Norwegian Meteorological Institute

METreport

No. 6/2020
ISSN 2387-4201
Meteorology

Verification of EPS forecasts using AROME-Arctic

Alertness project deliverable

Andrew Singleton and Rafael Grote

Photo: Eirik Samuelsen

Norwegian Meteorological Institute

METreport

Title	Date
Verification of EPS forecasts using AROME-Arctic	June 5, 2020
Section	**Report no.**
Meteorology	6/2020
Author(s)	**Classification**
Andrew Singleton and Rafael Grote	● Free ○ Restricted
Client(s)	**Client's reference**
Norwegian Research Council	Project number 280573 'Advanced models and weather prediction in the Arctic: enhanced capacity from observations and polar process representations (ALERTNESS)

Abstract

This report describes work done as part of the Alertness product to assess the capabilities of the AROME-Arctic model as a high resolution Ensemble Prediction System (EPS). The AROME-Arctic EPS was run for two 3-week periods for winter (March 2018) and summer (July 2018) that coincided with the Year Of Polar Prediction Special Observing Periods 1 and 2. Forecasts from the AROME-Arctic EPS were verified against weather station observations and the verification scores compared with operational forecasts from the global IFSENS model run by the European Centre for Medium Range Weather Forecasts. The AROME-Arctic EPS model generally had better, or similar verification scores when compared with IFSENS for 2m temperature, 10m wind speed, 2m relative humidity and 12-hour precipitation accumulations. However, some biases in the perturbed ensemble members of the AROME-Arctic EPS were discovered - with a warm bias related to extremely cold weather and a dry bias in the summer. The scores shown in this report for AROME-Arctic EPS will act as a reference against which developments of the EPS through the Alertness project can be measured.

Keywords

Ensembles, Arctic, Verification, Weather Forecasting

Disciplinary signature

Responsible signature

Norwegian Meteorological Institute
Org.no 971274042
post@met.no
www.met.no / www.yr.no

Oslo
P.O. Box 43, Blindern
0313 Oslo, Norway
T. +47 22 96 30 00

Bergen
Allégaten 70
5007 Bergen, Norway
T. +47 55 23 66 00

Tromsø
P.O. Box 6314, Langnes
9293 Tromsø, Norway
T. +47 77 62 13 00

Contents

1 Introduction

The chaotic nature of the atmosphere means that there is inherent unpredictability. The ability of deterministic models to forecast the weather is therefore limited to how well it can model predictable scales and phenomena. Moreover, small errors will grow with increasing lead time (Lorenz 1963). In order to provide a complete forecast, estimates of the uncertainty must also be included. These estimates of uncertainty will help decision makers to make better informed decisions where the future is uncertain. In weather forecasting, ensemble prediction systems (EPS) have been developed in an attempt to model the uncertainty. Initially these systems were developed for the global scale with relatively low spatial resolution and long time horizons (Toth and Kalnay 1993; Molteni et al. 1996), but advances in computing power mean that EPSs are now being developed for convection permitting scales for short range forecasts (Hacker et al. 2011; Bouttier et al. 2012; Marsigli et al. 2014; Hagelin et al. 2017; Frogner et al. 2019).

There are many sources of uncertainty that are modelled by EPSs, chief among these are uncertainties in the model initial conditions. At the global scale, the EPS developed by the European Centre for Medium Range Weather Forecasts (ECMWF) uses a combination of singular vectors that identify perturbations that maximize error growth (Buizza and Palmer 1995) and ensembles of data assimilation (EDA) whereby observations used in the data assimilation are perturbed based on known error statistics (Buizza et al. 2008). Other methods for initial conditions perturbations exist, such as error breeding (Toth and Kalnay 1993) and the ensemble transform Kalman filter (Bowler et al. 2008). Convection permitting ensembles were initially made by simply downscaling selected members of a global EPS using a convection permitting model (e.g. Molteni et al. 2001), but more recent developments have included perturbations to the surface boundary conditions and EDA (Bouttier et al. 2016; Frogner et al. 2019).

Another significant source of uncertainty is errors due to the parameterization of sub grid scale physical processes, which are often based on empirical relationships - this is known as model uncertainty. One of the most used methods to model this uncertainty is to perturb the tendencies that come from the physics parameterizations, a process known as stochastically perturbed parameterization tendencies (SPPT). SPPT can be used at both global (Buizza et al. 1999) and convection permitting (Bouttier et al. 2012; Frogner et al. 2019) scales. SPPT is a very indirect way of accounting for the uncertainties in the model parameterization schemes and a new method has been developed whereby uncertain parameters within the parameterization schemes themselves are perturbed. This method is known as stochastically perturbed parameterizations (SPP) and has shown promising results (Ollinaho et al. 2017).

