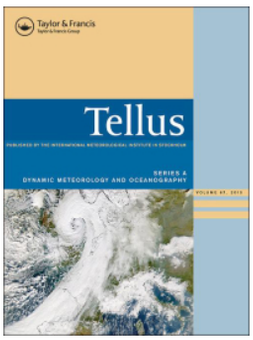


Taylor & Francis
Taylor & Francis Group

Tellus A: Dynamic Meteorology and Oceanography

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/zela20

On the configuration of a regional Arctic Numerical Weather Prediction system to maximize predictive capacity

Morten KØltzow, Rafael Grote & Andrew Singleton

To cite this article: Morten KØltzow, Rafael Grote & Andrew Singleton (2021) On the configuration of a regional Arctic Numerical Weather Prediction system to maximize predictive capacity, Tellus A: Dynamic Meteorology and Oceanography, 73:1, 1-18, DOI: 10.1080/16000870.2021.1976093

To link to this article: https://doi.org/10.1080/16000870.2021.1976093

Published online: 13 Sep 2021.

Tellus SERIES A DYNAMIC METEOROLOGY AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

On the configuration of a regional Arctic Numerical Weather Prediction system to maximize predictive capacity

By MORTEN KØLTZOW*, RAFAEL GROTE, and ANDREW SINGLETON, *Norwegian Meteorological Institute, Oslo, Norway*

(Manuscript Received 3 March 2021; in final form 26 August 2021)

ABSTRACT

Limitations to operational weather forecasts exist in terms of availability of computer (and human) resources combined with operational deadlines. For operational weather services it is therefore important to utilize their resources to maximize the predictive capability. This study shows how forecast quality in a state-of-the-art high-resolution regional Arctic Numerical Weather Prediction (NWP) system changes with varying configuration choices; (1) Ensemble Prediction System (EPS), (2) higher spatial resolution, (3) atmospheric initialization by assimilation of observations, (4) surface initialization by assimilation of observations and by (5) changing the regional domain and location. Results from such inter-comparisons are useful guidance for (Arctic) weather forecast systems, and can together with information on e.g. user-needs and post-processing capabilities be used to maximize the operational predictive capacity. All configuration choices have a significant impact on the forecast quality of near-surface parameters, but the impact varies with parameter, region, weather type, lead time and part of the forecast evaluated (e.g. average errors or rare events). Higher spatial resolution and EPS are expensive, but are still promising to further improve state-of-the-art regional Arctic high-resolution NWP systems. In particular when forecasting rare events regional EPS shows huge benefits. Assimilation of observations in the initialization process of the regional NWP system has also a positive impact on forecast quality. Finally, although less pronounced, the choice of the domain size and location also has a significant impact and should therefore be chosen carefully.

Keywords: Arctic regional Numerical Weather Prediction, operational forecast quality, Ensemble Prediction System, high-resolution, initialization

1. Introduction

It is anticipated that ship traffic, resource exploitation, tourism and other activities will increase in the Arctic over the coming years (WMO, 2017) and high quality weather forecasts are needed for safe operations (Jung et al., 2016). Arctic weather forecast capabilities have improved in recent decades (e.g. Jung and Leutbecher, 2007; Bauer et al., 2016), but Numerical Weather Prediction (NWP) systems still experience larger errors in the Arctic than at lower latitudes (e.g. Nordeng et al., 2007; Bauer et al., 2016; Gascard et al., 2017). It can be argued that the reduced Arctic forecast capabilities are due to the sparse conventional observations network, sub-optimal usage of remote sensing observations, the small spatial scales of many (high impact) Arctic-specific weather phenomena and that NWP systems are typically developed and tuned with a focus on mid and lower-latitude weather and less focus on e.g. the representation of cold surfaces.

The use of regional NWP systems can, compared with global systems, add value resulting from higher spatial resolution, optimized physics, and other configuration settings tailored to the targeted area. Such Added Value (AV) has been demonstrated in many studies including for Arctic short-range weather prediction (e.g. Müller et al., 2017; Yang et al., 2018; Køltzow et al., 2019). In both the Regional Climate Modelling community and in operational National Weather Services (NWSs) there is an awareness that forecast quality, i.e. the AV, from regional models depends on the configuration of the regional systems (e.g. domain size, location, initialisation, spatial resolution and deterministic vs. an ensemble

*Corresponding author. e-mail: famo@met.no

Citation: Tellus A: 2021, **73**, 1976093, https://doi.org/10.1080/16000870.2021.1976093

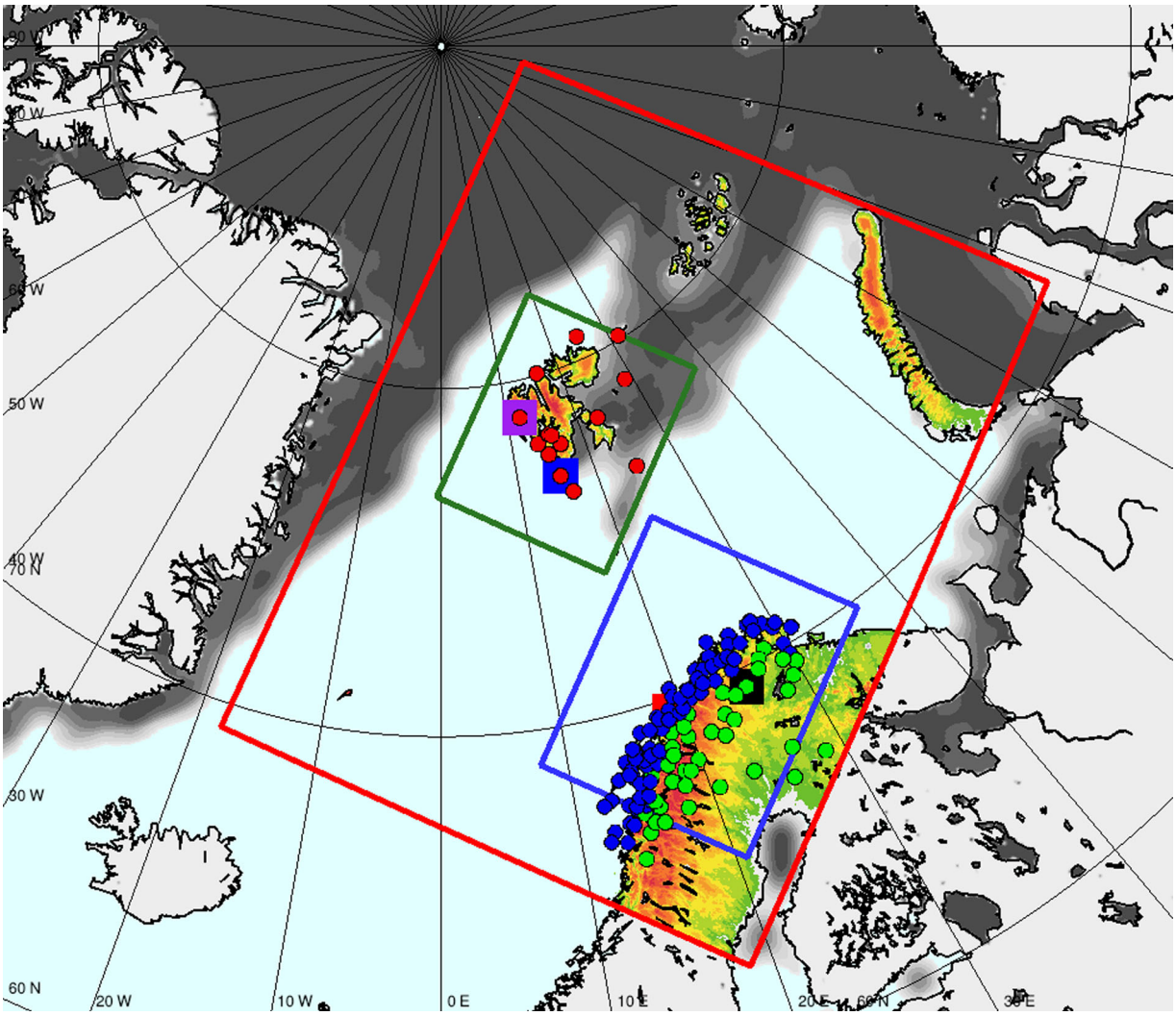


Fig. 1. Regional integration domains. The operational AROME-Arctic/CNTRL-experiment in red, the Svalbard domain in green and the North Norway domain in blue. The Observation sites used for evaluation; inland (green circles), coast and fjords (blue circles) and Svalbard stations (red circles). In addition, individual observation sites used for time series are marked with colored squares—Hornsund (blue), Ny-Ålesund (purple), Tromsø (red) and Karasjok (black). Orography in CNTRL-experiment shown in white/green (0–100 masl) to red colors (more than 1000 masl). Sea ice concentration (8 March 2018) from the operational IFS-HRES is shown in grey with darker shades indicating higher concentrations.

approach) in addition to the model code itself. Despite this, a pragmatic approach has often been chosen in the NWSs; a deterministic run for the area and lead times covering forecast obligations. Then the highest possible spatial resolution affordable can be found within the limitations of available computer power and operational deadlines. However, this approach is challenged with the advent of regional high-resolution EPSs in operational forecasting (e.g. Hagelin et al., 2017; Frogner et al., 2019a, 2019b). In regional high-resolution EPS studies, it has been shown that more members often are more beneficial than even higher spatial resolution (Hagelin et al., 2017; Raynaud and Bouttier, 2017). However, it is also well documented that increased spatial resolution in general improves forecast quality (e.g. Bauer et al., 2015) and it is easy to imagine that certain very local weather phenomena, e.g. connected with stationary forcings from complex topography or coast lines, will benefit more from higher spatial resolution than EPS for regional systems. It is well documented that in regional weather and climate model systems the AV of regional models is most pronounced in the presence of local small-scale forcing, but there is also a sensitivity to domain size, location and lateral boundary condition quality (e.g. Feser et al., 2011; Kristiansen et al., 2011; Rummukainen, 2016; Wang et al., 2016; Køltzow et al., 2019). For NWSs an important

Table 1. Summary of experiments.

