No. 2/2020
ISSN 2387-4201
Meteorology

METreport

# Verification metrics and diagnostics appropriate for the (maritime) Arctic

Alertness project deliverable

Morten Køltzow,
Matilda Hallerstig,
Rune Graversen,
Marius Jonassen,
Stephanie Mayer

| | |
|---|---|
| **Title**<br>Verification metrics and diagnostics appropriate for the (maritime) Arctic | **Date**<br>27/02/2020 |
| **Section**<br>Meteorology | **Report no.**<br>No. 2/2020 |
| **Author(s)**<br>Morten Køltzow, Matilda Hallerstig, Rune Graversen, Marius Jonassen, Stephanie Mayer | **Classification**<br>● Free ○ Restricted |
| **Client(s)**<br>Norwegian Research Council | **Client's reference**<br>Project number 280573 'Advanced models and weather prediction in the Arctic: enhanced capacity from observations and polar process representations (ALERTNESS) |

**Abstract**

This report summaries work in the Alertness project on verification metrics and diagnostics appropriate for the (maritime) Arctic. The importance of observation, interpolation and representativeness errors are discussed. Particular focus is on the wind-induced undercatchment of solid precipitation in observations and subgrid variability of weather parameters. Furthermore, it is given examples on how stratification of the verification can provide useful information in the Arctic. In addition, some new possibilities for verification of Arctic weather, i.e. rain-on-snow events, maritime and aviation icing and polar lows are explored.

**Keywords**
Arctic, weather forecasting, verification

Disiplinary signature

Responsible signature

# Verification metrics and diagnostics appropriate for the (maritime) Arctic

Deliverable 1.2
in Norwegian Research Council Project 280573 'Advanced models and weather prediction in the Arctic: enhanced capacity from observations and polar process representations (ALERTNESS).

Authors: Morten Køltzow (MET Norway), Matilda Hallerstig (NORCE and BCCR), Rune Graversen (UiT), Marius Jonassen (UNIS), and Stephanie Mayer (NORCE and BCCR).

## 1. Introduction

Verification of weather forecasts serve administrative and diagnostic purposes, i.e. verification can be used to both monitor forecast skill and to understand forecast errors. Different purposes require different verification strategies, but the basic steps of the verification process are the same; Choose a parameter, forecast systems, lead times, observational data, and appropriate diagnostics and metrics. A **diagnostic** can be derived from any geophysical data set, independently of any reference, and provide information in some condensed form. A **metric** is the quantitative comparison of a diagnostic to some reference (Massonet and Jung, 2017). A wide range of diagnostics and metrics already exist to provide a comprehensive picture of forecasts capability (Wilks, 2011, Joliffe and Stephenson, 2012).

The difference between forecasts and observations can be divided in model, observational, interpolation, and representativeness term (Kanamitsu and DeHaan 2011). As the forecast capabilities are increasing, the relative importance of the latter three terms are growing. For a number of applications it is important to distinguish the forecast model error from the other terms. However, this is relatively little studied and in particular in the Arctic (see Casati et al., 2017 and references within).

Ideally, independent high-quality observations should be used for verification, but in the Arctic a major challenge is the limited amount of reliable observations. Furthermore, Arctic surface observations are unevenly distributed, more prone to observation errors and representativeness issues than at mid-latitudes (Casati et al., 2017). Other observation sources can be used, e.g. remote sensing data, forecast- and re-analysis, but they all have limitations. In addition, a number of high-impact weather (HIW) happens or develop over the ocean in the Arctic and are less well observed. However, Casati et al. (2017) have identified some possible new avenues and key research foci in polar verification work: 1) account for observation uncertainty, 2) the impact of using model-dependent analysis, 3) process-based model diagnostics in key polar processes and 4) multivariate statistics, conditional verification and spatial verification. In addition there are Arctic specific processes that need special attention (e.g. maritime icing).

In summary, there exist numerous Arctic verification challenges. This report summarizes how the first 2 years of the Alertness project have contributed to new diagnostics and metrics appropriate for the Arctic environment. A number of these metrics and diagnostics have already been used in publications (Køltzow et al., 2019), and others are included in scientific work in preparation (Hallerstig et al., in prep, Køltzow et al., in prep). In section 2 the report presents work on observation, interpolation and representativeness errors, while several diagnostics and metrics are described in section 3. A short summary is given in section 4. A number of forecast systems and weather parameters are used to illustrate the use of diagnostics and metrics in this report and a short explanation of these with further references are given in Annex A and B, respectively.

# 2. Observation, interpolation and representativeness errors

The true performance of NWP systems become apparent only by taking interpolation, observational and representativeness errors into consideration. Quality control of observations used in the verification process is crucial, but it is difficult to identify and remove all erroneous observations. In the harsh Arctic environment, observations may be more prone to observational errors and quality control is more difficult, e.g. by buddy-check against nearby observations (Casati et al., 2017). A particular Arctic observation error is the wind-induced undercatch of solid precipitation which will be discussed in more detail in section 3.6.

A site measurement of pressure, temperature, wind speed or precipitation represents a point-observation, which differs from what the gridbox value in a NWP system represents and introduces a representativeness error. In Køltzow et al. (2019) the representativeness error was estimated for a case study. Two observations with a short distance in between are assumed to represent a grid box and the “perfect forecast” for that grid box is therefore the average of the observations (Figure 2.1). By verifying this “perfect forecast” together with several forecast systems, an estimate of the contribution from the representativeness error (i.e. the error of the perfect forecast) can be estimated. Table 2.1 shows that for this specific location and period (YOPP-SOP-NH1) the representativeness errors can be estimated to 6-11% for the large scale parameter MSLP and 19-35%, 36-42% and 15-20% for the more spatial and rapidly varying parameters T2m, WS10 and 24h precipitation, respectively. It is therefore recommended that estimation of the representativeness error is an integral part of the forecast verification exercise.

In Køltzow et al. (2019) the impact of interpolation method (i.e. forecast grid to observation point) was investigated by applying multiple interpolation methods. The impact varied between forecast system, region and parameter from being negligible to as high as 10%. It is therefore also recommended that any verification strategy should test the impact of interpolation method.

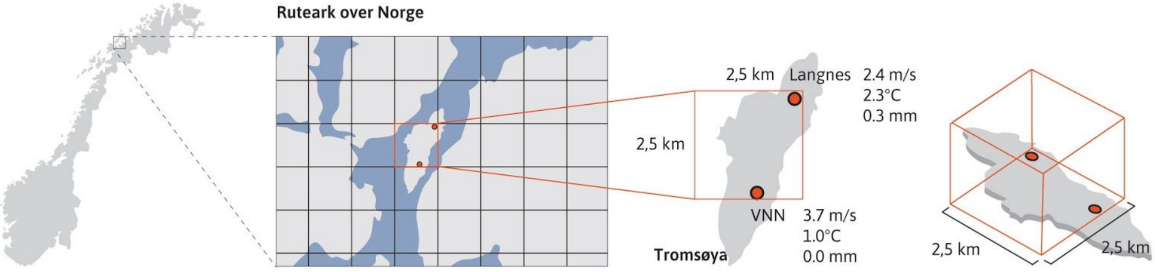
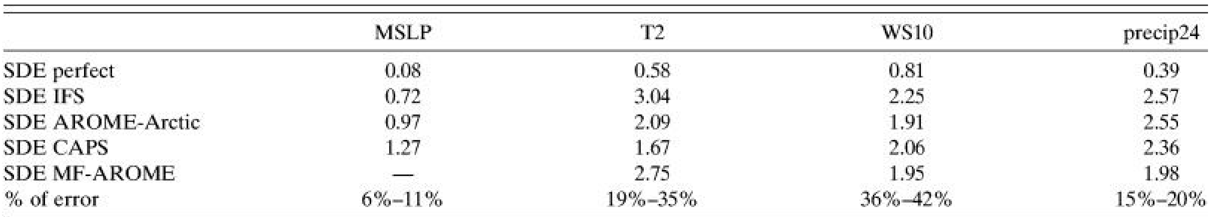


*Figure 2.1 Illustration of representativeness errors. In Tromsø, two stations do meteorological measurements within the same grid box. The perfect forecast would forecast the averaged observations, but will not verify perfectly. Photo: StorakerSchwartz/Alertness*



| | 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% |

***Table 2.1: From Køltzow et al. (2019); 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.*

# 3. Metrics and diagnostics

## 3.1 Geographical stratified verification

Verification statistics over large data sets are necessary to produce robust findings. However, this may hide important information and it is useful to stratify verification. In Figure 3.1.1 the systematic (bias) and unsystematic (standard deviation) forecast errors are stratified by geographical region for 4 different NWP systems during YOPP SOP-NH1. By doing this, important differences in model quality between regions, model systems and parameters are highlighted. To give information about statistical significance, 95% confidence intervals can be calculated by bootstrapping (not shown) and most differences are statistically significant. Even if stratified approaches have been used for a long time, this particular example taken from Køltzow et al. (2019)  is new and clearly shows some common and a few specific model weaknesses that need to be understood. For example, 2m air temperature in all model systems have large unsystematic errors inland, while the CAPS model has large systematic errors at Svalbard and IFS-HRES in the fjords. These errors were after investigations attributed to the representation of the stable boundary layer, sea ice representation and resolution issues in resolving the complex topography and land-sea-mask in the fjords, respectively. A more comprehensive discussion can be found in Køltzow et al. (2019).