In the Arctic, the observing system is relatively sparse and parameterization schemes are typically more uncertain since they are often based on empirical relationships obtained for less extreme climates. Furthermore, it has been shown that sea surface temperature (SST) products derived from satellite are particularly uncertain in the polar regions (Liu and Minnett 2016), and that modelling the atmospheric boundary layer is very sensitive to uncertainties in sea ice concentration (Seo and Yang 2013). All of this means that uncertainties in the Arctic region are likely to be considerable and that it is vitally important to model them as well as possible in order to provide forecasts with uncertainty estimates that allow forecast users to take the uncertainty into account. Indeed it has been shown that a simple downscaling of a global ensemble in

the Arctic region can provide better warnings than the global EPS in the case of a severe polar low event (Kristiansen et al. 2011).

Work Package 4 in ALERTNESS is dedicated to the development of a convection permitting EPS for the Arctic. The purpose of this report is to provide a reference for an EPS implementation of the the AROME-Arctic model against which the outcomes of model developments within the work package can be measured.

2 Reference model description

The model used in ALERTNESS is based on the operational implementation of the AROME-Arctic model (Müller et al. 2017), which is an implementation of the Harmonie-AROME model (Bengtsson et al. 2017) for the Arctic region around Norway. This model is put into the Harmonie EPS system (HarmonEPS: Frogner et al. 2019) to make ensemble forecasts. In addition to the features described in the aforementioned references, we have taken advantage of recent developments in the operational implementation of the AROME-Arctic model, such as the inclusion of modelling of snow on ice surfaces.

For ALERTNESS, we run the AROME-Arctic EPS with a horizontal grid length of 2.5 km on the domain shown in Fig 1 and use one control member and ten perturbed members. The setup of the EPS is very much as described in Frogner et al. (2019), though the salient features are described below.

Perturbations to the initial and lateral boundary conditions are provided using the scaled lagged average forecast (SLAF) approach, whereby differences between lagged forecasts with the same validity time taken from the operational integrated forecasting system (IFS) of ECMWF are added to, and subtracted from, the AROME-Arctic control member. The perturbations are scaled using the total energy norm (Keller et al. 2008) to ensure that each perturbation has roughly the same impact on the forecast. Upper air observations are assimilated for the control member only using 3DVAR - this includes radiosonde observations and all available satellite observations. Surface processes are modelled using SURFEX (Masson et al. 2013), which divides the surface into 4 tiles - nature, town, sea and inland water bodies, each with their own physics. Sea ice is modelled using a simple thermodynamic scheme (Batrak et al. 2018). Observations of 2m temperature and 2m relative humidity are assimilated in SURFEX for both the control and perturbed members using optimal interpolation. Perturbations of spatially correlated noise, with a specified standard deviation that depends on the parameter and a correlation length scale of 150 km, are applied to a number of surface variables. The spatially correlated noise perturbations are either added to the control member or multiplicatively scaled. The surface perturbations are summarised in Table 1.

Forecasts were run during the reference periods described in the next section to a lead time of 48 hours every 24 hours initialized at 00:00 UTC. Short forecasts to a lead time of 3 hours were run every 3 hours for the data assimilation cycling. For the purposes of this report, the model setup described here is referred to as ALERTNESS_ref.

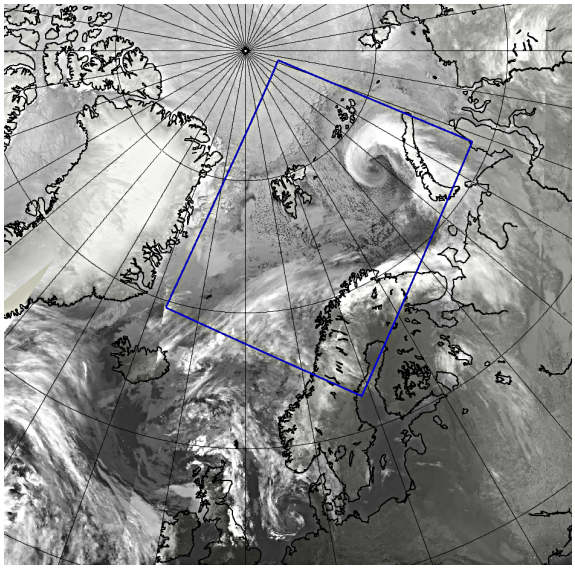

Figure 1: The domain used for ALERTNESS

Table 1: The magnitude and type of perturbation applied to the surface parameters. For type, × means that the perturbations are multiplicative and + means that the perturbations are additive.

Parameter	Standard deviation	Type
Vegetation fraction	0.1	x
Leaf area index	0.1	x
Thermal coefficient of vegetation	0.1	x
Surface roughness length over land	0.2	x
Albedo	0.1	x
Sea surface temperature	0.25	+
Soil temperature	1.5	+
Soil moisture	0.1	x
Snow depth	0.5	x
Surface fluxes over sea	0.2	x

3 Reference periods

Periods of 3 weeks during the The Year Of Polar Prediction (YOPP) Special Observing Periods (SOP) 1 and 2 were used to assess the performance of ALERTNESS_ref. The period used in SOP 1 was from 8 - 31 March 2018 and gives information about the performance of the model during the winter / transitioning into spring season, and the period used for SOP 2 was 10 July - 1 August 2018, thus providing details of the model performance during the summer season.

