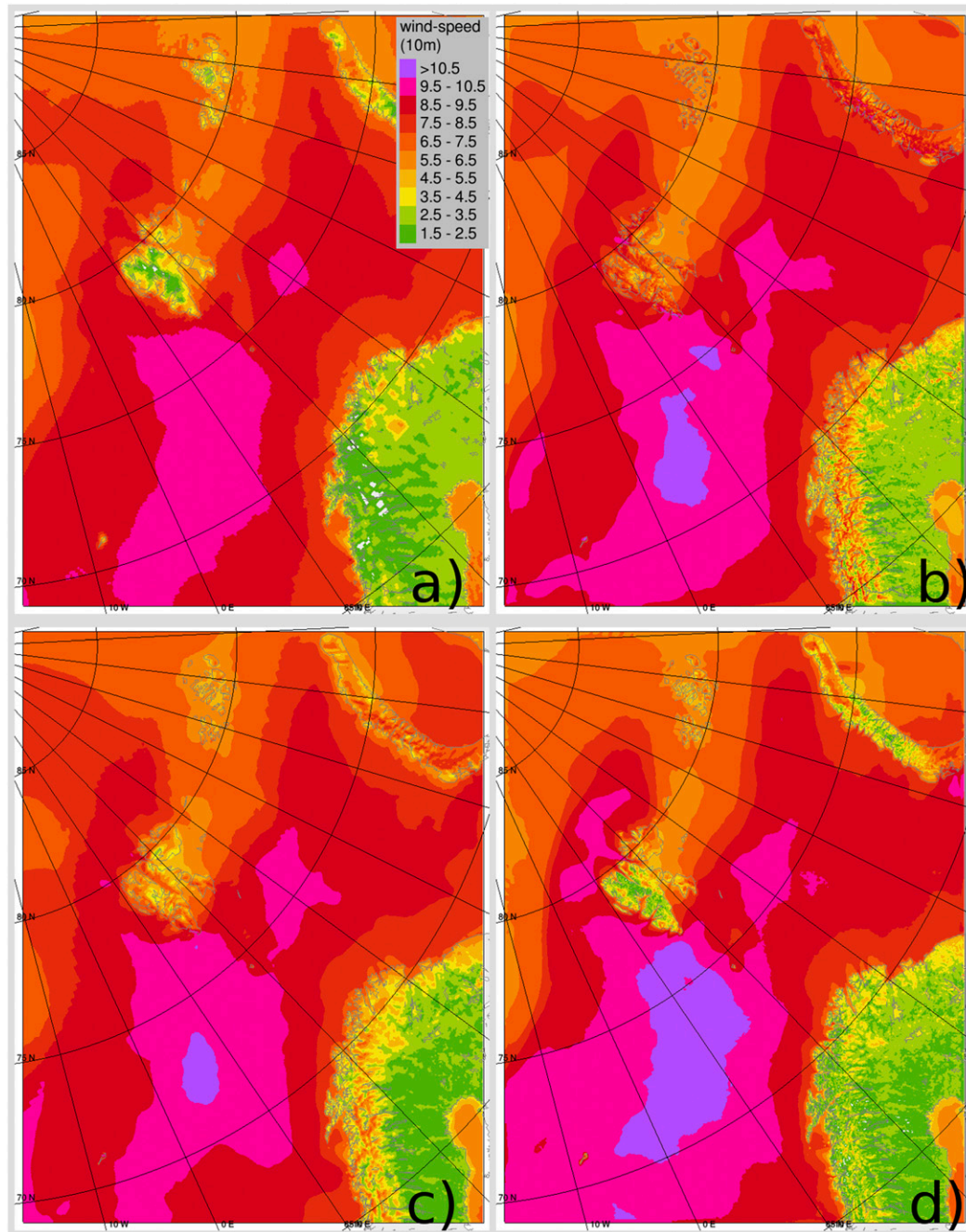


FIG. 10. Forecast average of 10-m wind speed for (a) IFS-HRES, (b) AROME-Arctic, (c) CAPS, and (d) MF-AROME. The averages are taken over YOPP SOP-NH1 for all lead times (hourly) from +25 to +48 h.

Water Resources and Energy Directorate. While the precipitation gauge-based estimate gives 19.7 mm, the increase in snow water equivalents gives 59.0 mm, indicating a substantial underestimation by the precipitation gauge in support of the adjusted accumulations in Fig. 12.

Ideally, the verification with adjusted precipitation should have included other metrics than only the

accumulated precipitation (e.g., skill scores). However, in single cases the undercatch is also influenced by particle shape, fall speed, and other microphysical properties in such a way that unrealistic errors will be introduced in skill verification. The adjustment algorithm therefore performs best averaged over many cases and is most appropriate for the estimation of systematic errors.

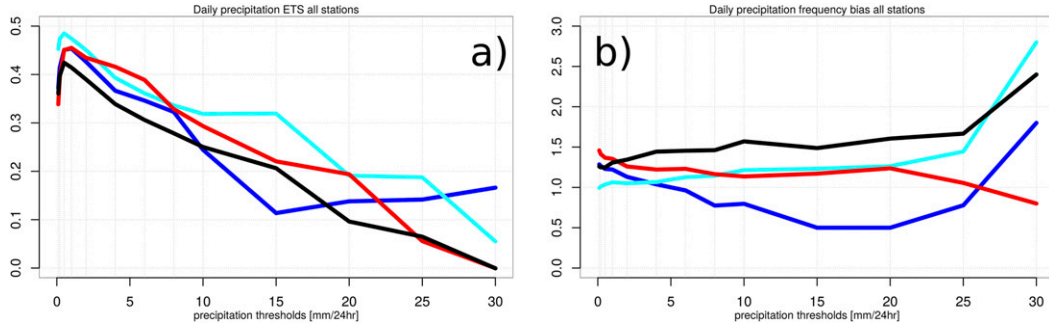


FIG. 11. (a) ETS and (b) FBI for accumulated daily precipitation (between lead time +18 and +42 h) over all stations used in the model intercomparison. Models are IFS-HRES (red), AROME-Arctic (blue), CAPS (black), and MF-AROME (cyan).

The gross features of forecasted spatial precipitation patterns are similar for all models (Fig. 13). All forecasts show maximum precipitation over steep topography at Svalbard, along the Norwegian coast and mountains, and at Nova Zemlja. However, the amplitude of the precipitation differs between models, that is, the high-resolution models produce higher maxima connected to the topography than IFS-HRES which has smoother precipitation (in agreement with Fig. 4). Another difference is that AROME-Arctic and MF-AROME have less precipitation over the ocean (and coast and fjords) than the other models. Note also that IFS-HRES has slightly more precipitation in sea ice covered areas which may be important for example, when forcing sea ice models.

e. Total cloud cover

The large-scale spatial patterns of TCC are similar in all forecast systems, but regional differences are found in the forecast climatologies (Fig. 14). All forecasts agree on a cloudy atmosphere over the ocean, but CAPS has less, while MF-AROME and IFS-HRES have a higher TCC. A noticeable difference in total cloud cover between the two AROME models are expected due to differences in their turbulence schemes (Bengtsson et al. 2017). Another noticeable feature is the maximum in cloud cover from IFS-HRES on the east side of the mountains at the border between Norway and Sweden not seen in the high-resolution models. The differences in forecasted cloud cover call for more investigations beyond the scope of this intercomparison study by using more appropriate cloud observations (e.g., satellite based measurements).

f. Observation, interpolation, and representativeness errors

The difference between forecasts and observations can be divided in model, observational, interpolation, and representativeness errors (Kanamitsu and DeHaan 2011). The actual performance of NWP systems will

become apparent only by taking the latter three components into consideration. In particular for short-term forecasts with relatively small forecast errors all components contribute significantly. We have tried to minimize the observational error by employing quality controlled observations and taking the undercatch of precipitation into account.