EXP	UPPER-AIR initialisation	SURFACE initialisation	Spatial resolution	Domain (see Fig. 1)	EPS members	Cost
CNTRL	Blending	OI	2.5 km, 65 L	Large	1	1.0
HIGHRES	Blending	OI	**1.25 km, 90 L**	Large	1	∼11
EPS	**3D-Var**	OI	2.5 km, 65 L	Large	**1 + 6**	∼7
ATMASS	**3D-var**	OI	2.5 km, 65 L	Large	1	∼1.1
DD	Blending	**no**	2.5 km, 65 L	Large	1	∼0.925
SVA	Blending	OI	2.5 km, 65 L	**Svalbard**	1	∼0.1
NN	Blending	OI	2.5 km, 65 L	**N. Norway**	1	∼0.15

Configurations unique for the different experiments in bold. Experiments; CNTRL (control experiment), EPS (1 + 10 member), ATMASS (upper-air assimilation), DD (Dynamical Downscaling), HIGHRES (finer horizontal and vertical resolution), SVA (small domain around svalbard) and NN (small domain around Northern Norway). The cost column is relative to the CNTRL experiment (EPS is relative to 1 EPS member). The regional domains are shown in Fig. 1 and large refers to the operational AROME-Arctic domain.

issue is therefore whether their regional NWP system(s) should use (increased) operational computer power to e.g. increase spatial resolution, domain size, lead time and/or to include more ensemble members.

At MET Norway, operational forecasts for the Arctic are based on the regional AROME-Arctic model system (Müller et al., 2017), a version of the HARMONIE-AROME (HIRLAM–ALADIN Research on Mesoscale Operational NWP in Euromed–Application of Research to Operations at Mesoscale) configuration (Bengtsson et al., 2017) with 2.5 km horizontal grid spacing. This particular forecast region (Fig. 1) includes large, relatively homogeneous areas of open ocean, but also regions with complex topography, coastlines, fjords and moving sea ice (e.g. Svalbard and northern Norway) combined with more continental inland areas in northern Norway, Sweden and Finland. The domain is therefore characterized by a variety of different weather phenomena, e.g. cold air outbreaks with extensive convective activity and the formation of polar lows, wind channeling in narrow fjords, valleys and coast lines, periods with cold events (e.g. less than −20 °C) replaced with warm air intrusions with the risk of heavy rainfall events during winter to mention a few (e.g. Kolstad et al., 2009; Atlaskin and Vihma, 2012; Esau and Repina, 2012; Rojo et al., 2019). We can therefore imagine that both a further increase in spatial resolution and an introduction of a regional EPS can contribute to improved forecasts for this Arctic region.

It is also well documented that weather forecasts produced by the regional Arctic NWP system are sensitive to the initialization of the model atmosphere and surface (e.g. Køltzow et al., 2019; Randriamampianina et al., 2019, 2021). The forecast quality is substantially improved by assimilation of observations in this specific region compared to simply starting regional integrations from coarser spatial resolution global model analyses.

The aim of this work is to improve our understanding of how forecast quality varies with different configuration choices for the operational regional Arctic NWP system used at MET Norway. The study will compare forecasts from configurations of the AROME-Arctic model system using EPS, with higher spatial resolution, with and without upper-air and surface assimilation and using smaller integration domains. To our knowledge an inter-comparison of these different configurations done in a systematic way as presented here has not been done earlier for Arctic regional NWP systems. This work will therefore provide guidance on the optimal choice of configurations for (Arctic) regional NWP systems within the limitations of operational computer capacity.

In the following section, we describe the model system and the performed set of experiments, the observation used for verification and outline the inter-comparison strategy. In Sec. 3 the results are presented and discussed while Sec. 4 summarizes and concludes the work.

2. Experiments, observations and inter-comparison strategy

2.1. Model configuration and experiments

At MET Norway, short-range forecasts for the Arctic are based on the AROME-Arctic model system (Müller et al., 2017), which is a version of the HARMONIE-AROME (HIRLAM–ALADIN Research on Mesoscale Operational NWP in Euromed–Application of Research to Operations at Mesoscale) configuration (Bengtsson et al., 2017). All experiments presented here, use the same model description (dynamics and physics), cycling strategy (3 h initialisation), and apply 6–9 h old operational forecasts from the Integrated Forecasting System High Resolution (IFS HRES) at ECMWF as lateral boundary conditions (LBCs) in a similar way to the operational

version of AROME-Arctic at the time of the experiments. However, the experiments differ in their system configurations (details in Table 1) to make it possible to compare the impact of EPS (all EPS members vs. EPS control member); increased spatial resolution (HIGHRES vs. CNTRL experiment); assimilation in the atmosphere (ATMASS vs. CNTRL experiment); surface assimilation (CNTRL vs. DD); and changes in the location and size of the regional domains (SVA (Svalbard) and NN (North Norway) vs. CNTRL experiment).

The CNTRL experiment is similar to the operational AROME-Arctic at the time of the experiments with an exception for the initialisation of the atmosphere. Here, the CNTRL experiment takes the large scales from a 6 h old operational IFS HRES, while non-hydrostatic parameters and hydrometeors are taken from the previous model cycle (3 h earlier) in a blending process, i.e. no assimilation of observations in the atmosphere are done in the CNTRL experiment. This is done to avoid the requirement of generating a new B matrix needed for the assimilation process when making changes in spatial resolution or domains in other experiments. However, surface assimilation (temperature, humidity and snow) is done based on optimal interpolation of analyse-increments based on observed snow depth, and 2 m air temperature and relative humidity.

The HIGHRES experiment is identical to CNTRL except that the grid length has been reduced to 1.25 km and 90 vertical layers (2.5 km and 65 vertical layers in CNTRL, ~14 below 500 m), where the additional extra vertical layers are evenly distributed through the troposphere. No other changes have been done to adapt to the new grid spacing, e.g. physics parameterizations, with an exception for a shorter time step. The cost of HIGHRES in terms of computer resources is approximately 11 times that of CNTRL (4 times more grid cells × 2 because of a halfening of the time step × 90/65 due to more vertical layers).

In the EPS experiment, a 3D-Var upper air assimilation and a surface assimilation by optimal interpolation are applied to the control member. In addition, the EPS consists of 6 perturbed ensemble members generated by perturbing the initial conditions from the control member. Initial and lateral boundary condition perturbations are constructed by the "scaled lagged average forecasting" method (SLAF: Ebisuzaki and Kalnay, 1991; Hou et al., 2001) using the difference between two IFS-HRES forecasts initialized 6 h apart but valid at the same time. Different members are perturbed based on different lead times (e.g. difference between IFS-HRES +6 h and +12 h valid at the same time, and between +12 and +18 h, etc.). These perturbations are then scaled so all members on average have perturbations of approximately the same magnitude. Furthermore, a number of surface variables (including sea surface temperature, SST) are also perturbed by spatially correlated noise, with a specified standard deviation that depends on the parameter and a correlation length scale of 150 km. The configuration is similar to that which was used in the MetCoOp EPS (MEPS) at the time of the experiments, a shared operational high-resolution EPS for Scandinavia in cooperation between the Finnish, Swedish and Norwegian meteorological services (Frogner et al., 2019a, 2019b). To evaluate the impact of the full EPS, the forecast from the EPS control member is compared to the forecast from the full EPS (1 + 6 members). The cost of running the EPS is approximately 7 times the operational cost of AROME-Arctic.

The ATMASS runs differ from CNTRL by applying 3D-Var assimilation for the initialization of the atmosphere, i.e. the differences between ATMASS and CNTRL are due to the atmospheric assimilation process in the regional system. The Dynamical Downscaling (DD) experiment differs from CNTRL by not doing surface assimilation (i.e. differences between DD and CNTRL can be attributed to the surface initialization). In DD the soil and snow are initialized by interpolation from IFS-HRES which apply another surface scheme and resolution. The total cost of the atmospheric and surface assimilation is relatively modest with less than 10% reduction of the CNTRL. The two experiments Svalbard (SVA) and North Norway (NN) are similar to the CNTRL except that they only run on a subdomain of the operational AROME-Arctic domain (i.e. the differences are due to different domain size and location). In terms of computational costs SVA and NN are cheaper than CNTRL (only 10 and 15% of CNTRL, respectively). All the regional domains are shown in Fig. 1.

All experiments are done for a winter period (8–31 March 2018) which was chosen because it is a part of the Year of Polar Prediction (Jung et al., 2016) Special Observing Period 1. 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 during March. However, the pressure patterns were not consistent through the period and no particular temperature anomalies were present (Køltzow et al., 2019). Furthermore, the North Atlantic Oscillation, which is the dominant mode of variability in the region on synoptic time scales (Woollings et al., 2015) was not extreme in any way.

2.2. Observations

With end-user needs in mind we evaluate the different experiments with a focus on near-surface weather

parameters. We use high quality flagged observations from the quality controlled (Kielland, 2005) observation data base at MET Norway (frost.met.no) for 2 m air temperature (T2m), 10 m wind speed (WS10), 2 m Relative Humidity (RH2m) and one hour accumulated precipitation (precip). The precipitation observations have further been corrected for wind-induced undercatch, a substantial observation error for solid precipitation, by applying the universal transfer functions from Kochendorfer et al. (2017) following the approach of Køltzow et al. (2020). Only observation sites well inside each of the domains (>100 km away from the lateral boundaries) are used for verification. The observation sites used are shown in Fig. 1 and divided into the Svalbard region (14 observation sites), coast and fjord region (79 sites) and inland region (44 sites). These three regions experience different weather and hence also a possible difference in forecast accuracy. Not all sites observed precip, but most sites observed T2m, WS10 and RH2m.

2.3. Inter-comparison strategy

We evaluate the impact of the different configurational choices in three steps; (1) investigate the Potential Change in Forecast Quality (PCFQ), i.e. the differences between the experiments without being restricted by the availability of observations, and comparison with observations (2) by calculating the Mean Absolute Error (MAE) as a summary measure of the objective AV, and (3) by calculating the Brier Score (BS) to evaluate the forecast performance for tail (rare) events.

PCFQ follows the concept of Potential Added Value which has been used to investigate the usefulness of regional climate models compared to global climate models (e.g. Di Luca et al., 2012). However, in our study any differences between two forecasts can both lead to better or worse forecasts and we have therefore modified the naming. Here we define the PCFQ as any difference between two forecasts and use the Mean Absolute Difference (MAD) between two forecasts as an example. In addition, we use the ensemble spread as a measure of PCFQ in terms of predicting the forecast uncertainty. It should be noted that PCFQ is a necessary requirement, but not enough to claim objective AV comparing two forecasts. We limit our investigation of PCFQ to T2m and WS10. T2m is to a large degree constrained by the use of observed T2m in the surface assimilation and the fixed SST for each forecast. Opposite to this, observed WS10 is not assimilated and is more free to develop differences between two forecasts with different configurations.

To summarize the objective AV we compare experiments with its reference experiment by calculating the MAE Skill Score ($MAESS = 1 - MAE_{exp}/MAE_{ref}$), e.g. the impact of higher spatial resolution is given by $1 - MAE_{Highres}/MAE_{cntrl}$. Positive (negative) values show that the experiment performs better (worse) than the reference, i.e. the change in configuration adds (removes) value.