Time series of the mean, minimum and maximum values of 2m temperature, 10m wind speed, 2m relative humidity and 12 hour accumulated precipitation, taken from all of the available observations stations in the AROME-Arctic domain, were made to give an indication of the prevailing weather conditions during the two reference periods. SOP 1 (Fig. 2) was characterized by generally cold temperatures with the mean 2m temperature not going above $0°C$, the minimum as low as $-35°C$ and the maximum being close to $0°C$ for much of the period. Mean wind speeds were generally around 5 m/s, with the maximum typically around 15 m/s. There was a period around 17 - 19 March 2018 when the wind speed was at its maximum and this coincided with the time at which the maximum rainfall occurred.

During SOP 2 (Fig. 3), temperatures were particularly warm with day time mean temperatures in excess of $20°C$ and day time maximum temperatures higher than $30°C$. The wind speed was generally low with the maximum rarely above 15 m/s and the mean generally below 5 m/s. There were a number of heavy rain events observed during the period with largest maximum 12h precipitation accumulation of around 60mm observed on 29 July.

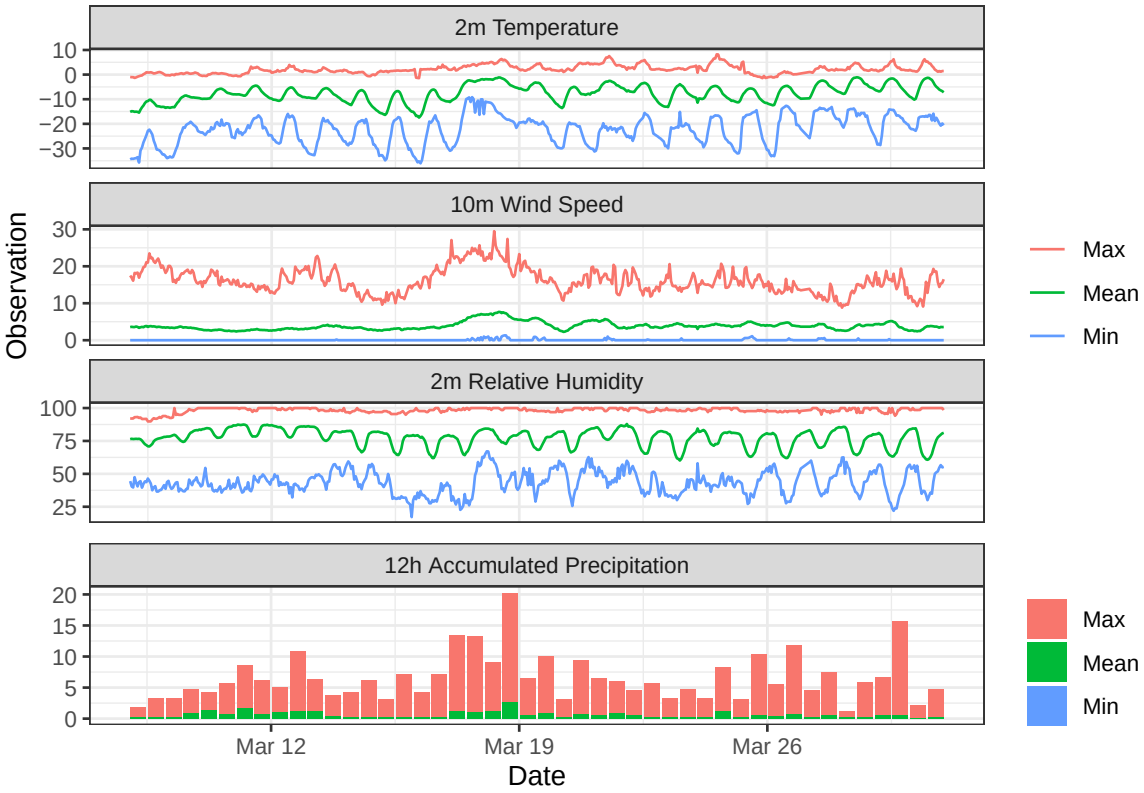

Figure 2: Time series of mean, maximum and minimum values of selected obseravations during SOP 1