A station measurement for MSLP, T2, WS10, and precip24 represents a point-observation, which differs from what the gridbox value in a NWP system represents. This is due to subgrid phenomena (e.g., small-scale

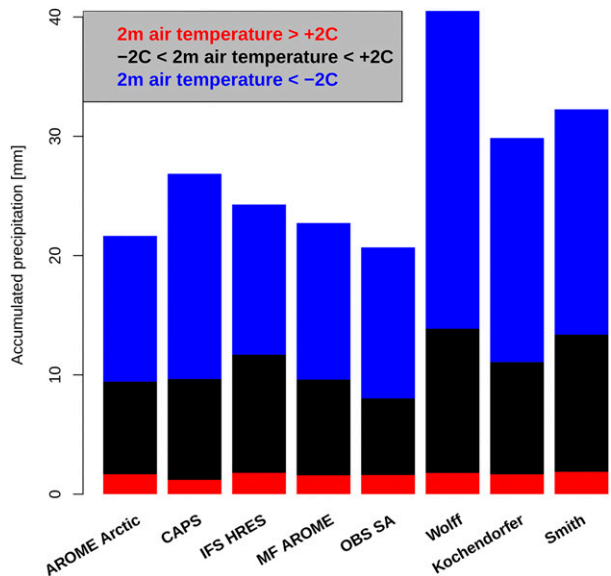


FIG. 12. Accumulated precipitation (estimated by temperature thresholds; rain in red, sleet in black, and solid precipitation in blue) for AROME-Arctic, CAPS, IFS-HRES, and MF-AROME with lead times from +25 to +48 h, observed precipitation from Geonor rain gauges with single-Alter shields, and observed precipitation corrected with Wolff et al. (2015), by Kochendorfer et al. (2017), and by Smith (2007). The accumulated precipitation amounts are averaged over 21 stations.

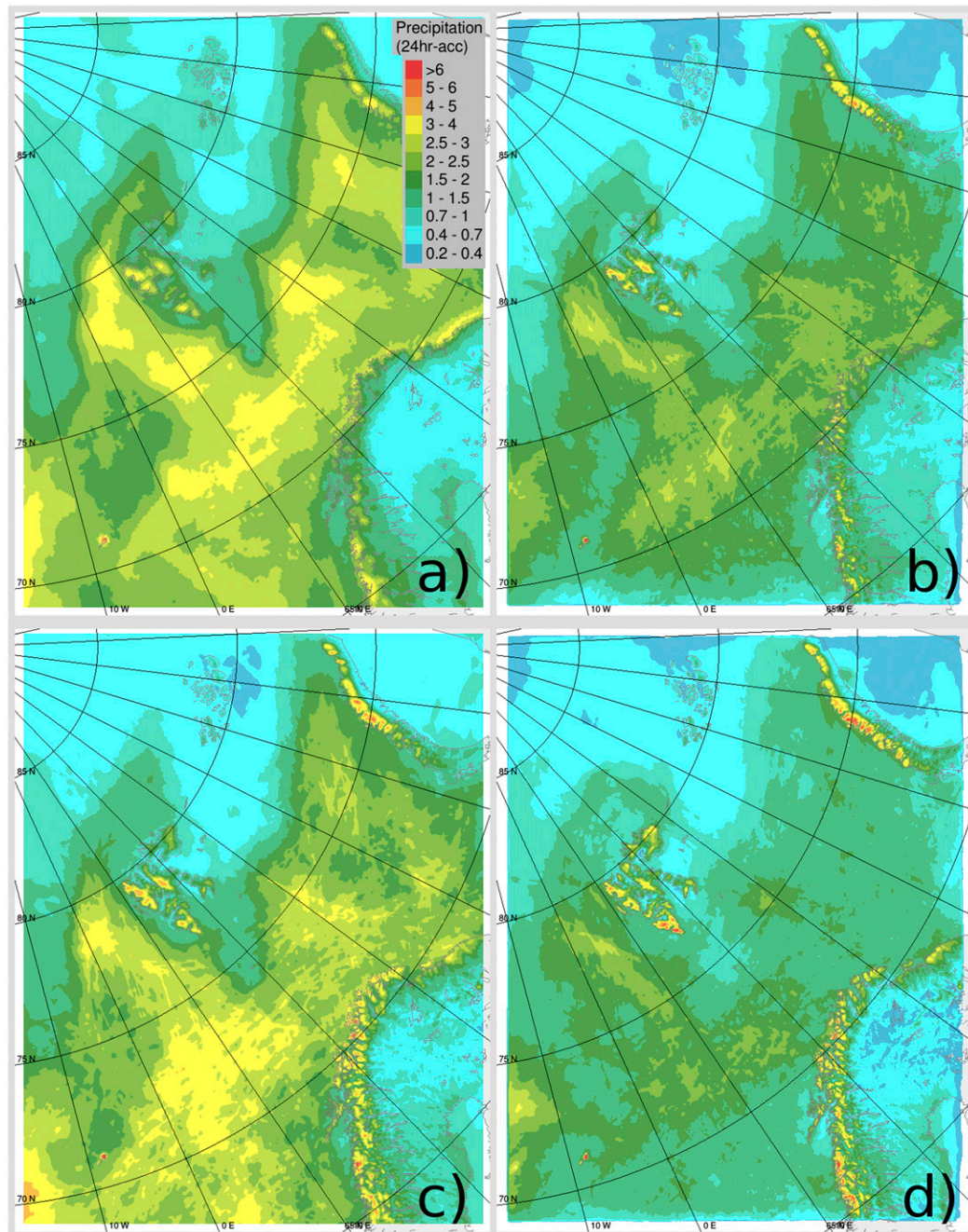


FIG. 13. Forecast average of daily precipitation for (a) IFS-HRES, (b) AROME-Arctic, (c) CAPS, and (d) MF-AROME. The averages are taken over YOPP SOP-NH1 for lead times between +25 and +48 h.

precipitation) and local effects, which cannot be reproduced by the model. Some representativeness issues are therefore present. To estimate these we include a simple example based on the approach of Göber et al. (2008). If several observations exist within a model grid box their average is assumed to represent an approximation of the grid box mean and will be treated as a “perfect” forecast.

However, the perfect forecast will not get perfect scores (e.g., SDE will not be 0 unless all observations are the same apart from constant differences), and the resulting error can be regarded as the representativeness error between a point and grid box average. Due to the sparse observational network a general estimate is difficult to establish. However, the two stations Tromsø (69.6536°N,