To also focus on high-impact weather we compare the ability of the different experiments to forecast cold/warm conditions, calm/windy conditions and dry/humid conditions defined by lower (higher) values than the observed 10%-tile (90%-tile) of T2m, WS10 and RH2m, respectively. For precip we compare the predictive capability for precipitation/no-precipitation decisions and precipitation above the 99%-tiles. Since these thresholds are defined by the observed percentiles in the investigated period, they are not necessarily high-impact events given the relatively short period and we therefore name them tail-events. To make it simple and possible to compare the different experiments we follow the approach of Frogner et al. (2019b), make probabilities for an event based on the number of members with exceedance of the given threshold in a grid box (i.e. probabilities from deterministic experiments are binary, either 0 or 100%) and calculate the BS for each experiment. Then the Brier Skill Score (BSS) is used to compare one experiment with its reference experiment similar to MAESS, e.g. the impact of higher spatial resolution on tail events is given by $(1 - BSS_{Highres}/BSS_{cntrl})$.

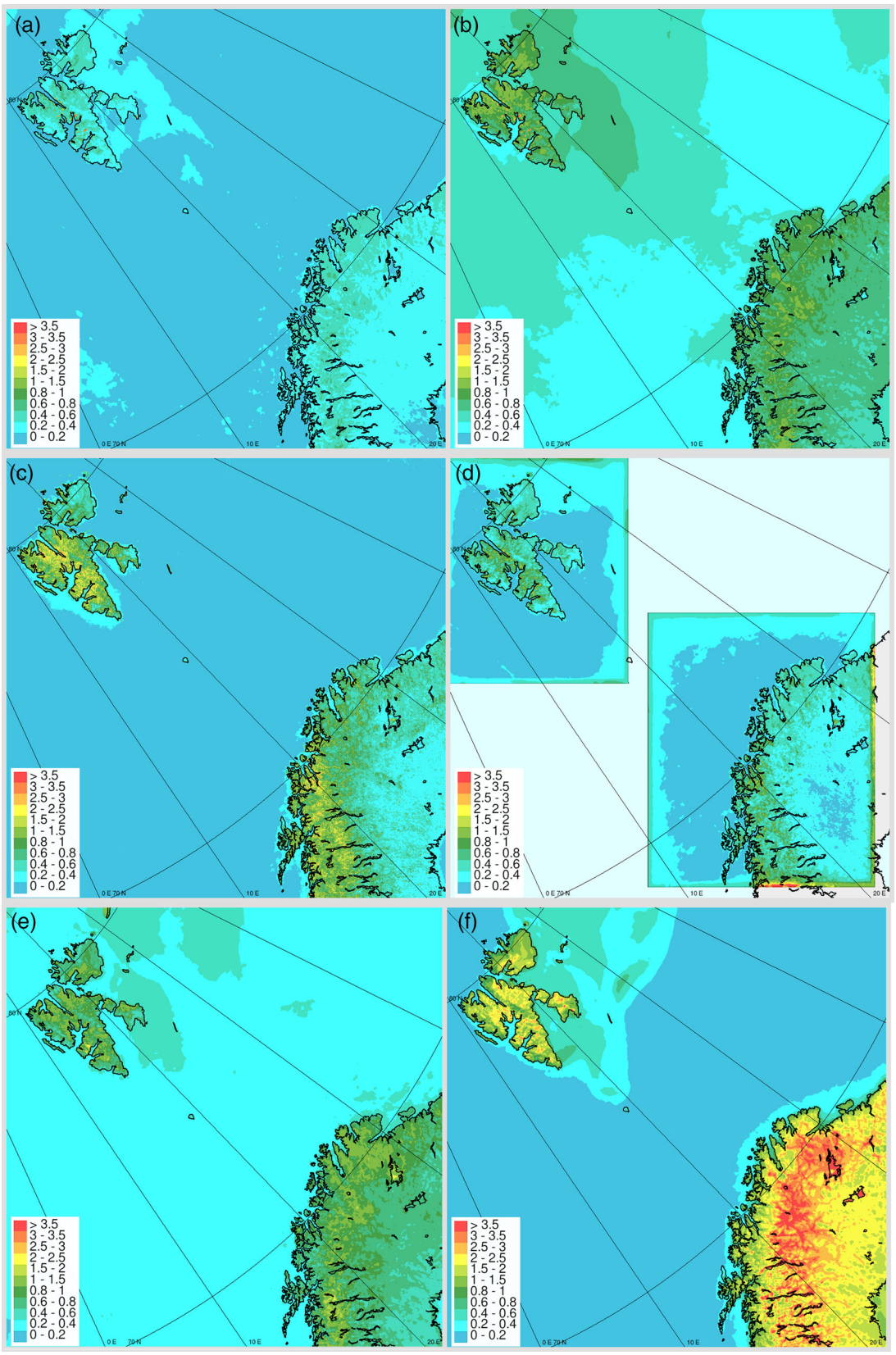
MAESS and BSS are not necessarily the best metrics to highlight impact for all parameters and configuration choices, but allow a useful inter-comparison between different configurations and parameters. By this they contribute to a useful first step to understand the impact of differences in configuration.

3. Results

In the following we use the inter-comparison strategy outlined in the previous section to identify the impact of different configuration choices. First we present examples of PCFQ focusing on T2m and WS10, then we discuss the objective AV for multiple near-surface weather parameters by comparing with observations, before we investigate the impact on forecasting observed tail events. Finally we include some additional evaluations to provide more robust results.

3.1. Potential change in forecast quality

Ensemble forecast output differs in many ways from a single deterministic forecast. Here we calculate the Mean Absolute Difference (MAD) between the mean of all members (EPS mean) and the non-perturbed control run of the EPS for T2m (Fig. 2a) and WS10 (Fig. 3a). In addition, we use the EPS spread of T2m and WS10 as


(a)
(b)
(c)
(d)
(e)
(f)
> 3.5
3 - 3.5
2.5 - 3
2 - 2.5
1.5 - 2
1 - 1.5
0.8 - 1
0.6 - 0.8
0.4 - 0.6
0.2 - 0.4
0 - 0.2
0 E 70 N
10 E
20 E

PCFQ in predicting the uncertainty of the forecasts (Figs. 2b and 3b, respectively). T2m shows higher PCFQ over land where each ensemble member has its own evolving surface during the forecast, than over the ocean where a fixed SST is used, which include relatively modest perturbations for the ensemble members (between −0.5 and 0.5°). Over the ocean the T2m spread shows a maximum east of Svalbard connected to situations with cold air flowing off the sea ice further north (see location of the sea ice in the beginning of the period in Fig. 1). However, the T2m PCFQ is largest in the observation sparse regions at Svalbard and higher elevated inland regions in Scandinavia, than in other parts of inland Scandinavia. Compared to the other configurations the impact of ensemble mean versus the control run is modest, and it is reasonable to assume that the main added-value from EPS is associated with the ensemble spread. For WS10 a modest local maxima in PCFQ is found in regions with higher elevations and complex topography as for T2m. However, opposite to T2m, the largest PCFQ for WS10 is present over the ocean and in particular along the Norwegain coast and in the southern part of the domain.

The T2m PCFQ of higher spatial resolution (MAD of HIGHRES against CNTRL, Fig. 2c) shows PCFQ larger than that of the EPS mean with a maximum in complex terrain, e.g. mountain and coastal areas of Norway and Svalbard. Less PCFQ is seen over the relatively flat and more homogeneous terrain of Sweden and Finland east of the Norwegian/Swedish mountains. Over the ocean only negligible differences are seen. Zooming in on individual mountains and fjords reveals that a large difference in PCFQ can be found between neighbouring grid cells (not shown) indicating that a part of the PCFQ is driven by local changes in the surface characteristics of the individual grid cell, e.g. changes in orography or land-sea-mask. Similar spatial patterns are also seen for the PCFQ of WS10 and it should be noted that for WS10 higher spatial resolution shows the most pronounced PCFQ over land measured by MAD for all experiments (Fig. 3).

The T2m PCFQ of applying small domains (MAD of SVA/NN against CNTRL, Fig. 2d) is small over the ocean, but noticeable over land. However, it is less than for HIGHRES, even if the spatial patterns have similarities. The PCFQ increases towards the interior of Svalbard and is higher at elevated areas in Norway and Sweden, and lower in the more flat homogeneous terrain in Sweden/Finland. The spatial patterns for WS10 PCFQ (Fig. 3d) are very similar to T2m. For small domains one can argue that the PCFQ should decrease in the interior of the domains, because the air masses would have had more time to adjust towards the climatology of the regional model. However, here the fixed SST constrains the T2m and the difference in wind is not very large over the flat ocean between the driving model and the regional model.

The T2m PCFQ of upper-air assimilation (MAD of ATMASS against CNTRL, Fig. 2e), is more pronounced than the PCFQ for the small domains and EPS mean, but less pronounced than for HIGHRES. However, the spatial patterns show similarities with highest PCFQ over Svalbard and elevated areas in general. Out of all of the configuration choices tested, the upper-air assimilation is the one who has the largest T2m PCFQ over the ocean. This seems reasonable since ATMASS is the only experiment that has the potential to, in a direct way, substantially change the atmosphere over the ocean, while the higher elevated areas are closer to and more affected by the free atmosphere than lower elevations which (during certain weather regimes) have less exchange with the large-scale air flow. For WS10, a local maxima in PCFQ is also found over the mountains on the border between Norway and Sweden. However, the largest PCFQ is seen at the Norwegian coast and the southwest coast of Svalbard.

Surface assimilation (MAD of CNTRL vs. DD, Fig. 2f) shows the largest T2m PCFQ of all the evaluated configuration choices. The maximum PCFQ is found inland close to the Norwegian/Swedish border, which during winter is an area dominated by stably stratified atmospheric conditions and cold temperatures known to be notoriously difficult to simulate properly in NWP systems (e.g. Atlaskin and Vihma, 2012). However, it is also an area with a relatively dense network of conventional observations where substantial adjustments during the assimilation process can be expected. Zooming in on the PCFQ reveals that the patterns of PCFQ are most pronounced in the lower part of the topography, probably because the location of most observation sites are in the valleys (not shown). However, opposite to this the largest PCFQ at Svalbard is found at higher elevations which is counter-intuitive, since most stations are located along the coastline in this area. This can therefore not be explained by assimilation increments. However, in DD a

Fig. 2. Potential Change in Forecast Quality measured by Mean Absolute Difference in T2m [°C] averaged over day 1 forecasts 8–31 March 2018 for (a) EPS: EPS mean vs. EPS control member, (c) higher spatial resolution: HIGHRES vs. CNTRL, (d) domain size/location: SVA/NN domains vs. CNTRL, (e) initialization atmosphere: ATMASS vs. CNTRL and (f) Initialization surface: CNTRL vs. DD. In addition, (b) predicting the forecast uncertainty: the average ensemble spread [°C].