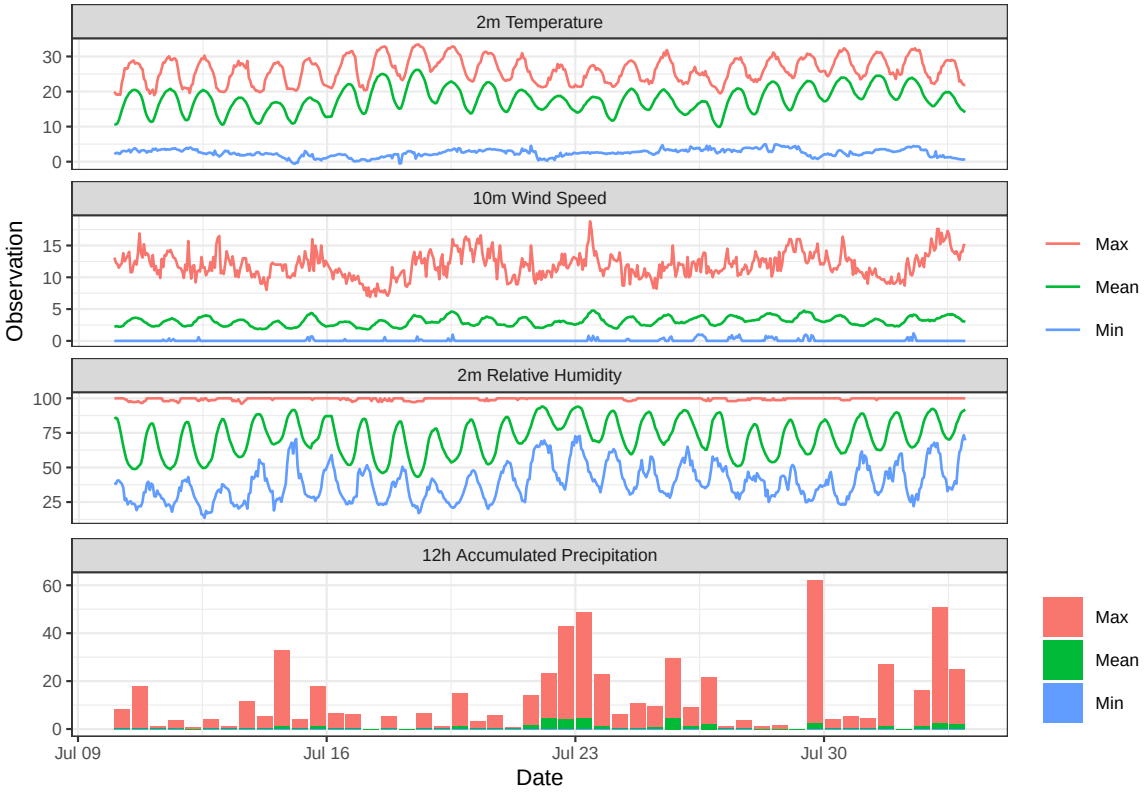

Figure 3: Time series of mean, maximum and minimum values of selected obseravations during SOP 2

4 Verification methodology

The performance of ALERTNESS_ref is measured against that obtained from the global EPS from ECMWF, IFSENS, using the same time period and observation stations. Model performance is measured using the following objective verification scores, which are described briefly here, but in more detail in Wilks (2011).

- The root-mean-square error (RMSE) of the ensemble mean of the forecast compared with observations.
- The ensemble spread, which is the standard deviation of the ensemble members around the ensemble mean. This reflects the uncertainty in the forecast that the ensemble is able to model. For a well calibrated ensemble, the ensemble spread should be equal to the RMSE.
- The continuous rank probability score (CRPS), which measures the distance of a continuous distribution function constructed from the ensemble forecast to the observed value. For a single ensemble member the CRPS reduces to the mean absolute error of the forecast. It is therefore negatively oriented with a perfect score being zero.
- Rank histograms (sometimes referred to as Talagrand diagrams), which show the distribution of observations into bins of ranked ensemble members. The shape of the rank histogram gives an indication of under (u shaped) or over (convex shaped) spread, or negative (weighted toward the right) or positive (weighted toward the left) bias. Here, the count of observations in each bin is given as the normalized frequency such that an ensemble with perfect spread would have a normalized frequency of 1, and the rank is given relative to the number of members in the ensemble, such that the rank is always between 0 and 1. Using the relative rank means that ensembles with different numbers of members can easily be compared.
- The Brier Score integrates the probabilistic distance, for a given threshold, between each member of the ensemble forecast and the binary probability of the observation. Its skill score, the Brier Skill Score, compares the Brier Score of the forecast with that of a reference forecast. Here we use the sample climatology as the reference forecast. A Perfect Brier Skill Score has a value of 1 and a value less than zero indicates no skill compared with the reference.
- The reliability measures the accuracy of the forecast probability for a threshold by comparing the forecast probability with the observed frequency. For a perfectly reliable forecast the observed frequency will be equal to the forecast probability. Note that reliability is conditioned on the forecast probability.
- The Relative Operating Characteristics (ROC) compare the hit rate and the false alarm rate for observed events for each probability. For a skillful forecast the ROC curve will be well above the diagonal with a high hit rate and a low false alarm rate. Note that the ROC is conditioned on the occurrence of events.
- The economic value of the forecast shows the increase in economic value, for a given threshold, over a reference forecast, here we use the sample climatology as with the Brier Skill Score, for users with a range of cost-loss ratios between 0 and 1. It is predicated on the fact that users will take protective action when the probability of the event is higher than or equal to the cost-loss ratio.