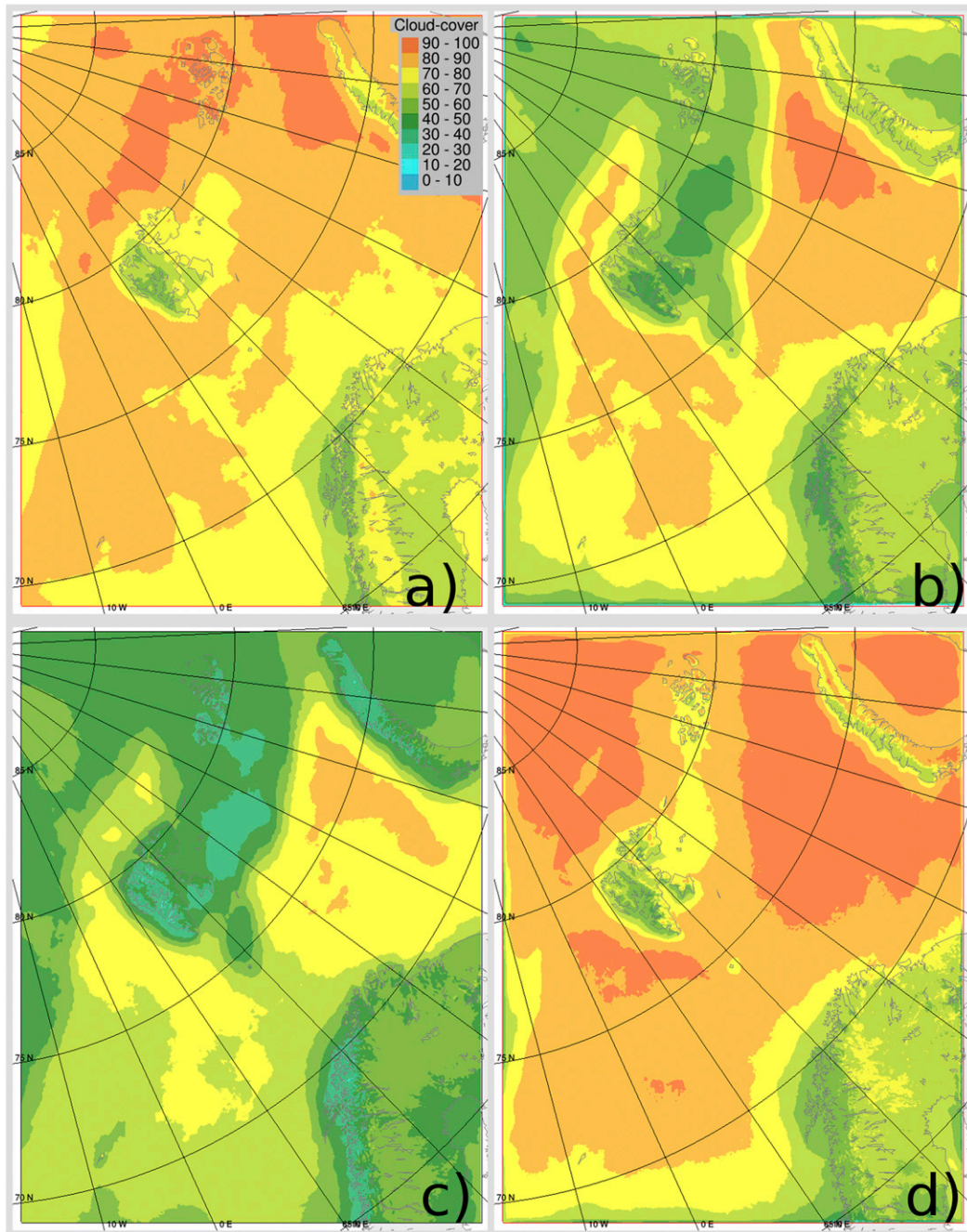


FIG. 14. Forecast average of total cloud cover for (a) IFS-HRES, (b) AROME-Arctic, (c) CAPS, and (d) MF-AROME. The averages are taken over YOPP SOP-NH1 for all lead times (hourly) from +25 to +48 h.

18.9368°E, 100 m MSL) and Tromsø Langnes (69.6767°N, 18.9133°E, 8 m MSL) are situated only 2.7 km apart. In Table 4 we verify a perfect forecast constructed by averaging these two observations and compare with the 4 NWP systems verified for the same observation sites. The representativeness part, estimated by the perfect forecast error divided by the NWP forecast error is

relatively small for MSLP (6%–11%), but higher for T2 (19%–35%), WS10 (36%–42%), and precip24 (15%–20%). Note that these are conservative estimates (for this kind of coastal location) since two stations are insufficient for generating a true grid box average. If the results from this example are more generally valid they can explain parts of the large (small) initial errors for

TABLE 4. SDE for a perfect forecast constructed by averaging observations following Göber et al. (2008) and for IFS-HRES, AROME-Arctic, CAPS, and MF-AROME during YOPP SOP-NH1. The last row shows the percentage of SDE from perfect forecast for the model with lowest/highest error.

	MSLP	T2	WS10	precip24
SDE perfect	0.08	0.58	0.81	0.39
SDE IFS	0.72	3.04	2.25	2.57
SDE AROME-Arctic	0.97	2.09	1.91	2.55
SDE CAPS	1.27	1.67	2.06	2.36
SDE MF-AROME	—	2.75	1.95	1.98
% of error	6%–11%	19%–35%	36%–42%	15%–20%

WS10 and T2 (MSLP) in Fig. 2 supported by Fig. 4 showing the rapid spatial decorrelation of wind speed. In addition, the better verification scores over ocean than over land discussed in section 4c can also be explained by representativeness issues. Haiden et al. (2012) have used the “perfect forecast” approach to estimate the effect of representativeness on precipitation scores such as ETS and FBI for a grid spacing of 25 km in Europe. They obtained a maximum achievable ETS around 0.75 and an FBI of 1.05. In summary, NWP forecasts perform better than the first impression given by verification statistics, and interpreting NWP output as point forecasts leads to scale mismatch effects that need consideration.

To estimate the sensitivity of the results to the interpolation method we calculate the root-mean-square error (RMSE) by using nearest grid point (used in all verification above) and bilinear interpolation methods. For MSLP the changes are negligible (less than 0.5%), while bilinear interpolation reduces errors for T2 (less than 4%), WS10 (less than 3%), precip24 (less than 2%), and TCC (less than 2%). Furthermore, we also upscale the three high-resolution models to a grid spacing comparable to IFS-HRES. In general, the RMSE changes by less than 5% with a few exceptions. For T2 the errors are reduced (in particular the nonsystematic part) by ~10% at the coast and island stations. An interpretation is that the high-resolution models have too sharp temperature gradients along the coast and a smoother field reduces the number of large errors. On the other hand, in the fjords and inland the systematic T2 error increases by 6%–7%. The interpretation is that the upscaling creates an undesirable mix of characteristics (e.g., fjords, valleys, mountains) in these areas. For WS10 (TCC) the errors increase (decrease) by less than 5%. Daily precipitation scores improve with upscaling inland (6%), while decreasing in fjords (5%).

4. High-impact weather case studies

To supplement the summary verification, we look in more detail at two high-impact cases during YOPP

SOP-NH1: 1) a mesoscale low pressure system in the Barents Sea and 2) a severe precipitation event at Svalbard.

a. Mesoscale low pressure system in the Barents Sea

In a southerly flow, a mesoscale disturbance with deep convection (Fig. 15a) and strong winds (Fig. 15b) developed south of the sea ice edge in the Barents Sea on 24 March 2018. Based on model analyses (a small spread between models exists) the low was located east of Bear Island at 1200 UTC (marked with L in Figs. 15a,b). All NWP systems develop a mesoscale disturbance with 24-h lead time (Figs. 16a–d) and also 48 h ahead (not shown). However, wind speed and minimum pressure and location vary. IFS-HRES forecasts (Fig. 16a) are less intense (higher minimum pressure and less windy) compared to the high-resolution models (Figs. 16b–d). The high-resolution models forecasted 25 m s^{-1} (AROME-Arctic), 24 m s^{-1} (MF-AROME), and 22 m s^{-1} (CAPS) as maximum wind speed, while maximum wind speed in ASCAT measurements are 22 m s^{-1} . In comparison IFS-HRES forecasted maximum wind speed of 20 m s^{-1} . At the Bear Island meteorological station (74.5°N , 19.0°E marked as red circle in Figs. 15 and 16) the maximum observed wind speed during the day is 19 m s^{-1} compared to 16 m s^{-1} from IFS-HRES, 18 m s^{-1} from AROME-Arctic and CAPS, and 19 m s^{-1} from MF-AROME. The observed wind speed is close to the observed maxima for a duration of 6 h and this is also seen, together with good timing of maximum wind, in all models with the exception of MF-AROME, which only gives a wind speed peak for 1 h. At Bear Island the minimum pressure in all forecasts is almost identical, about 2 hPa higher than observed.