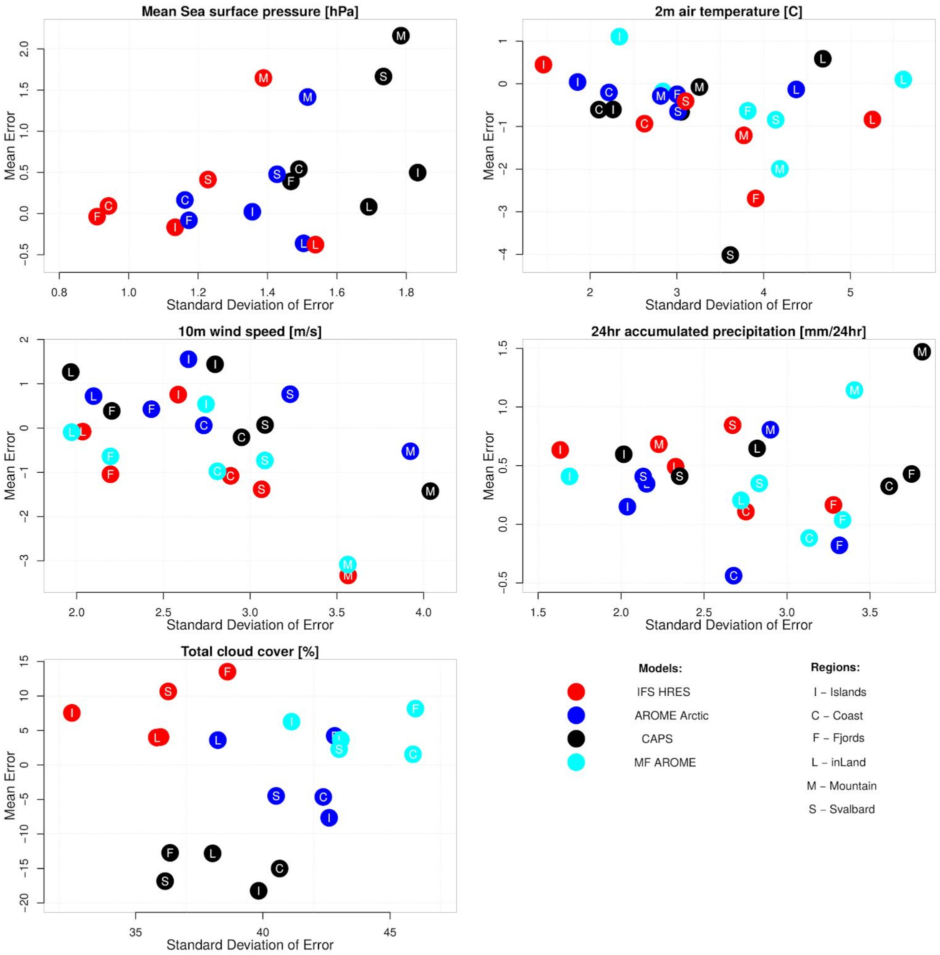


***Figure 3.1.1: From Køltzow et al. (2019); Fig. 3.*** *Mean error (bias) and SDE for MSLP, T2m, WS10, daily precipitation (precipitation), and TCC during YOPP SOP-NH1. Each circle represents one region and one model. Models are given by color: IFS-HRES (red), AROME-Arctic (blue), CAPS (black), and 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.*

## 3.2 Conditional verification of 2m air temperature

A new way to visualize conditional verification of T2m was made to investigate the large inland errors in temperature (shown in section 3.1), by stratifying the forecast errors on cloud cover, wind speed and time of the day (Figure 3.2.1). The increase in T2m forecast errors during calm, cloud-free conditions without the presence of solar radiation (smaller errors during daytime), points toward issues in the representation of the stable boundary layer as a common problem for all forecast systems. A more comprehensive description and discussion can be found in Køltzow et al. (2019).

Another way to diagnose model behaviour with respect to the stable boundary layer is to compare temperature drops during winter nighttime in observations and forecasts. The four forecast systems evaluated in Køltzow et al. (2019) are used to illustrate this. In Figure 3.2.2 the maximum 6 hourly temperature drop during nighttime from NWP forecasts and observations at the Sodankyla observation site are compared. All NWP systems share in common that they have to weak temperature drops, i.e. they are not sensitive enough to the forcing. Furthermore, in Table 3.2.1 the average 2-hourly temperature drops are calculated conditioned by cloud and wind conditions for observations, AROME-Arctic and IFS-HRES. It is evident that during cloudfree, calm conditions the temperature drops in observations are stronger (-2.6C/2h), than in AROME-Arctic (-1.7C/2h) and IFS-HRES (-0.8C/2h) similar to what was seen in Figure 3.2.2. In addition, it should be noted that calm conditions occurs more frequently in observations (~32% of time), than in AROME-Arctic (~15% of time) and IFS-HRES (~12% of time). The same is also true for cloud-free conditions which happen ~36%, ~30% and ~20% of the time for observations, AROME-Arctic and IFS-HRES, respectively. Basically, these diagnostics show that temperature drops most often happen in calm and cloud-free conditions which are too weakly represented and too rarely happen in the NWP forecasts.