5 Results

5.1 SOP1

5.1.1 2m temperature

Forecasts of 2m temperature are verified against all available stations within the AROME-Arctic domain. These stations are shown in Fig. 4. Forecasted 2m temperatures are adjusted for differences in height between the model elevation and the station elevation (shown in Fig. 4) using a simple lapse rate of $6.5°C/km$.

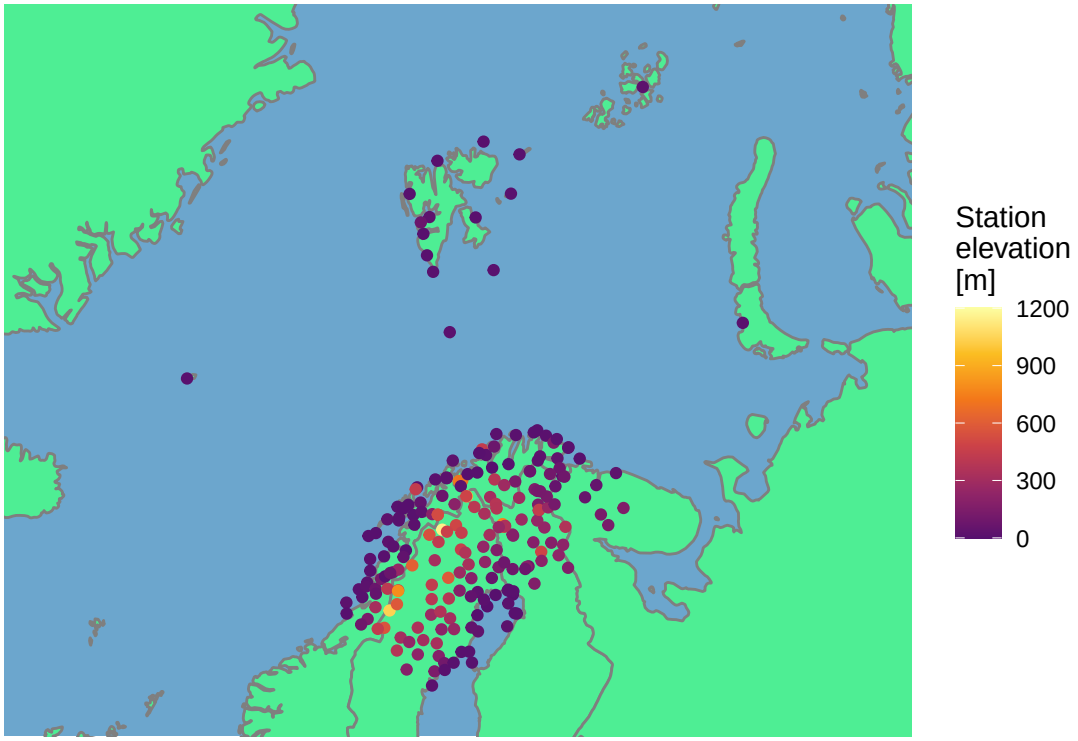

Figure 4: Stations used for 2m temperature verification.

Summary scores for SOP1 are shown in Fig. 5, comparing scores obtained from ALERTNESS_ref with those from IFSENS. Since each model run was started at 00 UTC, the time of day can be inferred from the lead time on the x-axis. It is immediately apparent that both ALERTNESS_ref and IFSENS have a distinct diurnal cycle in the forecast skill for 2m temperature. Both have larger RMSE (Fig. 5(a)) and CRPS (Fig. 5(b)) during the night time compared with the day time. ALERTNESS_ref has a larger spread, that remains broadly constant, throughout the forecast, while the spread for IFSENS slowly increases throughout the forecast (Fig. 5(a)). The CRPS for ALERTNESS_ref is lower than for IFSENS up to approximately hour 27 of the forecast, and is clearly lower during the day time (Fig. 5(b)). The superior skill for ALERTNESS_ref over IFSENS during the day time is also reflected in the RMSE (Fig. 5(a)). The bias of the ensemble mean (Fig. 5(c)) suggests a much stronger diurnal cycle in the 2m temperature forecast errors for IFSENS than for ALERTNESS_ref, with IFSENS showing a warm bias during the night and a cold bias during the day,

and ALERTNESS_ref showing a warm bias throughout that becomes weaker during the day time. The rank histogram (Fig. 5(d)) adds weight to the suggestion from Fig. 5(a) that ALERTNESS_ref has better spread for 2m temperature than IFSENS, with many of the observation ranks having a normalized count close to 1. Note that the logarithmic scale in 5(d)) under-emphasises the differences between ALERTNESS_ref and IFSENS.