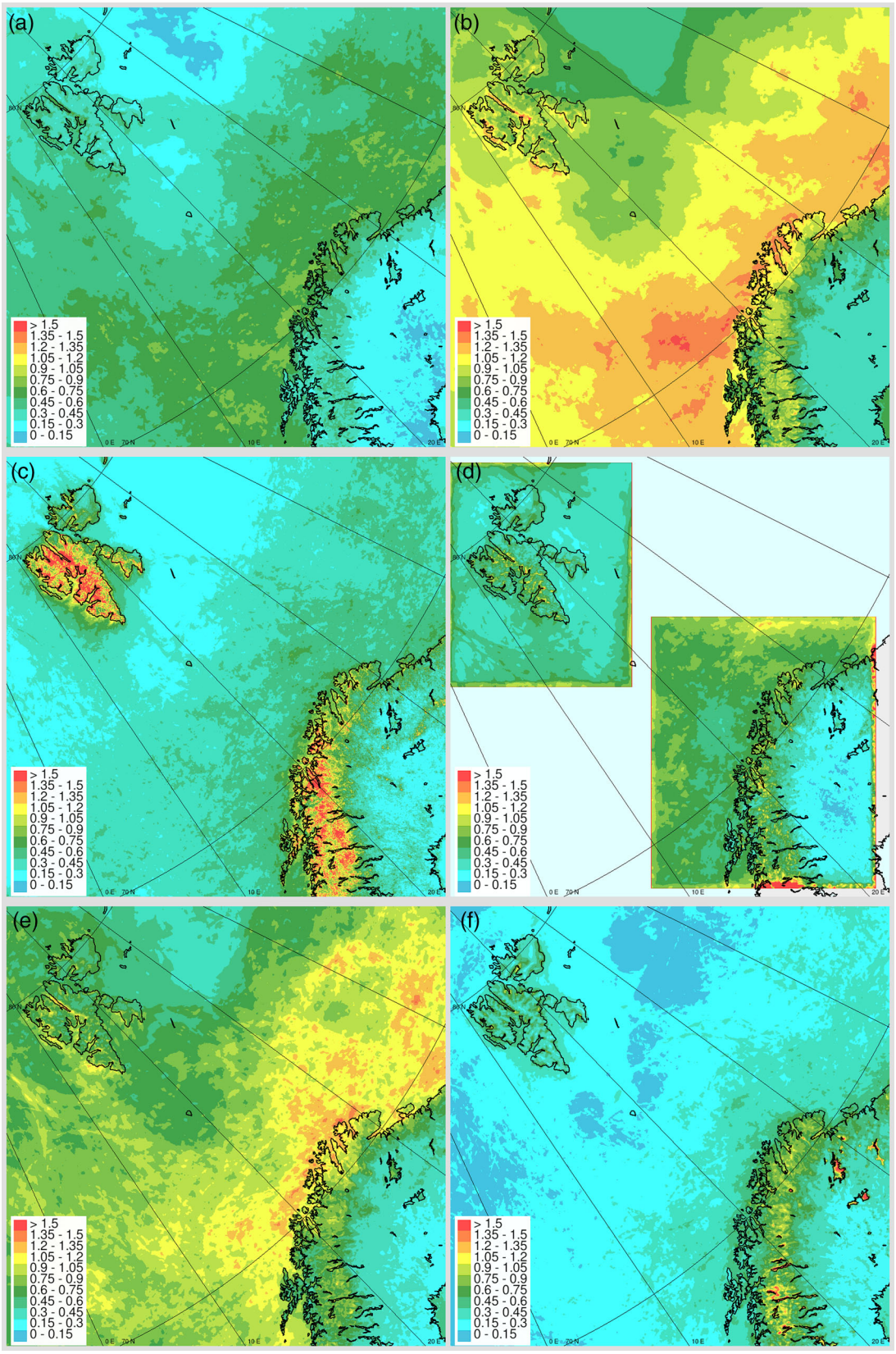


Fig. 3. As Fig. 2, but for 10 m wind speed, WS10 [m/s].

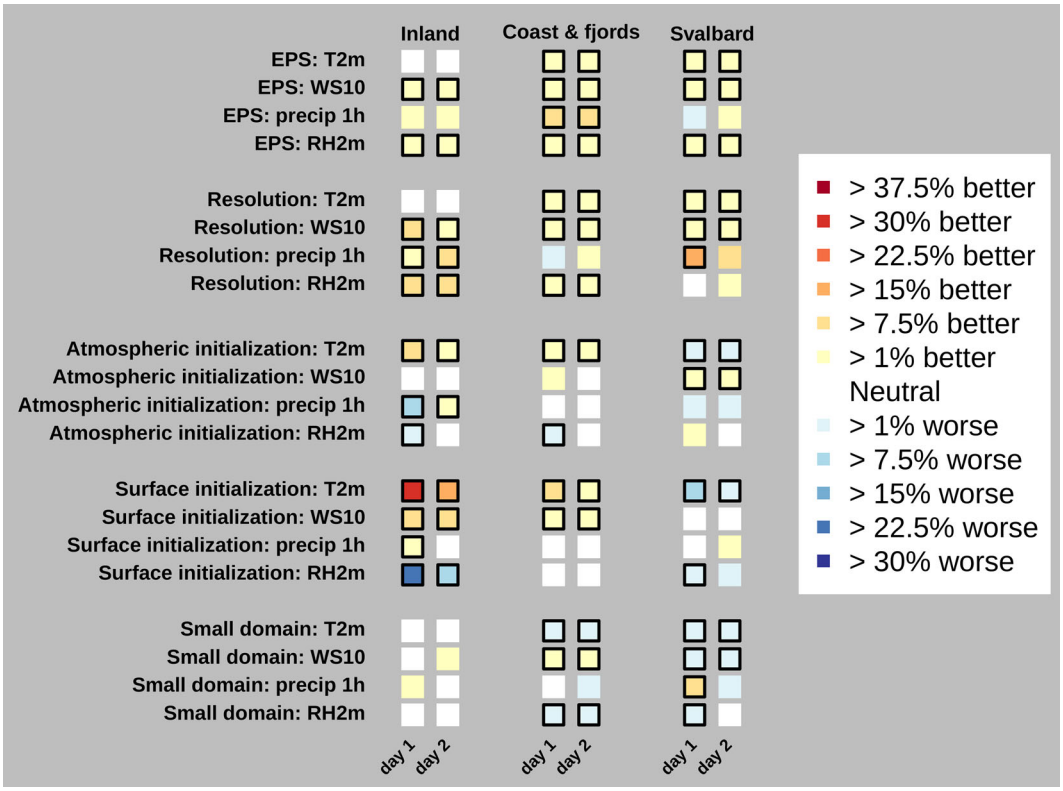


Fig. 4. Changes in forecast quality by **EPS** (EPS mean vs. EPS control member), higher spatial **Resolution** (HIGHRES vs. CNTRL), **atmospheric initialization** (ATMASS vs. CNTRL), **surface initialization** (CNTRL vs. DD) and using **small domains** (SVA/NN vs. CNTRL) measured by mean absolute error skill score for day 1 and day 2 forecasts. Parameters are T2m, WS10, precip and RH2m. Black frames indicate a significant difference in the verification score at the 95%-level calculated by bootstrapping.

restart from IFS-HRES fields is done for every forecast cycle which is not necessarily optimal due to differences in spatial resolution and in particular in this case since the regional and the global driving model have a different representation of the surface. Opposite to this, in CNTRL the first guess is taken from a previous cycle of AROME-Arctic, i.e. in observation sparse areas a recycling of the regional system is done between each new forecast. The PCFQ at the higher elevations in the interior of Svalbard is therefore also a measure of the difference between the regional high-resolution model and restarts from the global coarser spatial resolution IFS-HRES, which also explain parts of the pronounced differences in the inland regions of Norway and Sweden. The WS10 PCFQ (Fig. 3f) is small compared to T2m and in particular at Svalbard, but shows some of the same spatial patterns. Less PCFQ on WS10 by surface assimilation is expected since wind observations are not used in the assimilation process.

In all experiments some of the most pronounced areas with large PCFQ for T2m and WS10 are in regions with less observational coverage (e.g. interior of Svalbard, mountainous regions between Norway and Sweden). Since all experiments include surface assimilation (except DD) this can, in particular for T2m, at least partly be due to the fact that all experiments are constrained by the surface assimilation of temperature which reduces the PCFQ in observation-dense areas. On the other hand, this also presents a challenge to the objective evaluation against observations in the next section, in some of the regions with the largest differences we lack observations for a proper evaluation.

3.2. *Objective added value for near surface variables*

In this section we compare the experiments with observations to investigate the objective AV of T2m, WS10, RH2m and precip. Even if the coverage of observations are relatively good (Fig. 1) there are areas where the PCFQ for T2m and WS10 in the previous section was large, but few observations are available (e.g. interior of Svalbard and mountainous regions on the border between Norway and Sweden), making it difficult to provide robust estimates of the true AV in these regions. Figure 4 summarizes the AV in terms of improved MAESS for

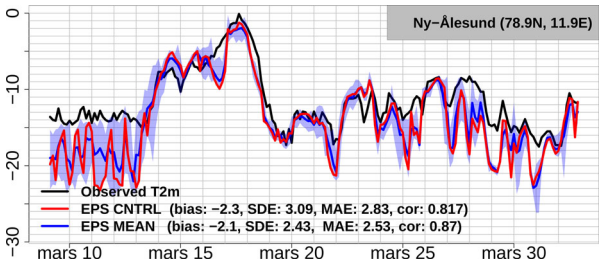


Fig. 5. Time series of T2m at Ny-Ålesund, observations (black), ensemble mean (blue), ensemble maximum/minimum (blue shading) and the non-perturbed control member (red). Forecasts are initialized at 00 UTC and lead times +27, +30, +33, +36, +39, +42, +45 and +48 h are used.

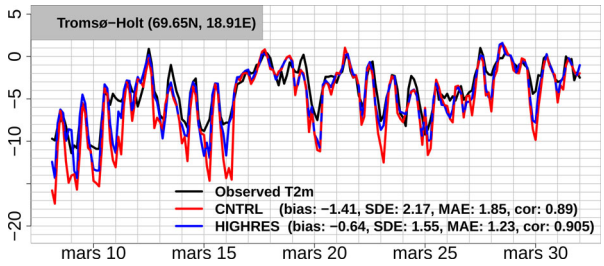


Fig. 6. Time series of T2m at Tromsø-Holt, observations (black), HIGHRES (blue) and CNTRL (red). Forecasts are initialized at 00 UTC and lead times +3, +6, +9, +12, +15, +18, +21 and +24 h are used.