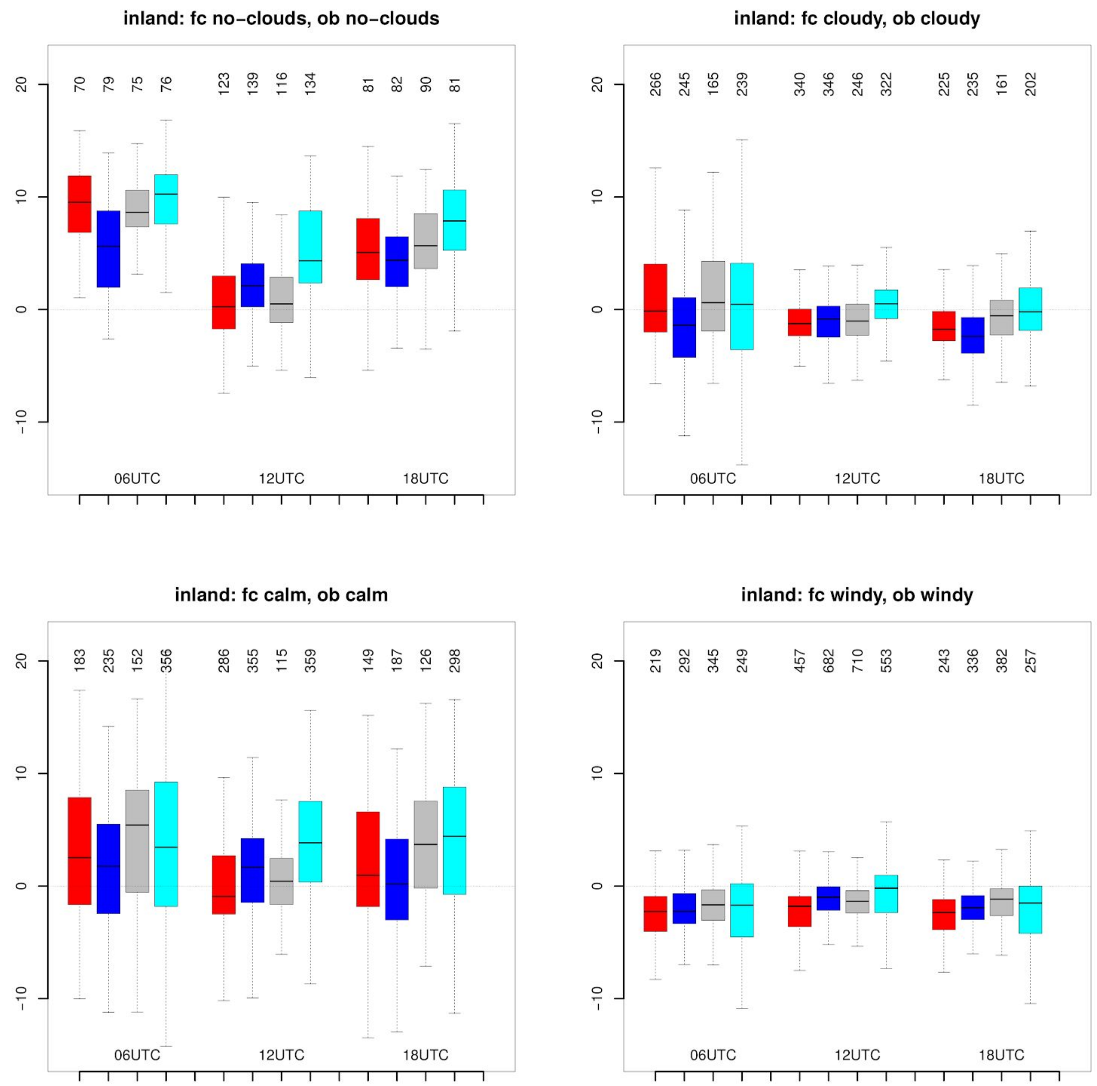


***Figure 3.2.1: From Køltzow et al (2019): Fig. 5.*** *Conditional verification of T2 for inland stations. Box-and-whiskers plot of T2 errors (forecasted minus observed) conditioned by (top) TCC and (bottom) wind. Cloud-free is defined as TCC less than 30% and cloudy as TCC larger than 70%. Calm conditions 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 the top, and outliers are omitted to increase readability in plots.*

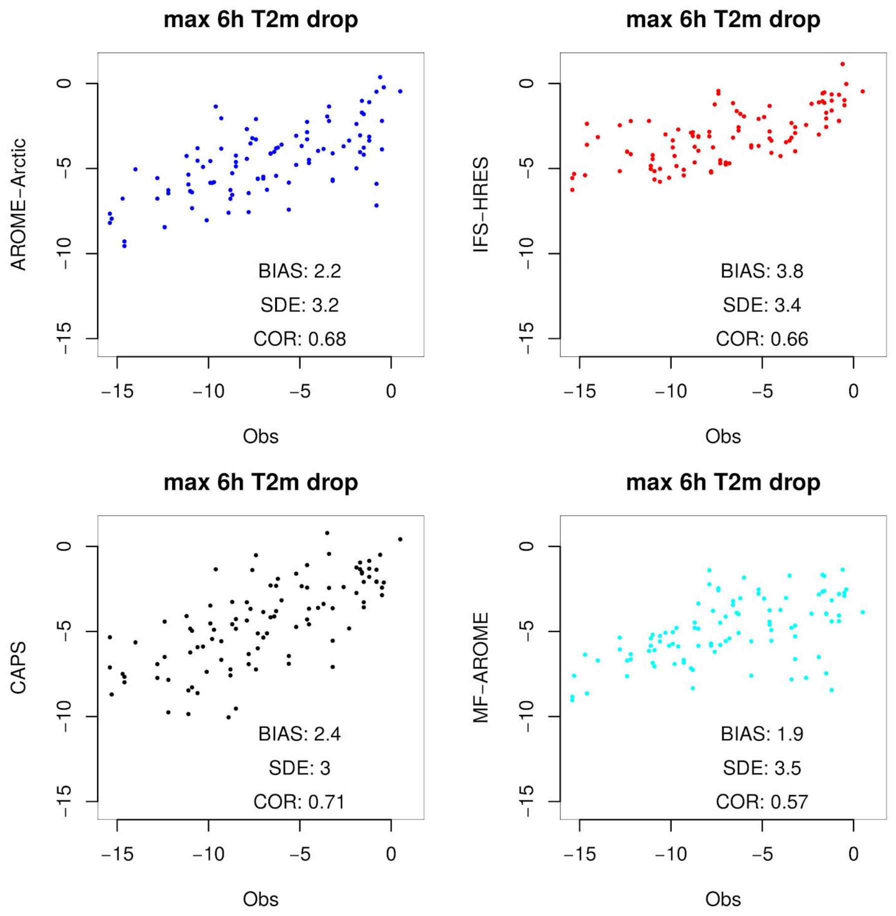


***Figure 3.2.2.*** *Maximum 6h temperature drop during nighttime (18-06 UTC) for four NWP systems (y-axis) vs observations (x-axis) during YOPP-SOP-NH1 (February and March 2018) at Sodankyla.*

| Conditional average T2m drops | WS =< 1.5 m/s | | | 1.5 < WS =<2.7 | | | WS > 2.7 | | |
|---|---|---|---|---|---|---|---|---|---|
| | OBS | AA | IFS | OBS | AA | IFS | OBS | AA | IFS |
| NN <= 2 | -2.6 | -1.7 | -0.8 | -1.9 | -0.8 | -0.5 | 0.3 | -0.1 | -0.1 |
| 2 > NN >=7 | -1.2 | -0.8 | -0.2 | -0.5 | -0.8 | -0.3 | 0.3 | 0.2 | 0.0 |
| NN = 8 | 0.3 | -0.1 | -0.1 | -0.3 | -0.4 | -0.2 | 0.2 | 0.1 | 0.2 |

***Table 3.2.1.*** *Average temperature drop (2 hourly) during nighttime (no solar radiation) conditioned by cloud cover and wind speed at Sodankyla during YOPP-SOP-NH1 (February and March 2018) for observations (OBS), AROME-Arctic (AA) and IFS-HRES (IFS).*

## 3.3 Rain-on-snow and warm-spell events

Rain-On-Snow (ROS) events can have major consequences, e.g. in Svalbard with substantial impact on infrastructure, society, and wildlife (e.g. Serreze et al. 2015; Hansen et al. 2014). In a climate context this phenomenon together with warm spell periods are relatively well studied although their definition varies (e.g. Pall et al., 2019, Cohen et al., 2015, Bienik et al., 2018, McCabe et al., 2006, Vikhamar-Schuler et al., 2016). However, to our knowledge, less work has been done on short-range prediction capabilities of such events.

To measure how well warm spell periods are forecasted we define melting hours as hours with 2m air temperature above 0°C and melting-hour intensity as the mean temperature during melting hours. These diagnostics can be extracted from observations and forecast systems. Our main interest is when this happens during cold periods, leading to re-freezing and implying consequences as described above. The comparison between forecasts and observations are done by standard verification metrics (e.g. Equitable Threat Score, Frequency Bias Index), but stratified by the observed monthly mean temperature. To measure how well ROS events are forecasted we define ROS hours as hours with precipitation and 2m air temperatures above 0°C, and ROS hours intensity as the average precipitation during ROS hours. Figure 3.3.1 shows verification of ROS and warm spell hours and intensity. The forecast skill of AROME-Arctic and IFS-HRES is relatively independent of the observed monthly average temperature, and AROME-Arctic has a higher skill than IFS-HRES. AROME-Arctic and in particular IFS-HRES underestimate the frequency and intensity of warm spell hours. The forecast accuracy and frequency of ROS events are higher for AROME-Arctic than IFS-HRES, highlighting the importance of model resolution. However, both forecast systems predict reasonably well the ROS intensity when the events are forecasted.