The location of the mesoscale disturbance is similar in IFS-HRES forecasts for +24 and +48 h (50–100-km misplacement). However, the location of the system varies more with lead time in the high-resolution models (not shown). A closer inspection of the wind pattern of MF-AROME (Fig. 16d) indicates a significant change in location compared with IFS-HRES, AROME-Arctic, and CAPS forecasts and available

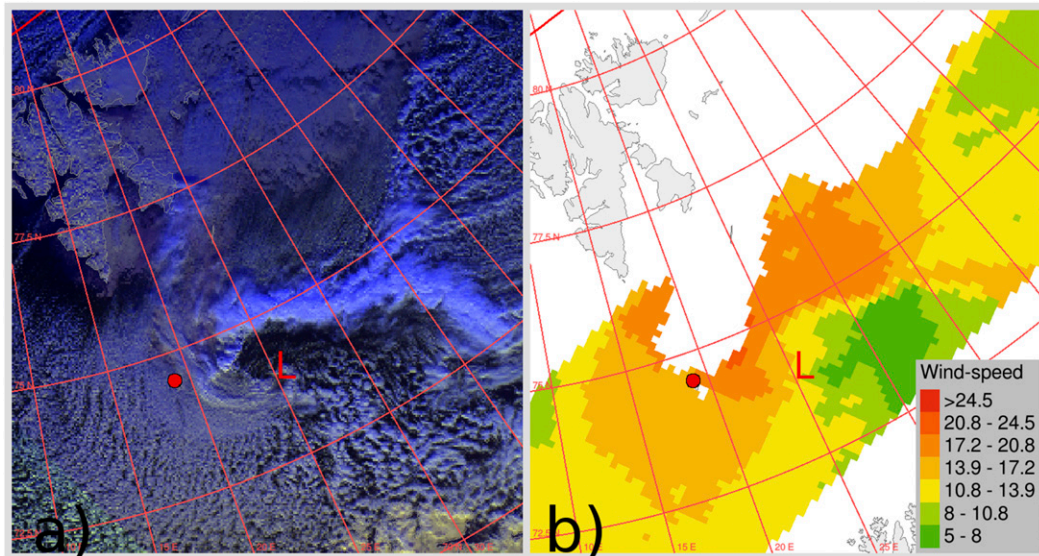


FIG. 15. Polar mesoscale low pressure system at 1200 UTC 24 Mar 2018: (a) NOAA satellite picture and (b) ASCAT wind speed (in white areas measurements are masked due to sea ice or land/islands). Low pressure center in IFS-HRES analysis is marked with “L” and Bear Island marked as a red circle. Every 5° longitude and every 2.5° latitude is plotted in red to make comparisons easier.

observations and analysis. To investigate this further, reruns of MF-AROME with initial conditions from AROME-Arctic were performed. Only changing the initial surface conditions (Fig. 16e) did not improve the location of the mesoscale disturbance. However, additionally changing the upper-air initial conditions in MF-AROME by using analysis from AROME-Arctic (Fig. 16f) improved the low pressure position significantly (misplacement reduced from approximately 230 to 90 km).

In this case all forecast systems simulate the mesoscale low pressure system. The benefit of IFS-HRES was more consistent forecasts of location for different lead times, while the high-resolution models better captured the highest wind speeds in agreement with earlier studies (e.g., McInnes et al. 2011). A bad location of the system in the +24-h forecast from MF-AROME was drastically improved by changes in the initial conditions.

b. Precipitation event at Svalbard

On 26 February 2018, a high pressure system over northern Scandinavia and a low pressure system west and north of Svalbard provided favorable conditions for the transport of heat and moisture (mainly below 800 hPa) toward Svalbard. This type of atmospheric large-scale setup is responsible for a majority of the high-impact precipitation events (rain on snow) at Svalbard, which have a substantial impact on infrastructure, society, and wildlife (Serreze et al. 2015; Hansen et al. 2014). The maximum precipitation measured was 61.0 mm in 36 h at

Ny-Ålesund (marked with A in Fig. 17). This might seem small compared to midlatitude extreme values, but 46.0 mm (measured in the first 24 h of the period) was the fourth-largest daily accumulated precipitation amount between August 2008 and August 2018. In addition, METAR temperature observations indicate that the majority of the precipitation was rain on frozen ground. Already on 28 February the daily mean temperature was close to -10°C and stayed below -10°C for the next two weeks, maintaining the surface ice conditions.

Precipitation forecasts for the Svalbard area (36-h accumulations) are shown in Fig. 17. All forecasts have a general agreement with the observations (Table 5) in that the highest precipitation amounts are in the northwest of Svalbard (point A; Ny-Ålesund), but which model is closest to observed values varies between observation sites (Table 5). Furthermore, the high-resolution models have more spatial details and higher maximum values than IFS-HRES (Fig. 17). However the local details are difficult to verify due to the lack of observations. One exception is the area around Longyearbyen (points B, C, and D), where there are sharp gradients in the observations from Platåberget, 450 m MSL (point B) 13.4 mm, Svalbard Airport (point C) 17.2 mm, and Adventdalen (point D) 2.8 mm. MF-AROME was able to capture some local differences with forecasts between 4.1 and 21.5 mm $(36\text{ h})^{-1}$ in the same area (see reduced precipitation in the Adventdalen east of points B, C, and D). It should be noted that even if local maximum precipitation values are higher in

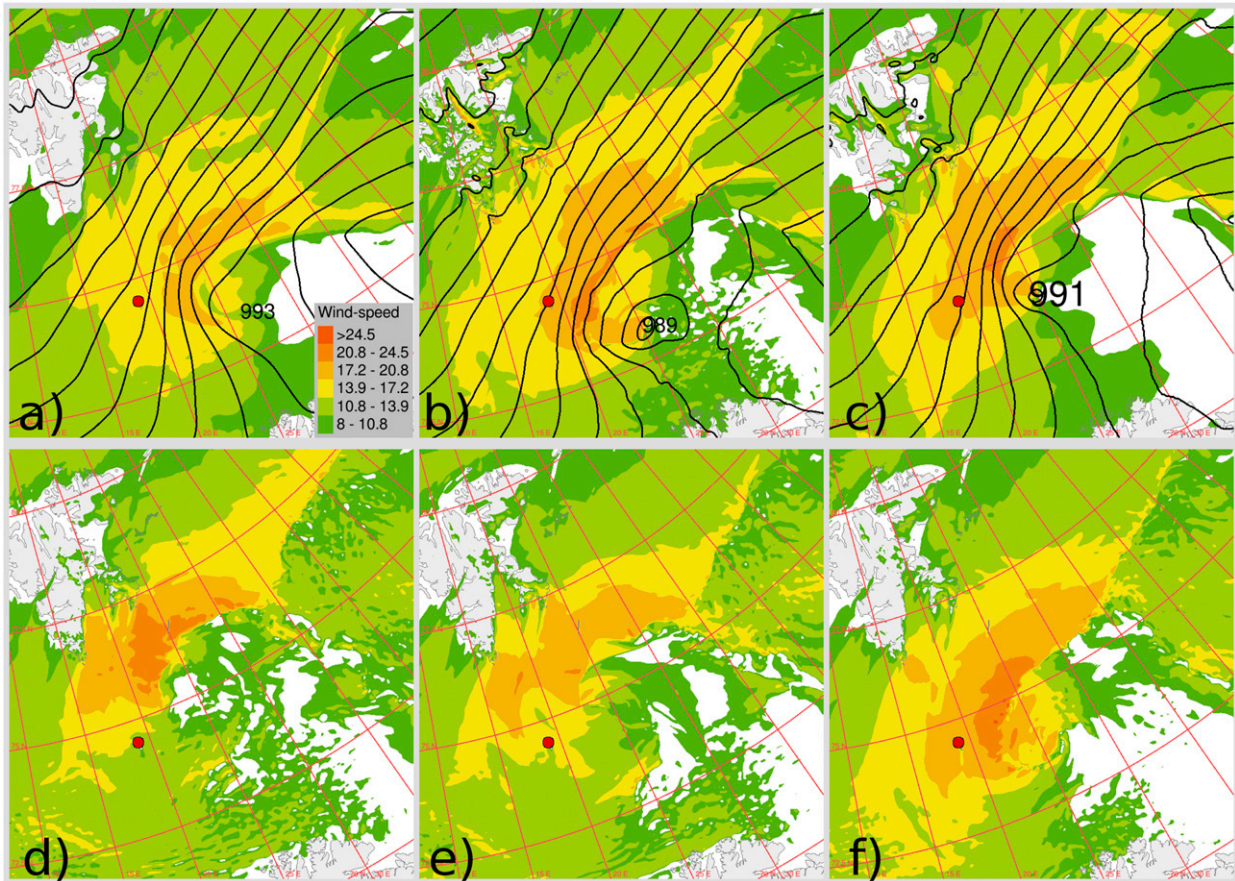


FIG. 16. MSLP and 10-m wind speed forecasts with +24-h lead time for (a) IFS-HRES, (b) AROME-Arctic, (c) CAPS, (d) MF-AROME, (e) MF-AROME with surface initial conditions from AROME-Arctic, and (f) MF-AROME with surface and upper-air initial conditions from AROME-Arctic. Notice that MSLP is not available from MF-AROME. The red circle indicates Bear Island.

the high-resolution forecasts the average precipitation over the entire Svalbard archipelago is 18%–26% higher in IFS-HRES.