T2m, WS10, precip and RH2m. Only changes that are significant (indicated by black frame) will be discussed.

The introduction of an ensemble (EPS-experiment), shows an improvement in MAE by using the EPS mean. For temperature, a modest improvement (1–7.5%) for Svalbard, coast and fjords is seen for day 1 and day 2 forecasts. The impact on WS10 is similar to temperature, but also present inland. For precip, EPS gives an improvement for coast and fjords (7.5–15%, day 1 and day 2). A consistent improvement for EPS is found for RH2m with 1–7.5% in all regions for both day 1 and day 2 forecasts. To further assess the impact of EPS on T2m forecasts we inspect the time series from Ny-Ålesund (Fig. 5), a location where we find pronounced differences between the ensemble and the control member. For a number of days the control forecast is too cold, with clear day-to-day variations in the temperature, and being colder than the observations (e.g. 10–13, 17, 22, 25 March). Also the EPS members have a cold bias for these days, but the EPS mean is more consistent and the uncertainty is instead reflected in the ensemble spread. These drops in forecasted temperature can be related to a model specific deficiency which occasionally is seen in the operational forecasts for Svalbard and include an unrealistic sudden drop in temperature during cold and calm conditions (Valkonen et al., 2020) However, the substantial temperature drop seen on 18 March in the forecasts is accompanied by little ensemble spread and is also observed, but with a slightly shifted timing. The results presented here confirm that the EPS mean is an efficient way to improve some verification scores since unpredictable small scales are filtered out (e.g. Owens and Hewson, 2018)

The higher spatial resolution, HIGHRES, also shows an improvement in MAE (Fig. 4) for all parameters. A modest improvement (1–7.5%) is found for T2m at the coast, fjords and Svalbard. As seen for PCFQ, also the T2m objective AV by HIGHRES varies spatially. At Tromsø-Holt (located at the island of Tromsø, see Fig. 1) HIGHRES improve the forecast by less underestimation of cold temperatures due to a better representation of the coast line (Fig. 6). However, at two other observation sites at the same island, located only a few kilometres away, the impact is neutral (not shown). A modest improvement (1–7.5%) is also found for WS10 in all regions, with the exception of an even larger improvement (7.5–15%) for day 1 inland. A substantial part of the latter is due to a reduced positive bias. Many observation sites are located in relatively sheltered areas, e.g. in valleys, resulting in an overestimation of WS10 in CNTRL while several of these locations (valleys) are better represented in HIGHRES which forecasts less windy conditions (not shown). Improvements are also found for precip, but varying in size and significance with lead time and region. HIGHRES also improve forecasts of RH2m inland (7.5–15%) and in coast and fjords (1–7.5%). The changes in RH2m are not following the same patterns as T2m, which imply that the origin of the improvements must also come from the specific humidity and not only changes in T2m. A general improvement by applying higher spatial resolution is not surprising and agrees with other related studies in the same areas (e.g. Køltzow et al., 2019; Valkonen et al., 2020).

The impact of assimilation of upper-air observations, ATMASS, has a clear positive impact on MAE for T2m inland, coast and fjords (1–7.5%, and 7.5–15% day 1 inland). However a modest degradation (1–7.5%) in forecast quality is found over Svalbard. For the observation site Tromsø (Fig. 7, located on the same island as Tromsø-Holt discussed above for HIGHRES, but at a higher elevation) the upper-air assimilation has a positive impact by reducing the cold bias which originates from specific days (e.g. 8–10, 15 March). For WS10, a modest improvement in wind over Svalbard is seen (1–7.5%) similar to what was reported by Kim et al. (2019). For precip the impact is small or modest and mixed. Inland, an improvement (deterioration) in precip is seen for day 2

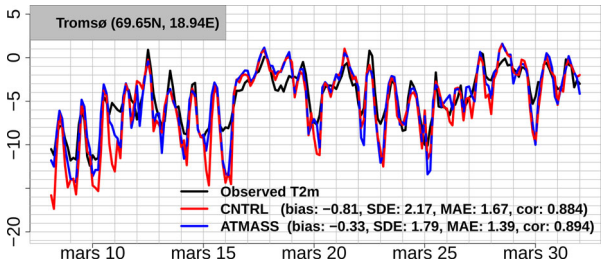


Fig. 7. Time series of T2m at Tromsø (close to Tromsø-Holt in Fig. 6), observations (black), ATMASS (blue) and CNTRL (red). Forecasts are initialized at 00 UTC and lead times +3, +6, +9, +12, +15, +18, +21 and +24 h are used.

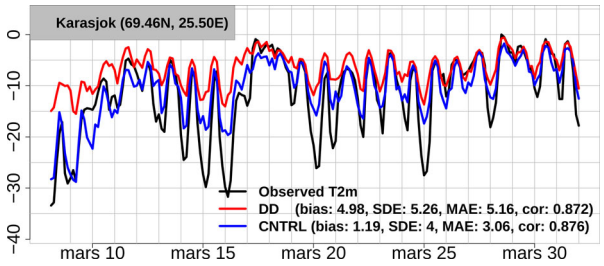


Fig. 8. Time series of T2m at Karasjok, observations (black), CNTRL (blue) and DD (red). Forecasts are initialized at 00 UTC and lead times +3, +6, +9, +12, +15, +18, +21 and +24 h are used.

(day 1) (1–7.5%). Mixed results for precip (as also seen for HIGHRES), are expected since robust results are hampered by the relatively short and dry period, fewer precipitation observations, and because the verification of solid precipitation is associated with large uncertainties (Køltzow et al., 2020). For RH2m a modest reduction (1–7.5%) in forecast skill is found for day 1 for inland, coast and fjords. The results presented here are in line with aggregated results for the entire domain from Randriamampianina et al. (2021), who also explain the deterioration in RH2m quality by an unbalanced adjustment of temperature and specific humidity in the assimilation process.

The impact of surface assimilation, CNTRL-experiment compared with DD, gives a substantial and positive impact on MAE for T2m for inland, coast and fjords. In particular inland the impact is large, but decreases slightly with lead time (30–37.5% and 15–22.5%, for day 1 and 2, respectively), while it is more moderate in coast and fjords (7.5–15% and 1–7.5% for day 1 and day 2, respectively). Figure 8 shows the time series of observations and forecasts at Karasjok, an inland site (see location in Fig. 1) located in the area of maximum PCFQ (Fig. 2f), with frequently observed temperatures below −20 °C during the period. The surface assimilation substantially improves the downscaled forecast, but is not able to fully follow the observed coldest temperatures. However, at Svalbard the surface assimilation reduces the forecast quality for T2m. To understand this requires more investigation, but we notice that IFS HRES which provides the initial conditions in DD verifies similar to AROME-Arctic on Svalbard, but substantial worse for inland, coast and fjords for this period (Fig. 3 in Køltzow et al., 2019). An improvement in MAE is also found for WS10 with 7.5–15% inland and 1–7.5% for coast and fjords. The improvement for precip is less pronounced, but present for day 1 inland (1–7.5%). The findings on improved forecasts by surface assimilation are in good agreement with earlier studies (e.g. Giard and Bazile, 2000). However, a substantial decrease in quality for RH2m is found inland (>30% and 7.5–15%, day 1 and 2, respectively) and for day 1 Svalbard (1–7.5%). The surface assimilation is done by applying T2m and RH2m increments to the surface properties (first guess minus observations), but when snow is present only the temperature is adjusted since the soil moisture is assumed to be uncoupled from the atmosphere. With only the temperature being adjusted an imbalance between temperature and humidity occurs which may deteriorate the RH2m. As discussed for the PCFQ also the cycling of the surface in CNTRL versus the re-start with IFS-HRES for each new forecast in DD will contribute to the differences, i.e. the deterioration in RH2m might also partly be due to differences in how well the humidity is represented in the global and regional model systems. It should be underlined, that as discussed in the previous section, the difference between CNTRL and DD is not only due to assimilation, but also due to different cycling strategies for the background or first guess surface fields in the experiments.

The use of small domains (SV, NN) shows no impact on MAE inland for any of the parameters. However, a moderate increase (1–7.5%) in T2m errors for coast, fjords and Svalbard is noticed. A time series from Hornsund at Svalbard (Fig. 9) shows how the small domain in given periods (e.g. 8–13, 18–21 March) has a larger cold bias than the large domain. A similar impact pattern as for T2m is seen for RH2m with the exception of a neutral impact day 2 Svalbard. The results for WS10 are mixed, a modest improvement is found in coast and fjords, while the opposite is true for Svalbard (both changes 1–7.5%). For precip the results are neutral except for improved (7.5–15%) MAE for day 1 Svalbard, but as discussed above the results for precip are less robust than for other parameters. In general, the results are consistent with previous findings that specific choices of location and size of the domain may have an impact on forecast quality (e.g. Kristiansen et al., 2011; Giorgi, 2019).

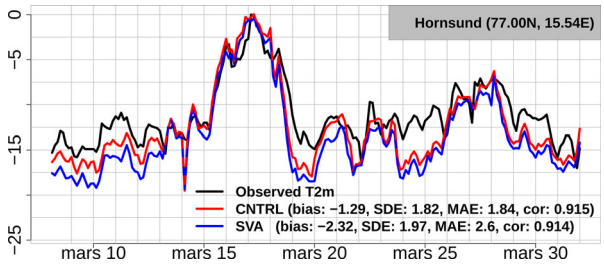


Fig. 9. Time series of T2m at Hornsund, observations (black), small domain (blue) and CNTRL (red). Forecasts are initialized at 00 UTC and lead times +3, +6, +9, +12, +15, +18, +21 and +24 h are used.

3.3. Impact on tail events

Only configuration changes with a significant impact on BSS (marked with a black frame in Fig. 10) in forecasting tail events will be discussed in the following. The absolute values of the thresholds used to define tail events are shown in Table 2 and correspond to the observed 10 and 90%-tile for T2m, WS10 and RH2m. For precip, precipitation/no-precipitation is chosen for the low threshold while the 99%-tile is used for the high threshold.

EPS shows a large positive impact for all parameters and regions and a substantially higher AV in forecasting tail events than the AV measured by MAE. The largest improvements are found for windy conditions (high WS10) inland, coast and fjords, precipitation/no-precipitation inland, coast and fjords, for high precipitation coast and fjords (high precip 1 h) and for dry (low RH2m) conditions in coast and fjords (all above 22.5%). These findings agree well with studies showing the benefits of EPS which enables forecasting small scales and rare events due to the spread between EPS members (e.g. Frogner et al., 2019b; Jung and Leutbecher, 2008).