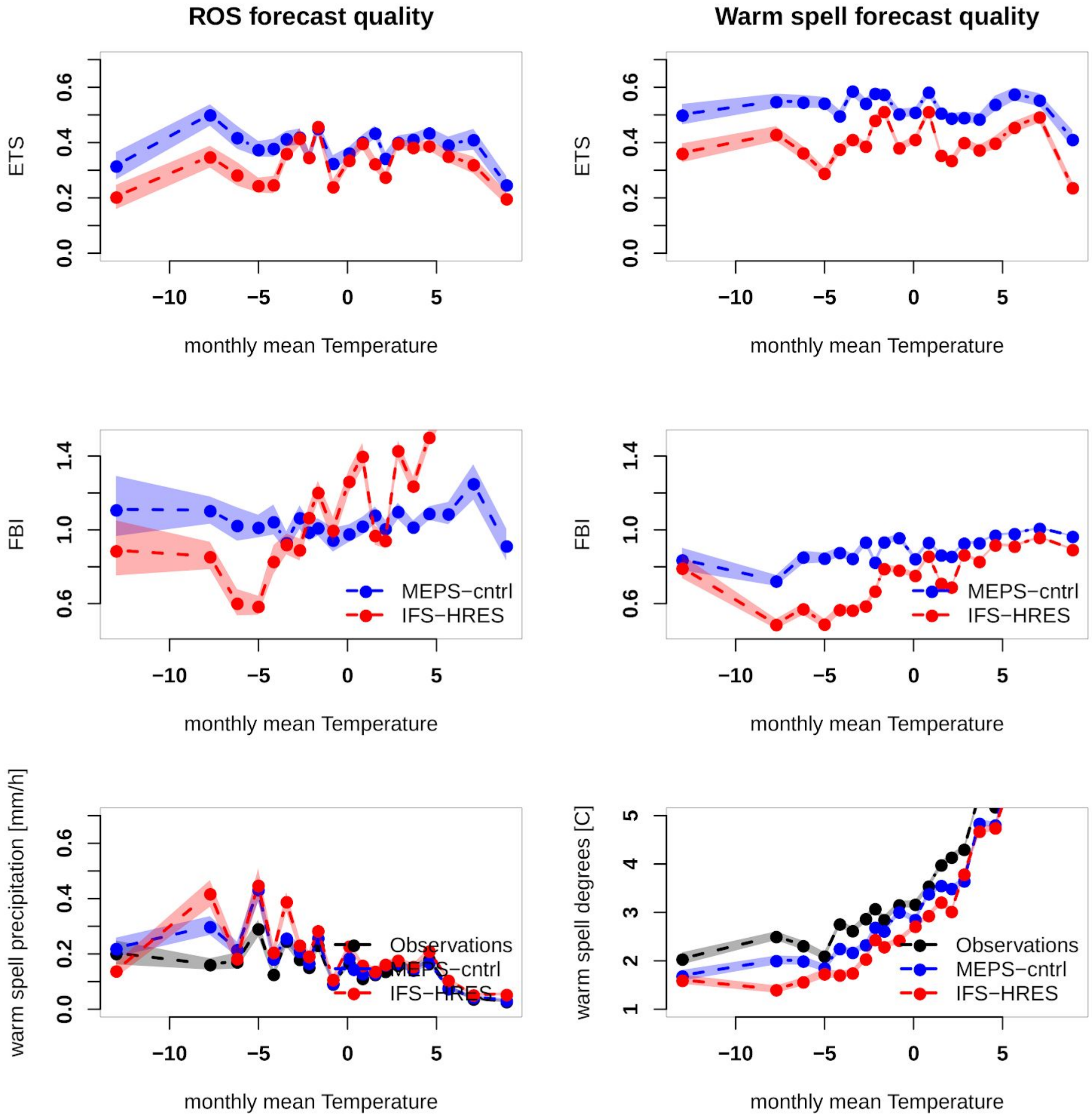


***Figure 3.3.1.*** *Forecast performance ROS events (left) and warm spell events (right) as a function of observed monthly mean temperature. At top row; ETS forecast skill for ROS hours (left) and melting hours (right), in the mid-row; Frequency Bias Index for ROS hours (left) and melting hours (right) and bottom row; mean precipitation during ROS events (left) and mean temperature during warm spell periods (right).*

## 3.4 Maritim icing

Ships and maritime installations exposed to sub-freezing conditions may experience icing with large potential consequences, in which ship-capsizing leading to human casualties are the most dramatic ones (Samuelsen, 2017a). Forecasting the phenomenon is therefore of high importance. The lack of regular observations is problematic and most verification studies are connected to limited data sets or case studies (e.g. Samuelsen et al., 2017b, Samuelsen., 2018). Maritime icing conditions are multi-parameter dependent or compound events, e.g. wave height, wind speed, and air temperature are all important. To circumvent the lack of direct icing observations and to focus on the input from atmospheric-forecast models we force an state-of-the-art ship-icing model with observed and forecasted atmospheric parameters and compare the icing-rate output.

We employ a simplified version of the icing model of Samuelsen et al. (2017a) in which the most important simplifications are the assumption of a constant brine freezing temperature of -4°C, and that wave height and period can be estimated based on wind speed following Zakrzewski (1987) and assuming a fetch of 100 nm. Since salt is expelled during freezing of saline water, the freezing temperature of the brine on a ship exposed to sea spray, is much lower than the freezing temperature of the sea water surrounding the ship. -4°C is an average value based on model calculations of the 37 icing events investigated in Samuelsen et al. (2017b). Estimating wave height and period based on wind speed and a constant fetch of 100 nm would probably lead to an overestimation of the waves and spray amount that the ship would be exposed to near the coast and sea ice when the wind is blowing from the coast and sea ice, i.e. in a fetch-limited regime. However, it is probably better to use the Zakrzewski (1987)-method, than using a relationship between wind and waves in a fully developed sea, which is an assumption that is commonly applied in other icing studies (e.g. Horjen, 2013). The advantage of using such empirical methods between wind and waves in this study is to avoid the introduction of a wave model, when studying the errors in an atmospheric model. Other simplifications are that we assume constant sea spray duration and frequency, estimate sea spray temperature as a linear combination of sea surface and air temperature, calculate long wave radiation from temperature, neglect short wave radiation, assume a constant angle between ship and waves, and ship and wind, and assume that the waves and winds follow the same direction. By doing this we can calculate icing rates with sea-surface temperature from IFS-HRES (based on the Operational Sea Surface Temperature and Sea Ice Analysis, OSTIA; Donlon et al. (2012) from the Met Office) and air temperature and wind speed from observations and forecasts as input. Following the icing intensity classification as defined by Samuelsen et al. (2017) we get no icing (icing rate < 0.05 cm/h), light icing (0.50 cm/h > icing rate >= 0.05 cm/h), moderate icing (1.34 cm/h > icing rate >= 0.50 cm/h) and severe icing (icing rate >= 1.34 cm/h).

In order to test this approach we identify exposed coastal stations (e.g. lighthouses) where wind speed and temperature are observed. We then calculate icing rates with both observations and forecasts as input to acquire forecasted and “observed” maritime icing. The icing model is also forced with combinations of forecasted and observed parameters to assess the impact of the individual forecast parameters. The approach was tested on the period December 2018 to March 2019 with the AROME-Arctic and IFS-HRES forecast systems (Table 3.4.1). AROME-Arctic and IFS-HRES underestimate the occurrences of maritime icing conditions, but AROME-Arctic is closer to the “observed icing”. In Figure 3.4.1a, the Hit Rate (HR) is plotted against the False Alarm Ratio (FAR) for light and moderate icing. The plot includes pure forecasts, but also combinations of forecasts and observations to investigate the individual forecast components (i.e. temperature and wind speed). For example, AROME-Arctic predictions of moderate icing show a high FAR and very small HR, i.e. not a very good prediction. By using observed temperature and AROME-Arctic wind speed the FAR is reduced by ~ 50%, but the HR is still low. However, by using observed wind speed and forecasted temperature the FAR is reduced to ~ 0.2 and the HR increased to > ~ 0.5. A close inspection of both models and light and moderate icing categories show similar behaviour which illustrates that the wind speed contributes to a larger part of the forecast errors.

The proposed approach can also be applied to Ensemble Prediction System (EPS) verification (WP4 in Alertness). In Figure 3.4.1b, reliability plots from two AROME-Arctic EPS experiments (with and without SST perturbations) are compared. Both experiments show similar behaviours and  produce reliable probabilities for light icing.  Furthermore, In Figure 3.4.2, time series of icing categories calculated from observations/forecasts from a coast guard ship in the Barents Sea during a cold air outbreak is shown. It is shown that the moderate icing “observed” 26 March is not forecasted due to forecast temperature issues, while the duration of the icing on the 29 March is not forecasted well by AROME-Arctic due to the forecasted wind speed.

| icing categories | no | light | moderate | severe |
|---|---|---|---|---|
| **“Observed”** | 82737 | 10801 | 274 | 0 |
| **AROME-Arctic** | 84654 | 9069 | 89 | 0 |
| **IFS-HRES** | 88373 | 5439 | 0 | 0 |

***Table 3.4.1.*** *Estimated icing categories based on observed and forecasted (AROME-Arctic and IFS-HRES) wind speed and temperature from exposed coastal stations northern Norway December 2018 to March 2019.*