It is important to correctly forecast precipitation type in these situations. Since direct observations of precipitation type are rare in time and space we use 2-m air temperature as a proxy. Evidently, such a proxy has limitations since it neglects information about the temperature and humidity profile. Averaged over the 36-h precipitation accumulation period the forecasts have negative temperature biases: IFS-HRES, -2.0°C ; AROME-Arctic, -1.4°C ; CAPS, -1.8°C ; and MF, -2.3°C , indicating too much solid precipitation and too little rain. If we assume that the precipitation will be rain when the temperature exceeds $+1^{\circ}\text{C}$ (Jennings et al. 2018), we find that the forecasts suggest that 70% (AROME-Arctic), 16% (IFS-HRES), 5% (CAPS), and 43% (MF-AROME) of the precipitation fell as rain at the observation sites. However, the METAR observations in Ny-Ålesund and Longyearbyen indicated rain for most of the period and if we replace the forecasted temperature with observed

temperature and keep the 1°C threshold we get approximately 80% as rain.

In summary, the potential added value of the high-resolution models for this case is associated with higher maximum precipitation and a redistribution of the precipitation patterns (forced by topography). In addition, the high-resolution models have the potential to improve precipitation type in complex terrain, compared to IFS-HRES.

5. Summary

In this study, short-range forecasts from one global (IFS-HRES) and three regional NWP systems (AROME-Arctic, CAPS, and MF-AROME) are compared in the European Arctic (Fig. 1). The model intercomparison seeks to establish a baseline or reference for Arctic forecasting capabilities of near-surface parameters as suggested by Jung et al. (2016). The forecast systems differ in model formulation, resolution, initialization methods and lateral boundary forcing (Table 1). IFS-HRES and

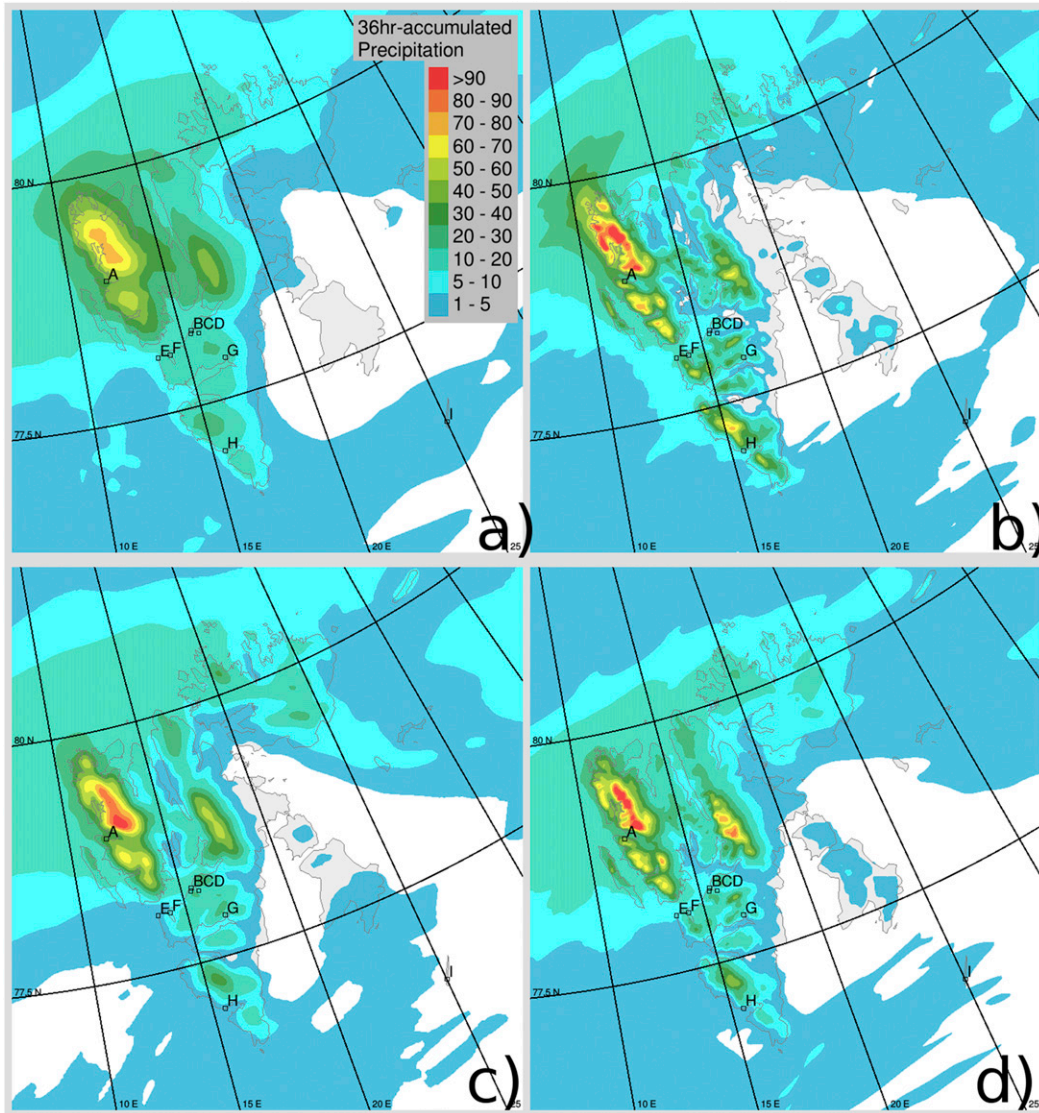


FIG. 17. Forecasts of 36-h accumulated precipitation for Svalbard in the period from 0600 UTC 26 Feb to 1800 UTC 27 Feb 2018 with lead times from +6 to +42 h: (a) IFS-HRES, (b) AROME-Arctic, (c) CAPS, and (d) MF-AROME. Observation sites are marked with letters as follows: Ny-Ålesund (A), Platåberget (B), Svalbard Airport (C), Adventdalen (D), Isfjorden Radio (E), Barentsburg (F), Sveagruva (G), Hornsund (H), and Hopen (I).

AROME-Arctic are operational systems (i.e., real-time multiple daily runs), which include data assimilation, while CAPS and MF-AROME are specific contributions to YOPP and initialized from global models. The lateral boundary conditions for the three regional systems are taken from different global forecast systems; IFS-HRES forces AROME-Arctic, GDPS forces CAPS, and ARPEGE forces MF-AROME. Differences in forecast characteristics, weaknesses, and strengths therefore can have a variety of sources which are not always easy to pinpoint. The comparison is performed for YOPP SOP NH1, a winter period with availability of extra radiosondes

in the Arctic which are expected to improve the actual forecast skill. The period includes a range of large-scale flow configurations and periods with both positive (February) and negative NAO values (March).

Forecast accuracy varies across regions, parameters, lead times, and NWP systems, and no NWP system is superior to the other systems in all aspects. However, compared to the other models, AROME-Arctic has the advantage of surface and upper-air assimilation (as IFS-HRES), high horizontal resolution (as MF-AROME and CAPS), and model development with a focus on the specific area of this comparison. These advantages are

TABLE 5. Observed and forecasted (from +6 to +42 h) 36-h accumulated precipitation from 0600 UTC 26 Feb to 1800 UTC 27 Feb 2018 at Ny-Ålesund (A), Platåberget (B), Svalbard Airport (C), Adventdalen (D), Isfjorden Radio (E), Barentsburg (F), Sveagruva (G), Hornsund (H), and Hopen (I).

36-h accumulated precip	A	B	C	D	E	F	G	H	I
Observed	61.0	13.4	17.2	2.8	10.1	15.0	18.1	1.6	0.4
IFS-HRES	37.0	16.7	16.7	16.7	9.0	11.4	16.8	8.5	1.3
AROME-Arctic	42.6*	12.7	12.6	15.2	5.5	12.7	7.6	4.5	1.0
CAPS	37.0	16.6	15.5	16.5	3.0	6.0	12.1	2.3	0.9
MF-AROME	29.3	7.5	5.4	2.5	13.7	14.6	12.9	8.3	1.4

reflected in the verification, where AROME-Arctic on average performs better than the other models.