Finer grid spacing (HIGHRES) has a more modest, but somewhat mixed impact. Warm conditions (high T2m) are improved inland (7.5–15%) and at the coast and fjords (1–7.5%). Also windy conditions are improved inland, coast and fjords (1–7.5%), while the forecasts of calm conditions (low WS10) shows a deterioration inland (15–22.5%). Forecasting high precipitation amounts show a substantial improvement at Svalbard (30–37.5%), while for precipitation/no-precipitation the results are mixed with a modest improvement (1–7.5%) inland, and a modest deterioration in coast and fjords (1–7.5%). Mixed results are also found for RH2m with an improvement in humid conditions inland (15–22.5%) and a deterioration for dry conditions in coast and fjords (7.5–15%) and Svalbard (15–22.5%).

The impact of ATMASS has a modest (1–7.5%) positive impact on both warm and cold conditions inland, and cold conditions in coast and fjords, but a reduction in forecast quality of cold temperatures at Svalbard. ATMASS also improves the representation of calm conditions at the coast and fjords and windy conditions at Svalbard (1–7.5%). For precipitation/no-precipitation and high precipitation amounts, the results are neutral except for a reduction in quality (1–7.5%) at Svalbard. For RH2m a modest deterioration (1–7.5%) is seen for humid conditions (high RH2m) inland and in coast and fjords, while a similar improvement is found for dry conditions (low RH2m) at Svalbard.

The impact of surface assimilation varies with parameters and regions. An improved representation of cold temperatures, as discussed in Sec. 3.2, are seen inland, coast and fjords (7.5–15%), while a deterioration is found for Svalbard (15–22.5%). Also for the warmest conditions inland a deterioration is found (7.5–15%). The latter is probably due to a general colder forecast, reducing the ability to forecast the sharp peak of maximum temperature during day time (e.g. Fig. 8). However, it should be noticed that the warm conditions in coast and fjords and at Svalbard are improved (1–7.5% and 7.5–15%, respectively). For WS10 the impact is mixed, with improved (deteriorated) windy (calm) conditions inland, but neutral elsewhere. Precipitation/no-precipitation inland is improved, but the impact for precipitation is elsewhere neutral. The impact on dry and humid conditions is in general negative and connected to the unbalanced assimilation of temperature and humidity discussed above.

The impact of using small domains (SV, NN) on tail events are mixed, but skewed towards a negative impact. Improvements are found for cold conditions inland (1–7.5%) and warm conditions at Svalbard (7.5–15%), while the opposite is seen for cold conditions in coast and fjords (1–7.5%) and Svalbard (7.5–15%). The windy conditions at the coast and fjords are improved (7.5–15%), while both windy and calm conditions at Svalbard are deteriorated (both 1–7.5%). No significant changes in the ability to forecast precipitation/no precipitation and high precipitation amounts are found. The ability to forecast humid conditions in coast and fjords (7.5–15%) and at Svalbard (22.5–30%), and dry conditions (1–7.5%) in coast and fjords are all worse with the small domains.

3.4. Additional evaluations

The resolution and EPS experiments were, in addition to the winter period, carried out for a summer period (10 July–1 August 2018). A summary of the results in terms of change in MAE (as presented in Fig. 4) and for tail events (as presented in Fig. 10) are shown in Fig. 11. The general picture is similar to the winter period with an overall improvement seen in MAE by EPS, while the results for the resolution experiment are more mixed, but

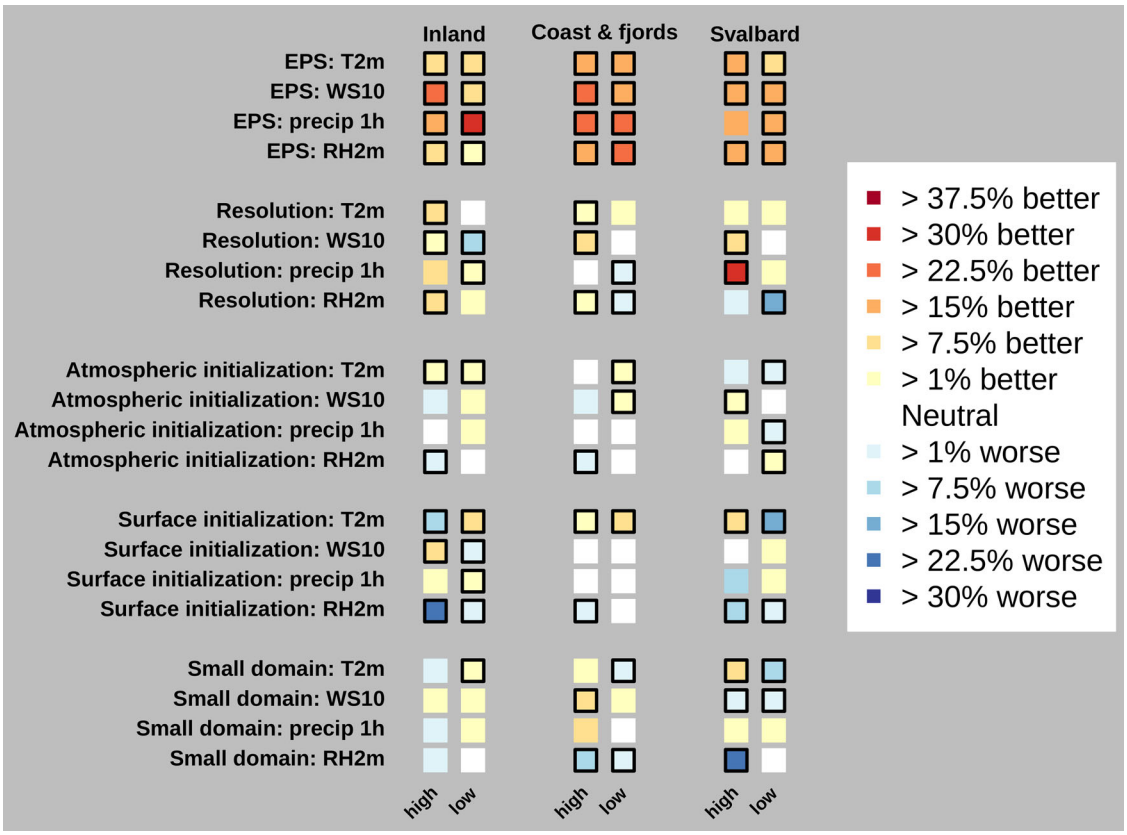


Fig. 10. Changes in forecast quality of tail events by **EPS** (EPS mean vs. EPS control member), higher spatial **Resolution** (HIGHRES vs. CNTRL), **atmospheric initialization** (ATMASS vs. CNTRL), **surface initialization** (CNTRL vs. DD) and using **small domains** (SVA/NN vs. CNTRL) measured by Brier Skill Score for observed low and high tail events. Parameters are T2m, WS10, precip and RH2m. Black frames indicate a significant difference in the verification score at the 95%-level calculated by bootstrapping.

Table 2. Absolute values of thresholds used to define observed tail events during the period 8–31.March 2018.

	Inland		Coast and fjords		Svalbard	
	Low %-til	High %-tile	Low %-tile	High %-tile	Low %-tile	high %-tile
T2m	≤−19.4 °C	≥−2.2 °C	≤−8.6 °C	≥0.8 °C	≤−18.0 °C	≥−7.0 °C
WS10	≤0.5 m/s	≥5.7 m/s	≤1.4 m/s	≥10.3 m/s	≤1.3 m/s	≥11.8 m/s
RH2m	≤59%	≥89%	≤56%	≥91%	≤54%	≥86%
precip	≥0.1 mm/h	≥1.2 mm/h	≥0.1 mm/h	≥3.0 mm/h	≥0.1 mm/h	≥0.5 mm/h

more positive than negative. However, compared to the winter period the quantitative AV for different regions and parameters are different. Also for the tail events in summer the results are qualitatively similar for EPS as in the winter period, but with a tendency towards higher added value. However, opposite to this, the added value by resolution appears to be lower in summer than winter. Possible explanations can be that more systematic errors are present in winter, for which increased resolution is beneficial. In summer, more small-scale features may lead to more unsystematic errors and double penalty issues in point verification, for which the EPS adds more value. Also tuning towards finding the optimal configurations for finer resolution and EPS can play a role.

To confirm the robustness of the EPS results for tail events we evaluate one year (2018) of forecasts from the operational MetCoOp EPS (Frogner et al., 2019a, 2019b). This regional system used in 2018 the same model version, resolution and perturbations techniques as used in the EPS experiments discussed above, and for a partly overlapping domain (Inland, coast and fjord sites). This allows us to evaluate for a much longer period, for different ensemble sizes (4, 7 and 10 members), and to include a comparison with a simple lagged 4-member ensemble

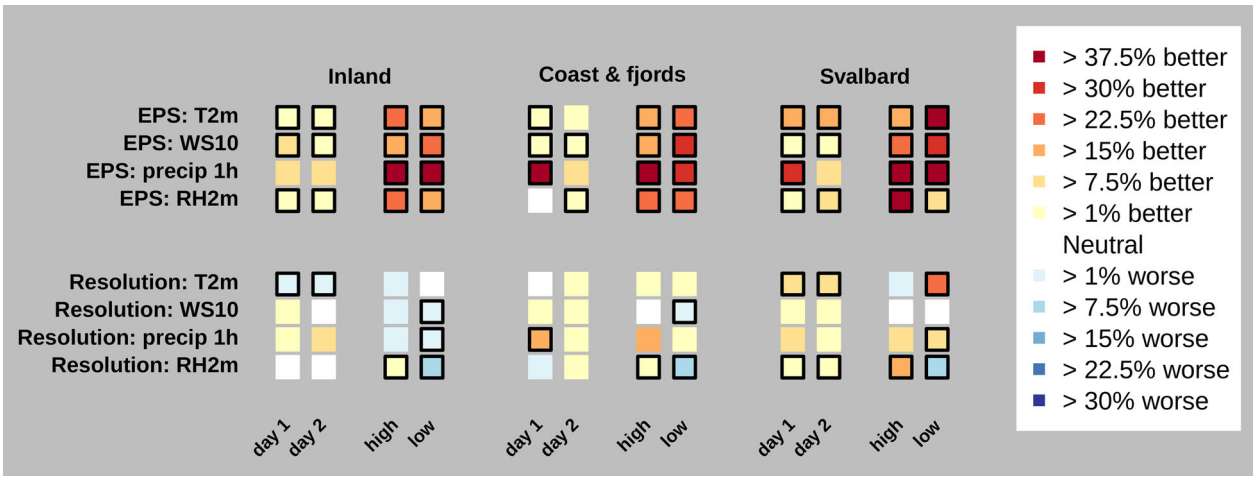


Fig. 11. Summer period (10 July–1 August, 2018), change in MAE for day 1 and day 2 forecasts and Brier Score for high and low tail events. Similar to what is shown in Figs. 4 and 10, respectively.