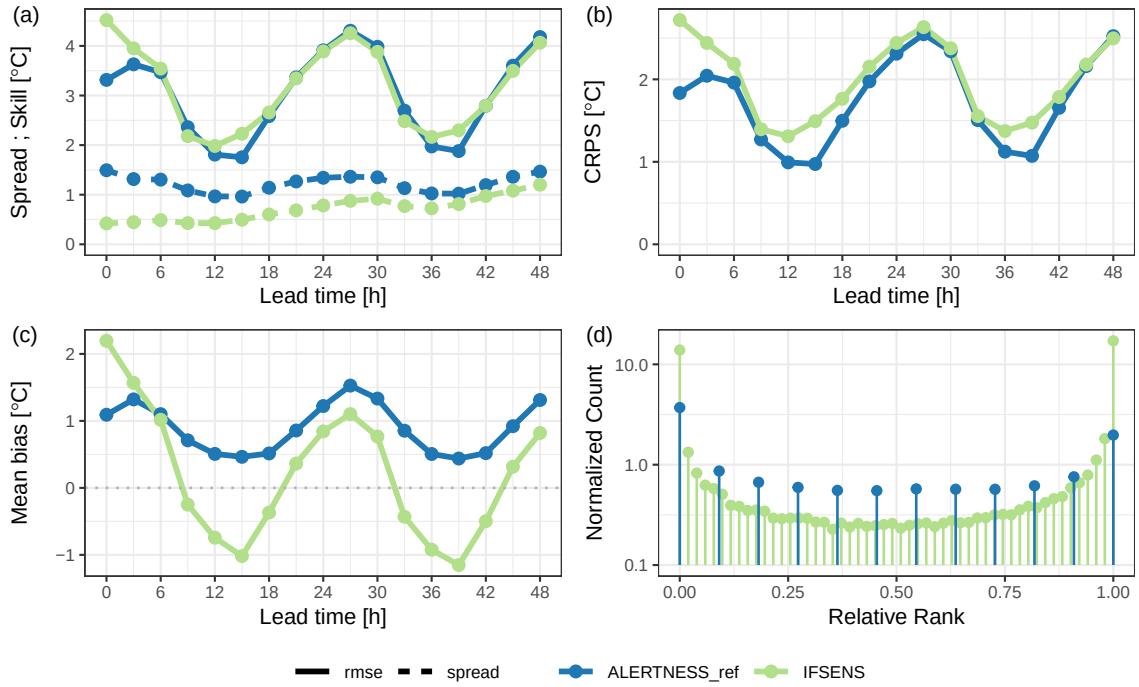

Figure 5: Summary verification scores for 2m temperature during SOP 1: (a) RMSE and spread, (b) CRPS, (c) Bias of the ensemble mean and (d) Normalized relative rank histogram.

To obtain further insight into the performance of the models for different temperatures, categorical scores were computed for different 2m temperature thresholds. In order to maintain a relatively consistent number of observations for each lead time, the thresholds were chosen to be the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed 2m temperatures that were available for each lead time. The thresholds used for each lead time are shown in Fig. 6. It is clear that there is a strong diurnal cycle on the coldest threshold that becomes weaker as the thresholds become warmer.

Fig. 7 shows the Brier Skill Score for each of the thresholds. The reference climatology used to compute the Brier Skill Score is the sample climatology obtained from the observations. For all thresholds, ALERTNESS_ref has a higher Brier Skill Score than IFSENS, particularly during the day. The strong diurnal cycles seen in the Brier Skill Score for both models for the lower percentiles is likely reflective of the diurnal cycles in the thresholds obtained for those percentiles.

The reliability and ROC are perhaps more useful for forecast users in helping to understand the quality of EPS forecasts. Fig. 8 shows the forecast reliability and ROC at 12 hours lead time and Fig. 9 the reliability and ROC at 24 hours lead time for ALERTNESS_ref and IFSENS for all percentiles. Results are only shown for lead times of 12 and 24 hours to be representative of day time and night time. Note that the results for