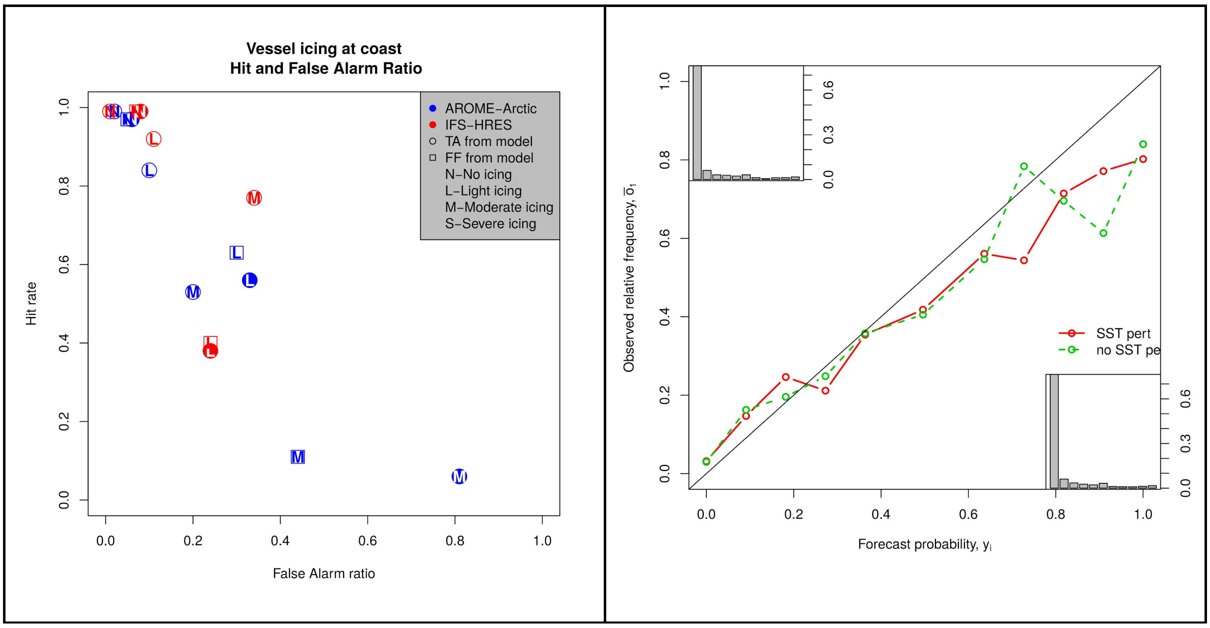


***Figure 3.4.1.*** *a) Hit rate versus False alarm ratio (left) for icing categories based on observed and forecasted (AROME-Arctic and IFS-HRES) wind speed and temperature from exposed coastal stations northern Norway December 2018 to March 2019. b) Reliability plot (right) based on two AROME-Arctic EPS experiments with (green) and without (red) SST perturbations for 10. March - 31. March 2018.*

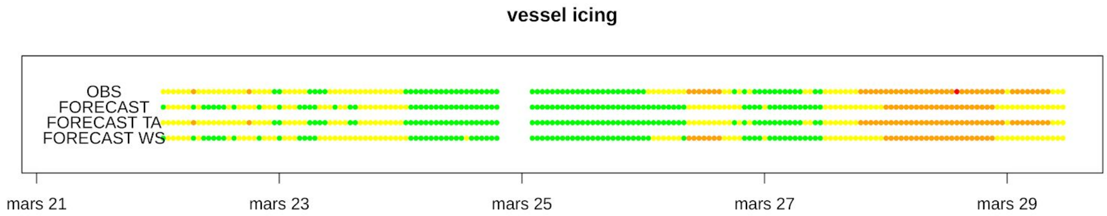


***Figure 3.4.2.*** *Time series of estimated icing categories from observations and forecasts for a coast guard ship during a cold air outbreak. No icing indicated by green, light icing indicated by yellow, moderate icing indicated by orange and severe icing indicated by red.*

## 3.5 Aviation icing

In-flight aircraft icing is a major hazard for aviation safety and happens when aircrafts enter clouds containing supercooled liquid water. Reports by pilots are the only widely available data source for evaluation. There are a number of limitations connected to the observations; observed icing intensity by pilots are subjective, depends on e.g. aircraft type and de-icing facilities, are not obligatory, "no icing" conditions are rarely reported and the data sets are biased towards areas with high traffic density. These limitations are discussed in more depth by Kalinka et al. (2017) which use similar pilot reports to evaluate aircraft icing forecasts over Europe. In addition, pilots take action to avoid icing conditions, i.e. they avoid areas where icing conditions are observed or forecasted. Since AROME is an important tool for the Norwegian meteorologists issuing aviation warnings, there is likely that aircrafts over Norway avoid, if possible, icing areas indicated by AROME and further bias the data set. Recently, e.g. in Bowyer and Gill (2018), icing products from satellite measurements has been used for verification. Such datasets make it possible to detect false alarms. However, also these products have limitations, e.g. coverage and the need for daylight which is a problem in the Arctic winter.

To our knowledge very limited evaluation of the AROME aircraft icing routines are done over Norway and adjacent areas. Here, we make steps towards better understanding of Nordic aircraft icing forecast capabilities by assessing AROME icing forecasts based on pilot reports. This work, can be used further in Alertness and also feed into work outside of the project making use of satellite based data as in Bowyer and Gill (2018).

A pilot report can look like; "ARS NH90 SEV ICE OBS AT 0920Z 5NM N OF FINNSNES FL090/070", which manually are translated into a more useful format including icing intensity (trace, fibula, moderate and severe), time, location (latitude/longitude) and height intervals (meters). In Figure 3.51a the forecasted icing intensity, conditioned by the observations is shown for December 2018 to February 2019. When no icing is observed the majority of the forecasts also indicate no icing. However, parts of the forecasts indicate risk for some icing. These forecasts are not necessarily wrong since this can be due to no aircraft flying in the area (either it is not scheduled or avoided due to issued warnings). When moderate or severe icing are observed, the forecast intensity also indicates icing risk in the majority of the forecasts, but on average the forecast intensity is less than the reported intensity. An interpretation of this is that the forecasts underestimate the icing intensity on occasions, but this might improve if we take neighbourhood information into account (e.g. verify against highest forecasted intensity in a neighbourhood). However, it is well known that NWP systems often contain too little supercooled cloud water and thereby is an underestimation of the icing intensity expected. Comparing the forecasts when moderate and severe icing have occured it seems like the forecasts are not able to distinguish properly between the two categories.

In Figure 3.5.1 the minimum and maximum height of observed and forecasted icing are compared. Most of the time the lowest observed level is higher than the lowest forecasted level which can be considered as a good forecast, i.e. most icing conditions are observed above the lowest forecasted level. A substantial part of the highest observed icing levels are above the highest forecasted level. This implies that often icing happens higher up in the atmosphere, at colder temperatures, than forecasted by AROME. Again, this might not be a surprise with the knowledge about NWP systems often contain too little supercooled cloud water.