(last four deterministic runs; initialized 0, 6, 12 and 18 hours earlier). For each month in 2018 thresholds for tail events are found with the %-tiles given in Table 1, and the added value by the different EPS configurations are calculated similarly as for Fig. 10. The added value is then averaged over all months and presented in Fig. 12. The highest added value is found for precip, but there is a substantial added value for all parameters and all EPS configurations (lagged, 4, 7 and 10 members) tested. In general the lagged EPS includes substantial added value and has only slightly less AV compared to an ordinary 4 member ensemble. The good performance of the lagged ensemble is at least partly due to the fact that the perturbed members in a deterministic verification are approximately similar in quality as a 12–24 hour old non-perturbed forecast (depending on parameter, metric, etc., not shown). This shows that a non-neglecting part of the added value provided by EPS is already available from the deterministic configuration and is in agreement with earlier studies (Bouallégue et al., 2013; Scheufele et al., 2014; Osinski and Bouttier, 2018). This also highlights that when studying the added value by an EPS system, a more correct picture is given when the EPS is compared with a simple lagged system constructed with already available information, and not only with the last deterministic run.

The AV increases with the increasing number of members, but the change in going from 7 to 10 members is modest in this particular set-up. The main features are similar for inland stations and for coast and fjord stations, but some differences are seen. For example, the AV for precipitation is larger inland than at the coast and fjords. A reasonable explanation for this is that the precipitation at the coast and fjord region is more steered by the complex terrain and coastline. Furthermore, the AV at coast and fjords is higher for low WS10 and high RH2m, which ultimately may have a positive impact on coastal wind and fog sensitive operations, respectively.

4. Summary and conclusions

The aim of this study has been to investigate, in a systematic way, how configuration choices of an Arctic high-resolution NWP system impact the forecast quality of near-surface parameters. From this we provide guidance for use of human and computer resources in the production of operational Arctic weather forecasts.

By configuration choices, we mean basic changes to the regional NWP system, for example as for AROME-Arctic used here the impact of (1) EPS: what is the AV of a $1+6$ member ensemble versus a single deterministic forecast, (2) Higher spatial resolution: what is the AV of applying half the grid spacing (1.25 km/90L vs. 2.5 km/65L), (3) Initialization of the atmosphere: what is the AV of applying data assimilation to initialize the atmosphere in the regional system compared to downscale the coarser spatial resolution global driving model, (4) Initialization of the surface: what is the AV of applying data assimilation to initialize the surface in the regional system compared to downscale the coarser spatial resolution global driving model, and (5) what is the impact of location and size of the regional domain applied. Points 1–4 are well known to contribute to forecast accuracy and are all an important part of the quiet revolution of NWP (Bauer et al., 2015). However, in an operational setting, they also compete for limited human and computer resources. The computational cost of EPS and improved spatial resolution is huge (7 times the cost of the CNTRL-experiment

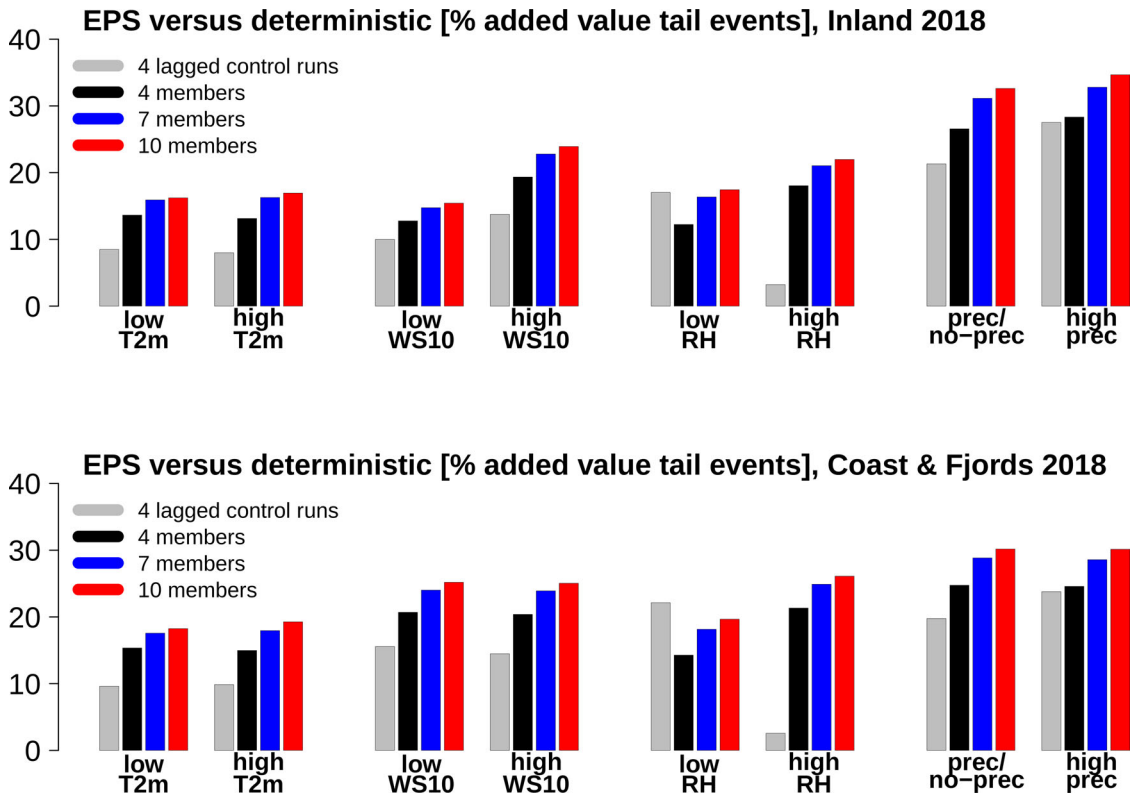


Fig. 12. Added value measured by Brier Skill Score for low/high tail events, defined for each month similar as in Fig. 10, and then averaged over all months, for inland (top) and coast and fjord (bottom) stations averaged over 2018.

in our configuration) and produces much more data that requires handling and storage. In addition, in particular EPS and the assimilation process adds complexity which require human resources to both develop and monitor. Point 5 above is one potential way to save computational costs, and thereby be able to add EPS, higher spatial resolution or better initialization. However, studies have shown how the choice of domain and location impact the results in regional NWP (e.g. Kristiansen et al., 2011) and regional climate models (e.g. Giorgi, 2019).

The evaluation of the impact of the different configuration choices was done by investigating the PCFQ of T2m and WS10 (i.e. what are the differences between the forecasts) and by quantifying the objective AV (i.e. comparison with observations) with focus on average impact and forecasting rare events. The concept of PCFQ makes it possible to inspect the impact also in regions which lack observations, e.g. the interior of Svalbard, high elevated mountain areas in Norway/Sweden, and over the ocean where the objective AV is more difficult to quantify. For more observation-dense regions the objective evaluation provides an overview of the impact on forecast accuracy.

The quantitative findings in this study represents the impact on this particular NWP system (AROME-Arctic), applied configurations, regions, period and the specific choices made related to EPS, higher spatial resolution, initialization and domain size and locations. Obviously, the experiments can be refined in numerous ways, e.g. by more sophisticated perturbation methods, better adaptation to higher spatial resolution, better choices/methods and more observations for assimilation and initialization, and a more careful tuning of the domain choices. This is exemplified by the presence of both improved and reduced forecast quality for given parameters and regions in the experiments. Furthermore, experiments for longer periods and seasons with more varying weather are necessary to confirm the details of the results. However, the presented results agree qualitatively with what has been reported in the literature for individual configuration choices, and we therefore believe that the systematic comparison presented provides some general guidance and recommendations for the configuration of regional (Arctic) NWP systems. The main findings of the study can be summarized as follows;

- All tested configuration choices have a significant impact on the forecast accuracy of near-surface parameters. However, the impact varies in amplitude and sign with configuration choices, parameters, region, weather type, and lead time.

- The largest impact was found by applying EPS, while higher spatial resolution, upper-air and surface initialization have impacts that are similar to each other in size, but differ by parameter and spatially. The smallest impact, but still significant, was found for domain size and location.
- The impact of EPS is always positive or neutral, but varies in space and time and by parameter. The most pronounced AV is found in forecasting tail events, i.e. the rare observed events during the period. It should be noted that a part of the AV by EPS can be obtained by constructing a simple lagged ensemble based on deterministic forecasts.
- The general impact of improved spatial resolution is positive or neutral for MAE averaged over all forecasts, while the results are both positive and negative in forecasting of tail events. The AV varies between neighbouring grid points, pointing towards that the origin of the AV partly comes from changes in the description of local surface characteristics like topography and land-sea-mask. The identified AV of improved spatial resolution is larger for the winter period, than during a summer period.
- For assimilation of atmospheric observations to initialize the regional NWP system we find a general positive impact on T2m and WS10, while the results for precipitation are neutral or mixed. A deterioration in quality is found for RH2m due to an unbalance in the assimilation process between increments for specific humidity and temperature.
- Surface assimilation has a general positive impact on T2m (in particular inland during cold stable conditions) and WS10, mixed or neutral impact on precipitation and a negative impact on RH2m. The latter is again related to an unbalanced assimilation of surface temperature and specific humidity in winter and the cycling process in the initialization.
- The use of the small domains have in general a neutral or negative impact on the forecast accuracy for T2m, WS10 and RH2m (with some exceptions), which is seen for both MAE and the capability to forecast tail events.

This study has shown that configuration choices for regional Arctic NWP systems matter. Furthermore, some areas with large PCFQ for T2m (interior of Svalbard and Norwegian/Swedish mountains) and WS10 (ocean close to the Norwegian coast and Norwegian/Swedish mountains) lack observations and make objective evaluation difficult.

The results are largely in line with earlier studies, but are new in the way it systematically compares the impact of the different choices for an Arctic regional domain. The introduction of EPS and moving towards (sub)kilometer-scale spatial resolution are both expensive, and compete with each other in an operational setting, but still a promising way to further enhance predictive capacity beyond state-of-the-art regional high-resolution Arctic convection permitting systems. Furthermore, for the initialization of regional models a dynamical downscaling of a coarser spatial resolution global model is not good enough, as it is required to do regional assimilation of observations to produce as accurate forecasts as possible. It should be noted that in the initialization experiments, it is not only the impact of assimilation that contributes to the presented differences, but also how the cycling of forecasts is done versus starting from the coarser spatial resolution global model. Finally the choice of the regional domain should be done with care.