There is a general agreement between models on the larger-scale patterns of average cloud cover, temperature maxima, and wind speed maxima over ocean areas; temperature minima over sea ice; and precipitation maxima connected to topographic and coastal forcing. IFS-HRES verifies best regarding MSLP, but all systems are in good agreement with observations (SDEs less than 1 hPa initially and 2 hPa or less after +48 h). Larger differences between forecasted and observed MSLP are mainly found in mountain areas, where it is problematic to reduce surface pressure to mean sea level as shown by Pauley (1998).

Several common model deficiencies are noted, although their magnitude varies between the different NWP systems. Problems associated with T2 forecasts inland in cloud-free and calm conditions during nighttime, related to the representation of the stable boundary layer are well known and studied (e.g., Sandu et al. 2013; Haiden et al. 2018b; Esau et al. 2018). Opposite to this all models show a cold bias under windy conditions. Another common deficiency is the low skill in distinguishing between freezing and nonfreezing conditions inland which is important for Arctic infrastructure, society, and wildlife (Hansen et al. 2014). For wind speed forecasts, the models find it difficult to reproduce the high spatial variations of WS10 over land and the high-resolution models forecast generally more wind than IFS-HRES [e.g., as seen in DuVivier et al. (2017) and Walsh et al. (2007)]. However, in particular over the ocean the skill is not necessarily improved by finer horizontal resolution alone (similar to results from Kalverla et al. 2019). Furthermore, adjusting for the undercatch of solid precipitation in observations reveals that most likely all forecast systems have too little precipitation in the area studied. This is an important finding because this feature is not apparent if undercatch in observations is not considered in the verification process (which often is the case).

For near-surface weather parameters (i.e., T2, WS10, precipitation) there are also several examples of

differences in local forecast skill between NWP systems, for example, a cold bias is found related to overestimation of sea ice in the surroundings of Svalbard for CAPS, while AROME-Arctic has a pronounced underestimation of precipitation at the coast and fjords (still under investigation). Furthermore, over land, IFS-HRES and MF-AROME underestimate the wind speed. It is particular that at higher elevations (e.g., mountain, inner part of Svalbard) the two models have less wind than AROME-Arctic, which on average has the highest wind speeds (Fig. 10). In addition, wind forecasts over ocean from MF-AROME and CAPS are less accurate than IFS-HRES and AROME-Arctic. The sensitivity to initial conditions is investigated in a rerun of MF-AROME. The original initial conditions (dynamical adaptation from the global ARPEGE model) are replaced with initial conditions from the AROME-Arctic data assimilation. For a case study with a mesoscale low in the Barents Sea the new initial conditions improve the location of the mesoscale low by 140 km. Similar runs with initial surface conditions from AROME-Arctic in MF-AROME runs reduce the MF-AROME T2 errors to the same level as AROME-Arctic and highlight the importance of surface assimilation as also shown in Randriamampianina et al. (2019).

The forecast climatologies also reveal that there are differences that are not evaluated in this study due to the sparseness of observations. This includes differences over areas covered by sea ice (e.g., T2, TCC, and precip24), ocean areas (TCC, precip24), and inland and mountain areas at Svalbard (e.g., WS10 and T2). A comparison of forecasted TCC with satellite based TCC estimates would be a natural extension of this work, together with the use of available field campaign, ship, and buoy data over the sea ice and ocean.

Regional high-resolution models can add value compared to global models by using finer resolution and domain-tailored process representations (Jung et al. 2016). In this study, the added value of the high-resolution models compared to IFS-HRES is most pronounced and significant for WS10 and T2 in regions with complex terrain and coast lines, as also found in

numerous other studies (e.g., Rummukainen 2016; Schellander-Gorgas et al. 2017). In contrast, in this study the added value is negligible or negative for some parameters and regions, for example, MSLP and total cloud cover in general, and for temperature and wind speed at islands. In addition, it is shown that the errors grow faster in the high-resolution models, indicating that the added value of high-resolution models depends on lead time.

In polar regions, the limited availability of reliable observations is one of the greatest challenges in the verification process (Casati et al. 2017). Furthermore, verification often compares grid box values with point observations. It is important to acknowledge that differences between forecasts and observations arise from observation, interpolation and representativeness errors in addition to model errors. In this study, it was found that observation errors and representativeness issues contribute substantially to the difference between forecasted and observed WS10, T2, and (solid) precipitation. We found large initial errors for WS10 (SDE $\sim 2.5 \text{ m s}^{-1}$) and T2 (SDE $\sim 3^\circ\text{C}$) indicating observation representativeness issues. In addition, an example from two observation sites situated close to each other shows that the subgrid variability, even for high-resolution models, for this particular example contributes a large part of the difference between predicted and observed WS10 ($\sim 40\%$), T2 ($\sim 25\%$), and daily precipitation ($\sim 15\%$). Furthermore, more skillful WS10 forecasts are seen over ocean (against ASCAT data) than over land (against SYNOP), which may be due to representativeness issues of wind observations (Wieringa 1996). As the forecast systems improve, and in particular for short-range forecasts, it is important to quantify and understand all error components and interpret results accordingly.

Acknowledgments. The work described in this paper has received funding from the European Union's Horizon 2020 Research and Innovation programme through Grant Agreement 727862 APPLICATE. The content is the sole responsibility of the author(s) and it does not represent the opinion of the European Commission, and the Commission is not responsible for any use that might be made of information contained. This study was also supported by the Norwegian Research Council Project 280573 'Advanced models and weather prediction in the Arctic: enhanced capacity from observations and polar process representations (ALERTNESS).' This work is a contribution to the Year of Polar Prediction (YOPP), a flagship activity of the Polar Prediction Project (PPP), initiated by the World Weather Research Programme (WWRP) of the World Meteorological Organization

(WMO). The scientific exchange and discussion within the Year of Polar Prediction verification team have significantly contributed to the work described in this paper.

REFERENCES

- Balsamo, G., A. Beljaars, and K. Scipal, 2009: A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the Integrated Forecast System. *J. Hydrometeorol.*, **10**, 623–643, <https://doi.org/10.1175/2008JHM1068.1>.
- Batrak, Y., E. Kourzeneva, and M. Homleid, 2018: Implementation of a simple thermodynamic sea ice scheme, SICE version 1.0-38h1, within the ALADIN–HIRLAM numerical weather prediction system version 38h1. *Geosci. Model Dev.*, **11**, <https://doi.org/10.5194/gmd-11-3347-2018>.
- Bauer, P., L. Magnusson, J.-N. Thépaut, and T. M. Hamill, 2016: Aspects of ECMWF model performance in polar areas. *Quart. J. Roy. Meteor. Soc.*, **142**, 583–596, <https://doi.org/10.1002/qj.2449>.
- Bélair, S., L. Crevier, J. Mailhot, B. Bilodeau, and Y. Delage, 2003: Operational Implementation of the ISBA land surface scheme in the Canadian regional weather forecast model. Part I: Warm season results. *J. Hydrometeorol.*, **4**, 352–370, [https://doi.org/10.1175/1525-7541\(2003\)4<352:OIOTIL>2.0.CO;2](https://doi.org/10.1175/1525-7541(2003)4<352:OIOTIL>2.0.CO;2).
- Bengtsson, L., and Coauthors, 2017: The HARMONIE–AROME Model Configuration in the ALADIN–HIRLAM NWP System. *Mon. Wea. Rev.*, **145**, 1919–1935, <https://doi.org/10.1175/MWR-D-16-0417.1>.
- Brasnett, B., 2008: The impact of satellite retrievals in a global sea-surface-temperature analysis. *Quart. J. Roy. Meteor. Soc.*, **134**, 1745–1760, <https://doi.org/10.1002/qj.319>.
- Buehner, M., and Coauthors, 2015: Implementation of deterministic weather forecasting systems based on ensemble-variational data assimilation at Environment Canada. Part I: The global system. *Mon. Wea. Rev.*, **143**, 2532–2559, <https://doi.org/10.1175/MWR-D-14-00354.1>.
- , A. Caya, T. Carrieres, and L. Pogson, 2016: Assimilation of SSMIS and ASCAT data and the replacement of highly uncertain estimates in the Environment Canada Regional Ice Prediction System. *Quart. J. Roy. Meteor. Soc.*, **142**, 562–573, <https://doi.org/10.1002/qj.2408>.
- Buizza, R., and Coauthors, 2017: IFS Cycle 43r3 brings model and assimilation updates. *ECMWF Newsletter*, No. 152, ECMWF, Reading, United Kingdom, 18–22, <https://www.ecmwf.int/en/newsletter/152/meteorology/ifs-cycle-43r3-brings-model-and-assimilation-updates>.
- Casati, B., T. Haiden, B. Brown, P. Nurmi, and J.-F. Lemieux, 2017: Verification of environmental prediction in polar regions: Recommendations for the Year of Polar Prediction. WWRP 2017-1, WMO, 44 pp.
- Davies, T., 2014: Lateral boundary conditions for limited area models. *Quart. J. Roy. Meteor. Soc.*, **140**, 185–196, <https://doi.org/10.1002/qj.2127>.
- Douville, H., J.-F. Royer, and J.-F. Mahfouf, 1995: A new snow parameterization for the Météo-France climate model: Part I: Validation in stand-alone experiments. *Climate Dyn.*, **12**, 21–35, <https://doi.org/10.1007/BF00208760>.
- Dutra, E., G. Balsamo, P. Viterbo, P. M. A. Miranda, A. Beljaars, C. Schär, and K. Elder, 2010: An improved snow scheme for the ECMWF land surface model: Description and offline validation. *J. Hydrometeorol.*, **11**, 899–916, <https://doi.org/10.1175/2010JHM1249.1>.