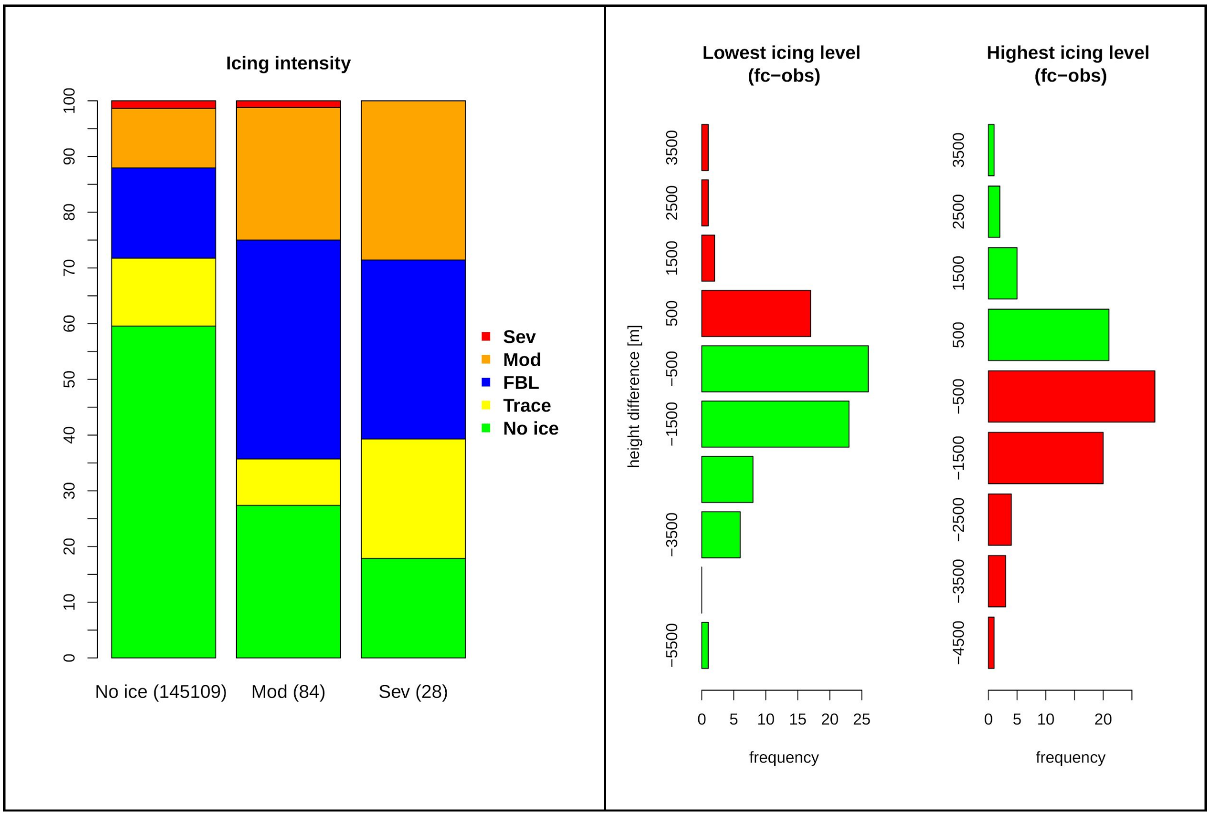


***Figure 3.5.1*** *Aviation icing, left; forecasted icing intensity (vertical) conditioned by observed icing reports by pilots, "No ice", "Mod(erate)" and "Sev(ere)" (horizontal). Number of cases given in parentheses. Right; difference between forecasted and observed highest and lowest levels (red bars indicate when observations are outside the forecasted height range.*

## 3.6 Verification of winter precipitation

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

Single-Alter shields are the commonly used equipment and a substantial undercatch of solid precipitation in windy conditions is experienced. The impact on forecast verification is illustrated in Figure 3.6.1 where forecast skill for liquid and solid precipitation is shown. For calm conditions (less undercatch) the forecast skills are similar for both precipitation phases, while in windy conditions (large undercatch for solid precipitation) the forecast skill is substantially lower for solid than for liquid precipitation due to erroneous observations.

A verification strategy for winter precipitation is to stratify the verification by wind speed or only include precipitation in calm conditions. However, this approach rejects large amounts of data and does not reflect all types of precipitating weather. A possible strategy is to adjust the measured observation for undercatch with so-called transfer functions. A number of transfer functions exist and aim at transfering the measured precipitation to what would have been observed with equipment less prone to undercatch (e.g. Kochendorfer et al., 2017, Wolff et al., 2015, Smith, 2007, Førland et al., 1998). An example of this is shown in Figure 3.6.2 where a number of different transfer functions are applied to estimate accumulated solid precipitation. Comparing AROME-Arctic and IFS-HRES forecasts with the raw unadjusted observations, shows an overestimation of solid precipitation with 22-23%. However, after adjusting the measured precipitation a forecast underestimation of solid precipitation emerges. This highlights the importance of taking wind-induced undercatch into account. However, there is a large spread between the different transfer functions, which indicates a high uncertainty in the estimates. Existing transfer functions are all functions of wind speed and most of them are also functions of temperature. On average they work well, but not necessarily for single cases, since e.g. the mass, shape and fall speed of the snow also have an impact on undercatch. As a result, applying transfer functions for skill assessment is difficult and may introduce an additional layer of noise in the results (not shown). Transfer functions are therefore most appropriate over large data samples, but show a considerable spread.

In practise both strategies should be combined and complemented with other observational data sets when possible. The latter is done in Figure 3.6.3 where precipitation is estimated from changes in snow water equivalent (SWE) from snow pillow measurements provided by the Norwegian Water Resources and Energy Directorate. Estimated precipitation at Filefjell December 2018 to March 2019 are approximately 3 times higher based on snow pillow data than the SA precipitation gauge. Also the adjustments of the SA gauge is lower than the snow pillow data estimates, while both MEPS and IFS-HRES forecasts are slightly higher than the snow pillow estimates. A comprehensive discussion of solid precipitation verification and the problem with undercatch, including stratification on wind speed, adjustment of observations and the use of snow pillow data, is the topic of an Alertness scientific publication in preparation (Køltzow et al., in preparation).

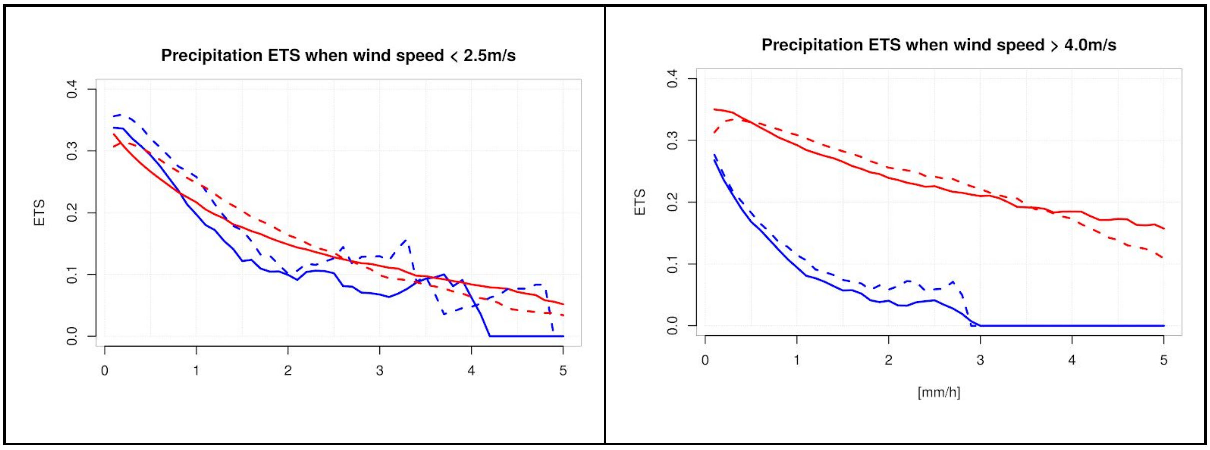


***Figure 3.6.1.*** *Equitable Threat Score for liquid (red) and solid (blue) 1h accumulated precipitation for calm (left) and windy (right) conditions. AROME-Arctic skill in solid lines and IFS-HRES skill in dashed lines.*

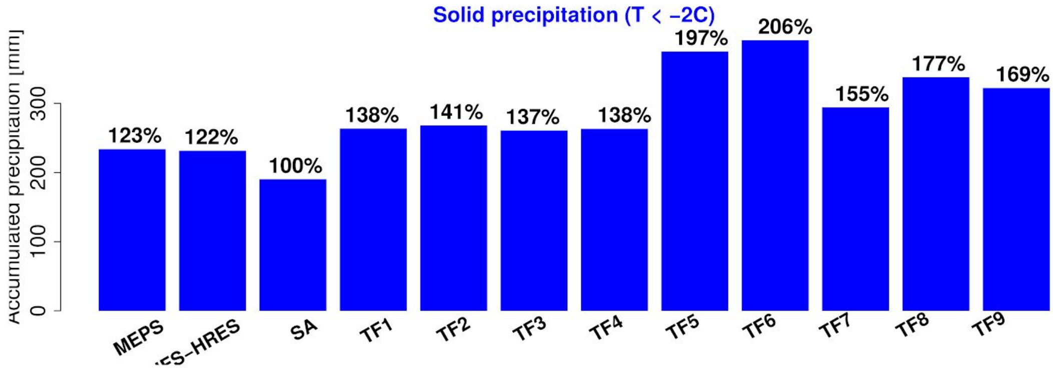


***Figure 3.6.2.*** *Accumulated solid precipitation (T2m < -2C) averaged over all Norwegian stations from May 2016 to April 2019 for MEPS, IFS-HRES, raw observations (SA) and adjusted SA based on different transfer functions. The %-number compares the precipitation amounts to the observed SA.*