It should be noted that it is the raw model output which is compared, no additional efforts are done to improve on any of the experiments, e.g. by bias-corrections or other types of post-processing, which in the end may change the difference in impact of the different configurations. The optimal configuration choice also depends on the ability to make and disseminate useful forecast products to a diversity of users. The latter is beyond the scope of this work, but is still a part of the considerations that need to be done by NWSs.

Funding

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, and the Norwegian Research Council Project 280573 'Advanced models and weather prediction in the Arctic: enhanced capacity from observations and polar process representations (ALERTNESS)'.

References

Atlaskin, E. and Vihma, T. 2012. Evaluation of NWP results for wintertime nocturnal boundary-layer temperatures over Europe and Finland. *Q. J. R. Meteorol. Soc.* **138**, 1440–1451. doi:10.1002/qj.1885

Bauer, P., Magnusson, L., Thépaut, J.-N. and Hamill, T. M. 2016. Aspects of ECMWF model performance in polar areas. *Q. J. R. Meteorol. Soc.* **142**, 583–596. doi:10.1002/qj.2449

Bauer, P., Thorpe, A. and Brunet, G. 2015. The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55. doi:10.1038/nature14956

Bengtsson, L., Andrae, U., Aspelien, T., Batrak, Y., Calvo, J. and co-authors. 2017. The HARMONIE–AROME model configuration in the ALADIN–HIRLAM NWP system. *Mon. Weather Rev.* **145**, 1919–1935. doi:10.1175/MWR-D-16-0417.1

Bouallégue, B., Z., S. E., Theis, C. and Gebhardt, 2013. Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorol. Z.* **22**, 49–59. doi:10.1127/0941-2948/2013/0374

Di Luca, A., de Elía, R. and Laprise, R. 2012. Potential for added value in precipitation simulated by high-resolution nested Regional Climate Models and observations. *Clim. Dyn.* **38**, 1229–1247. doi:10.1007/s00382-011-1068-3

Ebisuzaki, W. and Kalnay, E. 1991. Ensemble experiments with a new lagged average forecasting scheme. *WMO Research Activities in Atmospheric and Oceanic Modelling*. Report 15, 6.31–6.32. Geneva, Switzerland: WMO.

Esau, I. and Repina, I. 2012. Wind climate in Kongsfjorden, Svalbard, and attribution of leading wind driving mechanisms through turbulence-resolving simulations. *Adv. Meteorol.* **2012**, 1–16.

Feser, F., Rockel, B., von Storch, H., Winterfeldt, J. and Zahn, M. 2011. Regional climate models add value to global model data: A review and selected examples. *Bull. Amer. Meteor. Soc.* **92**, 1181–1192. doi:10.1175/2011BAMS3061.1

Frogner, I., Andrae, U., Bojarova, J., Callado, A. Escribà, P. and co-authors. 2019a. HarmonEPS - the HARMONIE Ensemble Prediction System. Wea. Forecast. 34(6), 1909–1937.

Frogner, I.-L., Singleton, A. T., Køltzow, M. Ø. and Andrae, U. 2019b. Convection-permitting ensembles: Challenges related to their design and use. *Q. J. R. Meteorol. Soc.* **145**, 90–106. doi:10.1002/qj.3525

Gascard, J.-C., Riemann-Campe, K., Gerdes, R., Schyberg, H., Randriamampianina, R. and co-authors. 2017. Future sea ice conditions and weather forecasts in the Arctic: Implications for Arctic shipping. *Ambio* **46**, 355–367. doi:10.1007/s13280-017-0951-5

Giard, D. and Bazile, E. 2000. Implementation of a new assimilation scheme for soil and surface variables in a global NWP model. *Mon. Wea. Rev.* **128**, 997–1015. doi:10.1175/1520-0493(2000)128<0997:IOANAS>2.0.CO;2

Giorgi, F. 2019. Thirty years of regional climate modeling: Where are we and where are we going next? *J. Geophys. Res. Atmos.* **124**, 5696–5723.

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N. M. and co-authors. 2017. The Met Office convective-scale ensemble, MOGREPS-UK. *Q. J. R. Meteorol. Soc.* **143**, 2846–2861. doi:10.1002/qj.3135

Hou, D., Kalnay, E. and Droegemeier, K. K. 2001. Objective verification of the SAMEX'98 ensemble forecasts. *Mon. Wea. Rev.* **129**, 73–91. doi:10.1175/1520-0493(2001)129<0073:OVOTSE>2.0.CO;2

Jung, T. and Leutbecher, M. 2007. Performance of the ECMWF forecasting system in the Arctic during winter. *Q. J. R. Meteorol. Soc.* **133**, 1327–1340. doi:10.1002/qj.99

Jung, T. and Leutbecher, M. 2008. Scale-dependent verification of ensemble forecasts. *Q. J. R. Meteorol. Soc.* **134**, 973–984. doi:10.1002/qj.255

Jung, T., Gordon, N. D., Bauer, P., Bromwich, D. H., Chevallier, M. and co-authors. 2016. Advancing polar prediction capabilities on daily to seasonal time scales. *Bull. Amer. Meteor. Soc.* **97**, 1631–1647. doi:10.1175/BAMS-D-14-00246.1

Kielland, G. 2005. KVALOBS - The quality assurance system of Norwegian Meteorological Institute observations. *Instruments and Observing Methods*. Rep. 82, WMO/TD-1265 3. https://library.wmo.int/pmb_ged/wmo-td_1265.pdf

Kim, D.-H., Kim, H. M. and Hong, J. 2019. Evaluation of wind forecasts over Svalbard using the high-resolution Polar WRF with 3DVAR. *Arct. Antarct. Alp. Res.* **51**, 471–489. doi:10.1080/15230430.2019.1676939

Kochendorfer, J., Rasmussen, R., Wolff, M., Baker, B., Hall, M. E. and co-authors. 2017. The quantification and correction of wind-induced precipitation measurement errors. *Hydrol. Earth Syst. Sci.* **21**, 1973–1989. doi:10.5194/hess-21-1973-2017

Kolstad, E. W., Bracegirdle, T. J. and Seierstad, I. A. 2009. Marine cold-air outbreaks in the North Atlantic: Temporal distribution and associations with large-scale atmospheric circulation. *Clim. Dyn.* **33**, 187–197. doi:10.1007/s00382-008-0431-5

Køltzow, M., Casati, B., Bazile, E., Haiden, T. and Valkonen, T. 2019. An NWP model intercomparison of surface weather parameters in the European Arctic during the year of polar prediction special observing period Northern Hemisphere 1. *Wea. Forecast.* **34**, 959–983. doi:10.1175/WAF-D-19-0003.1

Køltzow, M., Casati, B., Haiden, T. and Valkonen, T. 2020. Verification of solid precipitation forecasts from Numerical Weather Prediction models in Norway. *Wea. Forecast*. 35(6), 2279–2292.

Kristiansen, J., Sørland, S. L., Iversen, T., Bjørge, D. and Køltzow, M. Ø. 2011. High-resolution ensemble prediction of a polar low development. *Tellus A* **63**, 585–604. doi:10.1111/j.1600-0870.2010.00498.x

Müller, M., Batrak, Y., Kristiansen, J., Køltzow, M. A., Noer, G. and co-authors. 2017. Characteristics of a convective-scale weather forecasting system for the European Arctic. *Mon. Wea. Rev.* **145**, 4771–4787. doi:10.1175/MWR-D-17-0194.1

Nordeng, T. E., Brunet, G. and Caughey, J. 2007. Improvements of weather forecasts in polar regions. *WMO Bull.* **56**, 250–257.

Osinski, R. and Bouttier, F. 2018. Short-range probabilistic forecasting of convective risks for aviation based on a lagged-average-forecast ensemble approach. *Met. Apps.* **25**, 105–118. doi:10.1002/met.1674

Owens, R. G. and Hewson, T. D. 2018. *ECMWF Forecast User Guide*. Reading: ECMWF.

Randriamampianina, R., Bormann, N., Køltzow, M., Lawrence, H., Sandu, I. and co-authors. 2021. Relative impact of observations on a regional Arctic numerical weather prediction system. *Q. J. R. Meteorol. Soc*. 147: 2212–2232. doi:10.1002/qj.4018

Randriamampianina, R., Schyberg, H. and Mile, M. 2019. Observing system experiments with an Arctic mesoscale numerical weather prediction model. *Remote Sens.* **11**, 981. doi:10.3390/rs11080981

Raynaud, L. and Bouttier, F. 2017. The impact of horizontal resolution and ensemble size for convective-scale probabilistic

forecasts. *Q. J. R. Meteorol. Soc.* **143**, 3037–3047. doi:10.1002/qj.3159

Rojo, M., Noer, G. Claud, C. 2019. Polar low tracks in the Norwegian sea and the Barents Sea from 1999 until 2019. *PANGAEA*. doi:10.1594/PANGAEA.903058

Rummukainen, M. 2016. Added value in regional climate modeling. *WIREs Clim. Change* **7**, 145–159. doi:10.1002/wcc.378

Scheufele, K., Kober, K., Craig, G. C. and Keil, C. 2014. Combining probabilistic precipitation forecasts from a nowcasting technique with a time-lagged ensemble. *Met. Apps.* **21**, 230–240. doi:10.1002/met.1381

Singleton, A. and Grote, R. 2020. Verification of EPS forecasts using AROME-Arctic (Alertness project deliverable). *Met Norway report 02-2020*. Online at: https://www.met.no/publikasjoner/met-report/met-report-2020

Valkonen, T., Stoll, P., Batrak, Y., Køltzow, M., Schneider, T. M. and co-authors. 2020. Evaluation of a sub-kilometre NWP system in an Arctic fjord-valley system in winter. *Tellus A: Dynamic Meteorology and Oceanography* **72**, 1–21. doi:10.1080/16000870.2020.1838181

Wang, X., Steinle, P., Seed, A. and Xiao, Y. 2016. The sensitivity of heavy precipitation to horizontal resolution, domain size, and rain rate assimilation: Case studies with a convection-permitting model. *Adv. Meteorol.* **2016**, 7943845.

WMO. 2017. Navigating weather, water, ice and climate information for safe polar mobilities. WWRP/PPP 5, 74p. Online at: https://epic.awi.de/id/eprint/46211/1/012_WWRP_PPP_No_5_2017_11_OCT.pdf

Woollings, T., Franzke, C., Hodson, D. L. R., Dong, B., Barnes, E. A. and co-authors. 2015. Contrasting interannual and multidecadal NAO variability. *Clim. Dyn.* **45**, 539–556. doi:10.1007/s00382-014-2237-y

Yang, X., B. Palmason, K. Sattler, S. Thorsteinsson, B. Amstrup 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. Online at: http://www.umr-cnrm.fr/aladin/IMG/pdf/nl11.pdf