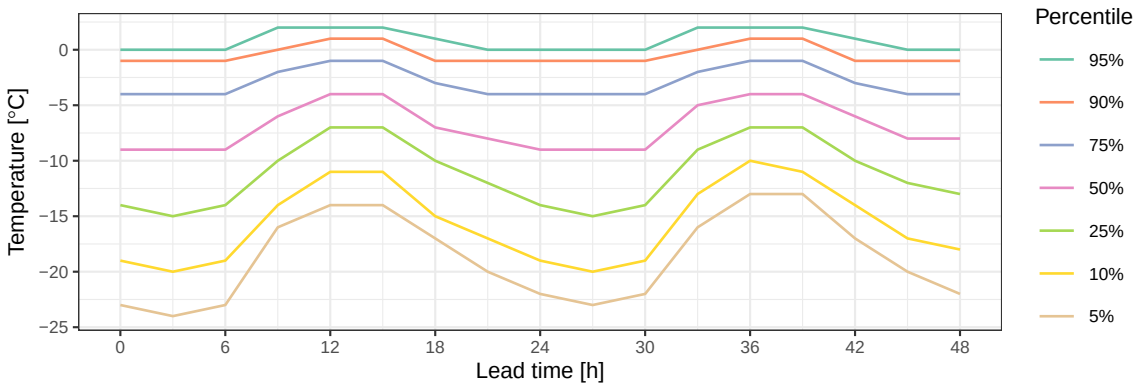

Figure 6: Thresholds used for categorical scores for 2m temperature during SOP 1 derived from the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values valid at each lead time.

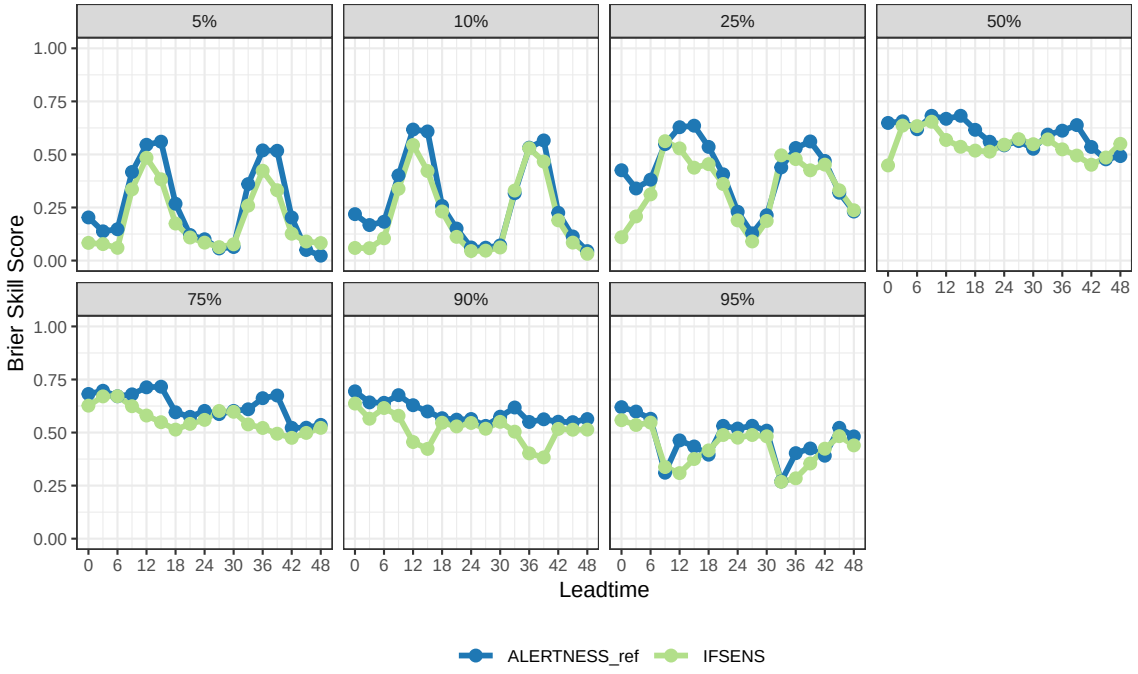

Figure 7: Brier Skill Score for 2m temperature during SOP 1 for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values. The sample climatology is used as the reference forecast.

lead times of 36 and 48 hours are broadly similar to those for 12 and 24 hours.

At 12 hours lead time, ALETRNESS_ref typically over forecasts the probabilities related to the colder percentiles, while IFSENS under forecasts the probabilities 8(a). Note that the probabilities are for forecasts greater than the threshold and thus over forecasting probabilities for 2m temperatures greater than the threshold means that temperatures colder than the threshold are being forecasted with probabilities that are too low. As the thresholds become warmer, ALERTNESS_ref becomes more reliable (closer to the diagonal), while IFSENS continues to under forecast the probabilities. The ROC 8(b) suggests that ALERTNESS_ref produces more false alarms than IFSENS, especially for the lower percentiles. Conversely this means fewer false alarms for events colder than the lower percentiles. As the thresholds become warmer, the hit rate for ALERTNESS_ref is considerably better than for IFSENS.

For 24 hours lead time, both ALERTNESS_ref and IFSENS struggle to be reliable for the lowest percentiles (9(a)), with noisy plots suggesting few forecasts below the thresholds. As temperatures become warmer, both models are similarly reliable, though in general ALERTNESS_ref forecasts lower probabilities than IFSENS. The ROC (9(b)) suggests a very large false alarm rate for both models for the colder thresholds. This means that when cooler temperatures occur, the models are typically forecasting temperatures higher than the threshold. It could be argued that, for the 5th and 10th percentiles, neither model has skill since the ROC curves are very close to the diagonal. As the thresholds become warmer the false alarm rate decreases along with the hit rate, with ALERTNESS_ref having a slightly higher hit rate than IFSENS.