- DuVivier, A. K., J. J. Cassano, S. Greco, and G. D. Emmitt, 2017: A case study of observed and modeled barrier flow in the Denmark Strait in May 2015. *Mon. Wea. Rev.*, **145**, 2385–2404, <https://doi.org/10.1175/MWR-D-16-0386.1>.
- Esau, I., M. Tolstykh, R. Fadeev, V. Shashkin, S. Makhnorylova, V. Miles, and V. Melnikov, 2018: Systematic errors in northern Eurasian short-term weather forecasts induced by atmospheric boundary layer thickness. *Environ. Res. Lett.*, **13**, 125009, <https://doi.org/10.1088/1748-9326/aaecfb>.
- Gascard, J.-C., K. Riemann-Campe, R. Gerdes, H. Schyberg, R. Randriamampianina, M. Karcher, J. Zhang, and M. Rafizadeh, 2017: Future sea ice conditions and weather forecasts in the Arctic: Implications for Arctic shipping. *Ambio*, **46**, 355–367, <https://doi.org/10.1007/s13280-017-0951-5>.
- Göber, M., E. Zsótér, and D. S. Richardson, 2008: Could a perfect model ever satisfy a naïve forecaster? On grid box mean versus point verification. *Meteor. Appl.*, **15**, 359–365, <https://doi.org/10.1002/met.78>.
- Haiden, T., M. J. Rodwell, D. S. Richardson, A. Okagaki, T. Robinson, and T. Hewson, 2012: Intercomparison of global model precipitation forecast skill in 2010/11 using the SEEPS score. *Mon. Wea. Rev.*, **140**, 2720–2733, <https://doi.org/10.1175/MWR-D-11-00301.1>.
- , M. Janousek, J. Bidlot, R. Buizza, L. Ferranti, F. Prates, and F. Vitart, 2018a: Evaluation of ECMWF forecasts, including the 2018 upgrade. ECMWF Tech. Memo. 831, 52 pp., <https://doi.org/10.21957/ldw15ckqi>.
- , I. Sandu, G. Balsamo, G. Arduini, and A. Beljaars, 2018b: Addressing biases in near-surface forecasts. *ECMWF Newsletter*, No. 157, ECMWF, Reading, United Kingdom, 20–25, <https://www.ecmwf.int/en/newsletter/157/meteorology/addressing-biases-near-surface-forecasts>.
- Hansen, B. B., and Coauthors, 2014: Warmer and wetter winters: Characteristics and implications of an extreme weather event in the High Arctic. *Environ. Res. Lett.*, **9**, 114021, <https://doi.org/10.1088/1748-9326/9/11/114021>.
- Hanssen-Bauer, I., E. J. Førland, H. Hisdal, S. Mayer, A. B. Sandø, and A. Sorteberg, 2019: Climate in Svalbard 2100. NCCS Rep. 1/2019, 205 pp., <https://www.miljodirektoratet.no/globalassets/publikasjoner/M1242/M1242.pdf>.
- Jennings, K. S., T. S. Winchell, B. Livneh, and N. P. Molotch, 2018: Spatial variation of the rain–snow temperature threshold across the Northern Hemisphere. *Nat. Commun.*, **9**, 1148, <https://doi.org/10.1038/s41467-018-03629-7>.
- Jolliffe, I. T., and D. B. Stephenson, Eds., 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Sciences*. Wiley-Blackwell, 292 pp.
- Jung, T., and M. Leutbecher, 2007: Performance of the ECMWF forecasting system in the Arctic during winter. *Quart. J. Roy. Meteor. Soc.*, **133**, 1327–1340, <https://doi.org/10.1002/qj.99>.
- , and M. Matsueda, 2016: Verification of global numerical weather forecasting systems in polar regions using TIGGE data. *Quart. J. Roy. Meteor. Soc.*, **142**, 574–582, <https://doi.org/10.1002/qj.2437>.
- , and Coauthors, 2016: Advancing polar prediction capabilities on daily to seasonal time scales. *Bull. Amer. Meteor. Soc.*, **97**, 1631–1647, <https://doi.org/10.1175/BAMS-D-14-00246.1>.
- Kalverla, P., G.-J. Steeneveld, R. Ronda, and A. A. M. Holtslag, 2019: Evaluation of three mainstream numerical weather prediction models with observations from meteorological mast IJmuiden at the North Sea. *Wind Energy*, **22**, 34–48, <https://doi.org/10.1002/we.2267>.
- Kanamitsu, M., and L. DeHaan, 2011: The Added Value Index: A new metric to quantify the added value of regional models. *J. Geophys. Res.*, **116**, D11106, <https://doi.org/10.1029/2011JD015597>.
- Karpechko, A. Y., 2018: Predictability of sudden stratospheric warmings in the ECMWF extended-range forecast system. *Mon. Wea. Rev.*, **146**, 1063–1075, <https://doi.org/10.1175/MWR-D-17-0317.1>.
- Kielland, G., 2005: KVALOBS - The quality assurance system of Norwegian Meteorological Institute observations. Instruments and Observing Methods, Rep. 82, WMO/TD-1265, 3(13), https://library.wmo.int/pmb_ged/wmo-td_1265.pdf.
- Kochendorfer, J., and Coauthors, 2017: The quantification and correction of wind-induced precipitation measurement errors. *Hydrol. Earth Syst. Sci.*, **21**, 1973–1989, <https://doi.org/10.5194/hess-21-1973-2017>.
- Kristjánsson, J. E., and Coauthors, 2011: The Norwegian IPY–THORPEX: Polar lows and Arctic fronts during the 2008 Andøya Campaign. *Bull. Amer. Meteor. Soc.*, **92**, 1443–1466, <https://doi.org/10.1175/2011BAMS2901.1>.
- Marzban, C., S. Sandgathe, H. Lyons, and N. Lederer, 2009: Three spatial verification techniques: Cluster analysis, variogram, and optical flow. *Wea. Forecasting*, **24**, 1457–1471, <https://doi.org/10.1175/2009WAF2222261.1>.
- McInnes, H., J. Kristiansen, J. E. Kristjánsson, and H. Schyberg, 2011: The role of horizontal resolution for polar low simulations. *Quart. J. Roy. Meteor. Soc.*, **137**, 1674–1687, <https://doi.org/10.1002/qj.849>.
- McTaggart-Cowan, R., C. Girard, A. Plante, and M. Desgagné, 2011: The utility of upper-boundary nesting in NWP. *Mon. Wea. Rev.*, **139**, 2117–2144, <https://doi.org/10.1175/2010MWR3633.1>.
- Mittermaier, M., 2012: A critical assessment of surface cloud observations and their use for verifying cloud forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1794–1807, <https://doi.org/10.1002/qj.1918>.
- Müller, M., Y. Batrak, J. Kristiansen, M. A. Køltzow, G. Noer, and A. Korosov, 2017: Characteristics of a convective-scale weather forecasting system for the European Arctic. *Mon. Wea. Rev.*, **145**, 4771–4787, <https://doi.org/10.1175/MWR-D-17-0194.1>.
- Noilhan, J., and S. Planton, 1989: A simple parameterization of land surface processes for meteorological models. *Mon. Wea. Rev.*, **117**, 536–549, [https://doi.org/10.1175/1520-0493\(1989\)117<0536:ASPOLS>2.0.CO;2](https://doi.org/10.1175/1520-0493(1989)117<0536:ASPOLS>2.0.CO;2).
- Nordeng, T. E., G. Brunet, and J. Caughey, 2007: Improvements of weather forecasts in polar regions. *WMO Bull.*, **56**, 250–257.
- Pauley, P. M., 1998: An example of uncertainty in sea level pressure reduction. *Wea. Forecasting*, **13**, 833–850, [https://doi.org/10.1175/1520-0434\(1998\)013<0833:AEouis>2.0.CO;2](https://doi.org/10.1175/1520-0434(1998)013<0833:AEouis>2.0.CO;2).
- Randriamampianina, R., H. Schyberg, and M. Mile, 2019: Observing system experiments with an Arctic mesoscale numerical weather prediction model. *Remote Sens.*, **11**, 981, <https://doi.org/10.3390/rs11080981>.
- Rasmussen, R., and Coauthors, 2012: How well are we measuring snow: The NOAA/FAA/NCAR winter precipitation test bed. *Bull. Amer. Meteor. Soc.*, **93**, 811–829, <https://doi.org/10.1175/BAMS-D-11-00052.1>.
- Rummukainen, M., 2016: Added value in regional climate modeling. *Wiley Interdiscip. Rev.: Climate Change*, **7**, 145–159, <https://doi.org/10.1002/wcc.378>.
- Samuelsen, E. M., 2018: Ship-icing prediction methods applied in operational weather forecasting. *Quart. J. Roy. Meteor. Soc.*, **144**, 13–33, <https://doi.org/10.1002/qj.3174>.
- Sandu, I., A. Beljaars, P. Bechtold, T. Mauritsen, and G. Balsamo, 2013: Why is it so difficult to represent stably stratified conditions