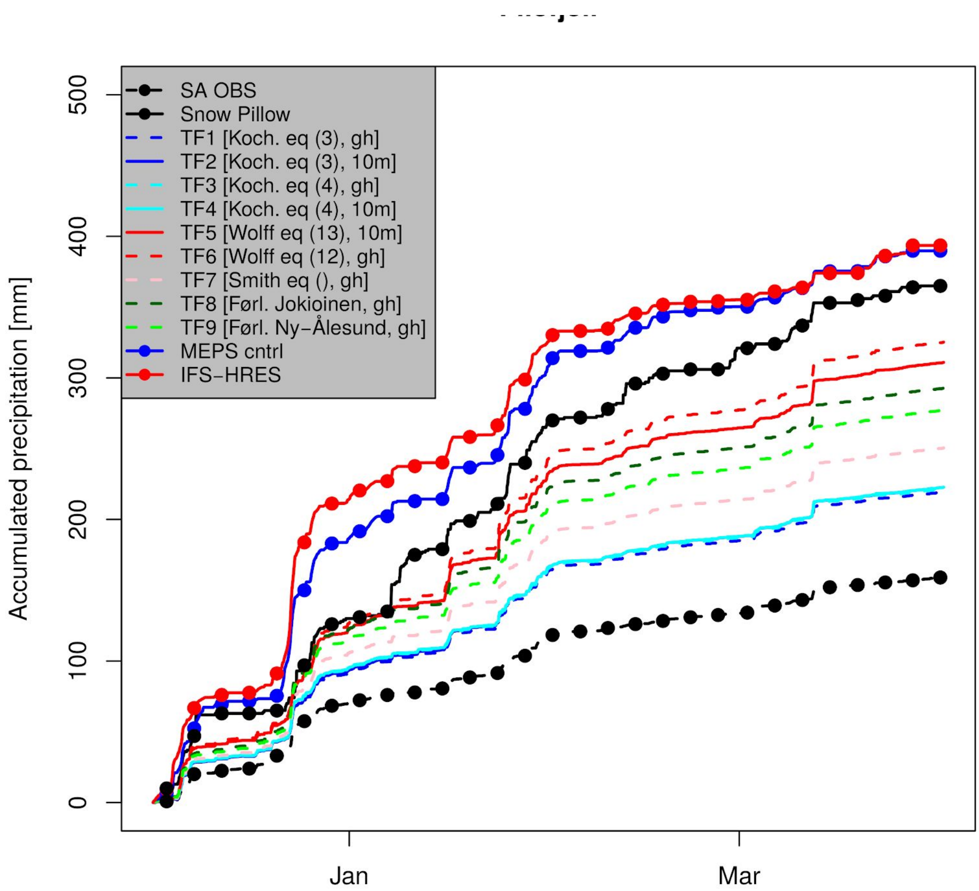


***Figure 3.6.3.*** *Precipitation* ***Filefjell*** *(WMO: 1364, 61.1780N, 8.1125E, 956 masl) December 2017 to March 2018. SA precipitation gauge (standard measurement equipment) in dashed black line with circles, TF1 to TF9 are adjusted measurements and estimated precipitation from snow water equivalent changes (snow pillow) in solid black line with circles. Forecasted precipitation from AROME-Arctic (blue) and IFS-HRES (red) in solid lines with circles.*

## 3.7 Polar low verification

Polar lows are small, intense weather systems that form over open, Arctic waters north of the main polar front. Due to their small scale and rapid development they are challenging to forecast. They can cause sudden changes from calm and sunny weather to heavy snow showers with strong, gusty winds and high waves and can therefore cause danger for operations at sea. In addition, these weather conditions can lead to maritime and aviation icing, turbulence and increased avalanche danger. To assist communities to be prepared for these hazards, it is important to provide the best possible forecasts. It is possible to use coastal observation stations to monitor forecast performance during landfall of polar lows, but to evaluate their full lifetime, we need observations with good spatial and temporal coverage over the open arctic waters. For this, remote sensing such as scatterometer and SAR satellites are most appropriate. The method described here (Hallerstig et. al., in preparation) focuses on scatterometer, but could potentially be adapted for SAR. Because the satellite observations have a coarser resolution than Arome-Arctic, we regridded the model from a 2.5 km horizontal grid to a 12.5 km grid to match the observations and avoid effects due to different resolutions. By manipulating the Arome-Arctic data this way, its maximum wind speed decreased by 5 m/s, but the overall structure and results of the study did not change. We also need to filter the data so that only observations and model data that correspond to each other in space and time are used. For the scatterometer data, this means that observations that occur between model output times are removed. A time window of +/- 30 min is used. For model data, at each timestep grid points that have no satellite data nearby are discarded. Next, the maximum value for each grid point during the polar low life time is plotted. Through this kind of composite we can get an overview of the total model performance in relation to satellite observations (Figure 3.7.1).

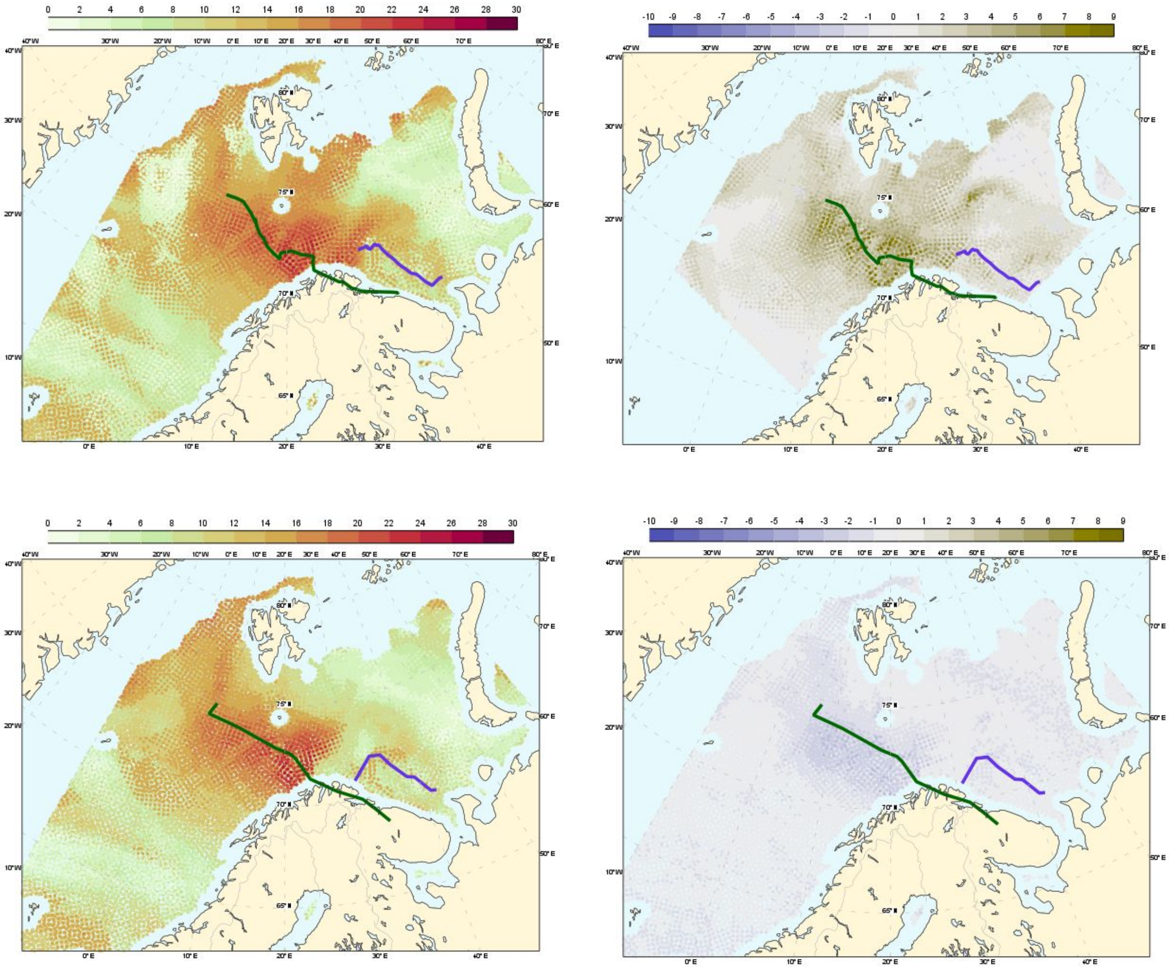


***Figure 3.7.1:*** *Example of scatterometer composites for a dual polar low event 20161208. Upper row shows a comparison to Arome-Arctic, lower row shows ECMWF IFS HRES. Purple and green lines show the tracks of the polar lows as forecasted by the models. Left: Composite of scatterometer winds after the data was filtered to match model output. Red shades show higher wind speeds. Right: Difference between scatterometer composites and the model. Blue shades show that model had less wind than observed by the satellite. Olive shades show that model had more wind than observed by the satellite.*

To compare the evolution over time between models, we created time series along the polar low tracks, following the minimum sea level pressure at each time step. For each step, the maximum wind speed within a certain radius from the low pressure center is indicated. The radius needs to be large enough to capture maximum wind speed associated with the polar low, but a too large radius will include features not directly connected to the polar low, such as another, stronger polar low, or the synoptic scale low center that often occurs together with polar lows. The optimal radius will be different for each case, and was chosen manually. For example, the stronger polar low in figure 3.7.1 (green line) had a radius of 250 km, while the weaker polar low shown by the purple line had a radius of 125 km. These plots allow for a comparison of intensity and timing between model experiments (figure 3.7.2).