- in numerical weather prediction (NWP) models? *J. Adv. Model. Earth Syst.*, **5**, 117–133, <https://doi.org/10.1002/jame.20013>.
- Schellander-Gorgas, T., Y. Wang, F. Meier, F. Weidle, C. Wittmann, and A. Kann, 2017: On the forecast skill of a convection-permitting ensemble. *Geosci. Model Dev.*, **10**, 35–56, <https://doi.org/10.5194/gmd-10-35-2017>.
- Seity, Y., P. Brousseau, S. Malardel, G. Hello, P. Bénard, F. Bouttier, C. Lac, and V. Masson, 2011: The AROME-France convective-scale operational model. *Mon. Wea. Rev.*, **139**, 976–991, <https://doi.org/10.1175/2010MWR3425.1>.
- Serreze, M. C., A. D. Crawford, and A. P. Barrett, 2015: Extreme daily precipitation events at Spitsbergen, an Arctic Island. *Int. J. Climatol.*, **35**, 4574–4588, <https://doi.org/10.1002/joc.4308>.
- Sheridan, P., S. Smith, A. Brown, and S. Vosper, 2010: A simple height-based correction for temperature downscaling in complex terrain. *Meteor. Appl.*, **17**, 329–339, <https://doi.org/10.1002/met.177>.
- Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, **132**, 3019–3032, <https://doi.org/10.1175/MWR2830.1>.
- Smith, C. D., 2007: Correcting the wind bias in snowfall measurements made with a Geonor T-200B precipitation gauge and Alter wind shield. *14th Symp. on Meteorological Observation and Instrumentation*, San Antonio, TX, Amer. Meteor. Soc., 1.5, https://ams.confex.com/ams/87ANNUAL/techprogram/paper_118544.htm.
- Spengler, T., C. Claud, and G. Heinemann, 2017: Polar Low Workshop summary. *Bull. Amer. Meteor. Soc.*, **98**, ES139–ES142, <https://doi.org/10.1175/BAMS-D-16-0207.1>.
- Sultana, K. R., S. R. Dehghani, K. Pope, and Y. S. Muzychka, 2018: A review of numerical modelling techniques for marine icing applications. *Cold Reg. Sci. Technol.*, **145**, 40–51, <https://doi.org/10.1016/j.coldregions.2017.08.007>.
- Verhoef, A., M. Portabella, and A. Stoffelen, 2012: High-resolution ASCAT scatterometer winds near the coast. *IEEE Trans. Geosci. Remote Sens.*, **50**, 2481–2487, <https://doi.org/10.1109/TGRS.2011.2175001>.
- Vionnet, V., I. Dombrowski-Etchevers, M. Lafaysse, L. Quéno, Y. Seity, and E. Bazile, 2016: Numerical weather forecasts at kilometer scale in the French Alps: Evaluation and application for snowpack modeling. *J. Hydrometeor.*, **17**, 2591–2614, <https://doi.org/10.1175/JHM-D-15-0241.1>.
- Walsh, K. J., M. Fiorino, C. W. Landsea, and K. L. McInnes, 2007: Objectively determined resolution-dependent threshold criteria for the detection of tropical cyclones in climate models and reanalyses. *J. Climate*, **20**, 2307–2314, <https://doi.org/10.1175/JCLI4074.1>.
- Warner, T. T., R. A. Peterson, and R. E. Treadon, 1997: A tutorial on lateral boundary conditions as a basic and potentially serious limitation to regional numerical weather prediction. *Bull. Amer. Meteor. Soc.*, **78**, 2599–2618, [https://doi.org/10.1175/1520-0477\(1997\)078<2599:ATOLBC>2.0.CO;2](https://doi.org/10.1175/1520-0477(1997)078<2599:ATOLBC>2.0.CO;2).
- Wieringa, J., 1996: Does representative wind information exist? *J. Wind Eng. Ind. Aerodyn.*, **65**, 1–12, [https://doi.org/10.1016/S0167-6105\(97\)00017-2](https://doi.org/10.1016/S0167-6105(97)00017-2).
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.
- WMO, 2017: Navigating weather, water, ice and climate information for safe polar mobilities. WWRP/PPP 5, 74 pp., https://epic.awi.de/id/eprint/46211/1/012_WWRP_PPP_No_5_2017_11_OCT.pdf.
- Wolff, M. A., K. Isaksen, A. Petersen-Øverleir, K. Ødemark, T. Reitan, and R. Brækkan, 2015: Derivation of a new continuous adjustment function for correcting wind-induced loss of solid precipitation: Results of a Norwegian field study. *Hydrol. Earth Syst. Sci.*, **19**, 951–967, <https://doi.org/10.5194/hess-19-951-2015>.
- Woollings, T., C. Franzke, D. L. R. Hodson, B. Dong, E. A. Barnes, C. C. Raible, and J. G. Pinto, 2015: Contrasting interannual and multidecadal NAO variability. *Climate Dyn.*, **45**, 539–556, <https://doi.org/10.1007/s00382-014-2237-y>.
- Yamagami, A., M. Matsueda, and H. L. Tanaka, 2018: Medium-range forecast skill for extraordinary Arctic cyclones in summer of 2008–2016. *Geophys. Res. Lett.*, **45**, 4429–4437, <https://doi.org/10.1029/2018GL077278>.
- Yang, X., and Coauthors, 2018: IGB, the upgrade to the joint operational HARMONIE by DMI and IMO in 2018. ALADIN-HIRLAM Newsletter, No 11, ALADIN Consortium, Brussels, Belgium, 93–96, <http://www.umr-cnrm.fr/aladin/IMG/pdf/nl11.pdf>.