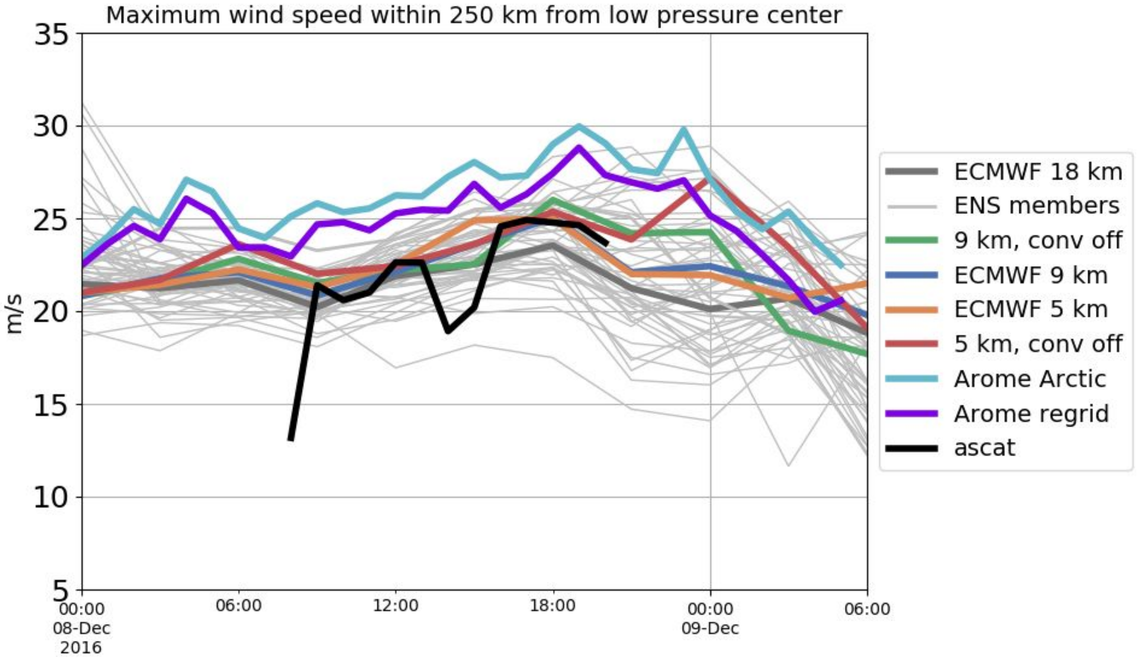


*Figure 3.7.2: Maximum wind speed at each time step for the larger polar low in figure 3.7.1. Black line shows scatterometer winds where these were available, grey lines are the ECMWF IFS ensemble and colored lines show model experiments.*

# 3. Summary

Arctic surface observations are unevenly distributed, more prone to observation errors, and representativeness issues than at mid-latitudes. Hence, the Arctic verification challenges are many. In this report, it is explained how Alertness contributes to a better understanding of Arctic forecast capabilities by a set of new diagnostics and metrics or by new use, visualization or condition of existing methods.

It has been illustrated how some observation, interpolation, and representativeness errors can be considered. Furthermore, it has been demonstrated how temperature diagnostic can highlight the problematic NWP behaviour in connection with the stable boundary layer. Furthermore, new metrics and diagnostics have been suggested for high-impact weather like polar lows, rain-on-snow, vessel and aviation icing.

The metrics and diagnostics described in this report will, together with already existing tools be used to evaluate Alertness NWP simulations, e.g. in Task 1.3.

# Annex A: Forecast systems

The NWP systems used to illustrate verification and metrics and diagnostics in this report 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, and AROME with Météo-France setup (MF-AROME) with 2.5-km grid spacing (Seity et al. 2011). Køltzow et al. (2019) give an overview in the differences in the model system formulations of these systems.

# Annex B: Weather parameters - abbreviations

| | |
|---|---|
| MSLP | Mean Sea Level Pressure |
| T2m | 2m air temperature |
| WS10m | 10m wind speed |
| precip1 | 1 hour accumulated precipitation |

# References

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.

Bieniek, P.A., U.S. Bhatt, J.E. Walsh, R. Lader, B. Griffith, J.K. Roach, and R.L. Thoman, 2018: Assessment of Alaska Rain-on-Snow Events Using Dynamical Downscaling.*J. Appl. Meteor. Climatol.,* **57**, 1847–1863, https://doi.org/10.1175/JAMC-D-17-0276.1

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.

Cohen, J., Ye, H., and Jones, J. ( 2015), Trends and variability in rain-on-snow events, *Geophys. Res. Lett.*,42, 7115– 7122, doi:10.1002/2015GL065320.

Donlon, C. J., M. Martin, J. Stark, J. Roberts-Jones, E. Fiedler, and W. Wimmer, 2012: The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Remote Sens. Environ.*, **116**, 140–158, https://doi.org/10.1016/j.rse.2010.10.017.

Førland et al., 1998, in Goodison B., Louie P.Y.T. and Yang D., 1998: WMO Solid Precipitation Measurement Intercomparison, Final Report, WMO IOM Report No. 67, WMO/TD – No. 872.

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.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2012: Forecast Verification: A Practitioner's Guide in Atmospheric Sciences. Wiley-Blackwell, 292 pp.

Hallerstig, M. et al., in preparation

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.

Horjen, I., 2013, Numerical modeling of two-dimensional sea spray icing on vessel-mounted cylinders Cold Reg. Sci. Technol., 93, pp. 20-35, 10.1016/j.coldregions.2013.05.003

Jolliffe, I.T., and D.B. Stephenson, 2012: Forecast Verification: A Practitioner's Guide in Atmospheric Science, 2nd Edition, Wiley and Sons Ltd, 274 pp.

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.

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.

Køltzow, M., B. Casati, E. Bazile, T. Haiden, and T. Valkonen, 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. Forecasting,* **34**, 959–983, https://doi.org/10.1175/WAF-D-19-0003.1

Køltzow et al. in preparation

Massonet F. and T. Jung, 2017, APPLICATE, metrics and all the rest, applicate.eu; https://applicate.eu/images/APPLICATE_metrics_final.pdf

McCabe, G. J., M. P. Clark, and L. E. Hay(2007), Rain-on-snow events in the western United States, *Bull. Am. Meteorol. Soc.*, 88, 319–28.

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

Pall, P., L.M. Tallaksen, and F. Stordal, 2019: A Climatology of Rain-on-Snow Events for Norway. *J. Climate,* **32**, 6995–7016, https://doi.org/10.1175/JCLI-D-18-0529.1

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

Samuelsen, E., M., 2017a, PhD thesis; Prediction of ship icing in Arctic waters -- Observations and modelling for application in operational weather forecasting, UiT -- The Arctic University of Norway, https://munin.uit.no/handle/10037/11801

Samuelsen, E. M., K. Edvardsen and R. Graversen, 2017b: Modelled and observed sea-spray icing in Arctic-Norwegian waters. Cold Regions Science and Technology. 134, 54-81.

Samuelsen, E. M., 2018, Ship-icing prediction methods applied in operational weather forecasting. Q.J.R. Meteorol. Soc., 144: 13-33. doi:10.1002/qj.3174

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.

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.

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.

Vikhamar-Schuler, D., K. Isaksen, J.E. Haugen, H. Tømmervik, B. Luks, T.V. Schuler, and J.W. Bjerke, 2016: Changes in Winter Warming Events in the Nordic Arctic Region. *J. Climate,* **29**, 6223–6244, https://doi.org/10.1175/JCLI-D-15-0763.1

Wilks, D.S., 2011: Statistical Methods in the Atmospheric Science, Third edition, Elsevier, 676 pp.

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.

Zakrzewski, W.P, 1987, Splashing a ship with collision-generated spray. Cold Reg. Sci. Technol., 14 (1) (1987), pp. 65-83