The verification score that may be of most use to end users of probabilistic forecasts is the economic value. 10 shows the economic value for 2m temperature for forecast from ALERTNESS_ref and IFSENS during SOP 1 for lead times of 12 and 24 hours. It is clear that at a lead time of 12 hours (10(a)), ALERTNESS_ref offers more value to end users than IFSENS where decisions are based on 2m temperature. Most striking, is that for all percentiles, ALERTNESS_ref can provide more value for users with lower cost-loss ratios. For a lead time of 24 hours (10(b)), both models offer a similar level of value to users with all cost-loss ratios, while neither model provides much value for the lowest percentiles.

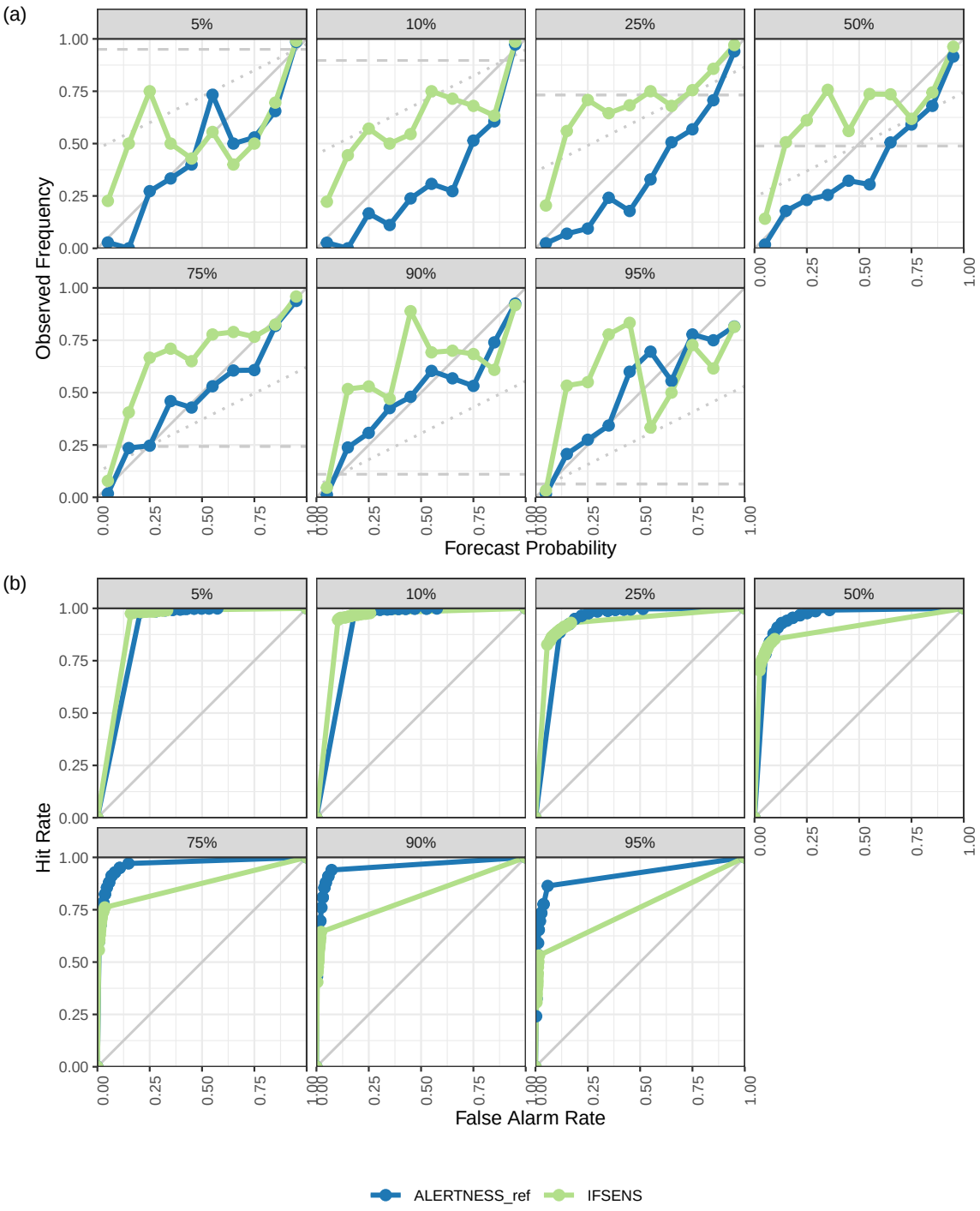

Figure 8: Verification for 2m temperature during SOP 1 at a lead time of 12 hours and for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values for (a) reliability and (b) ROC.

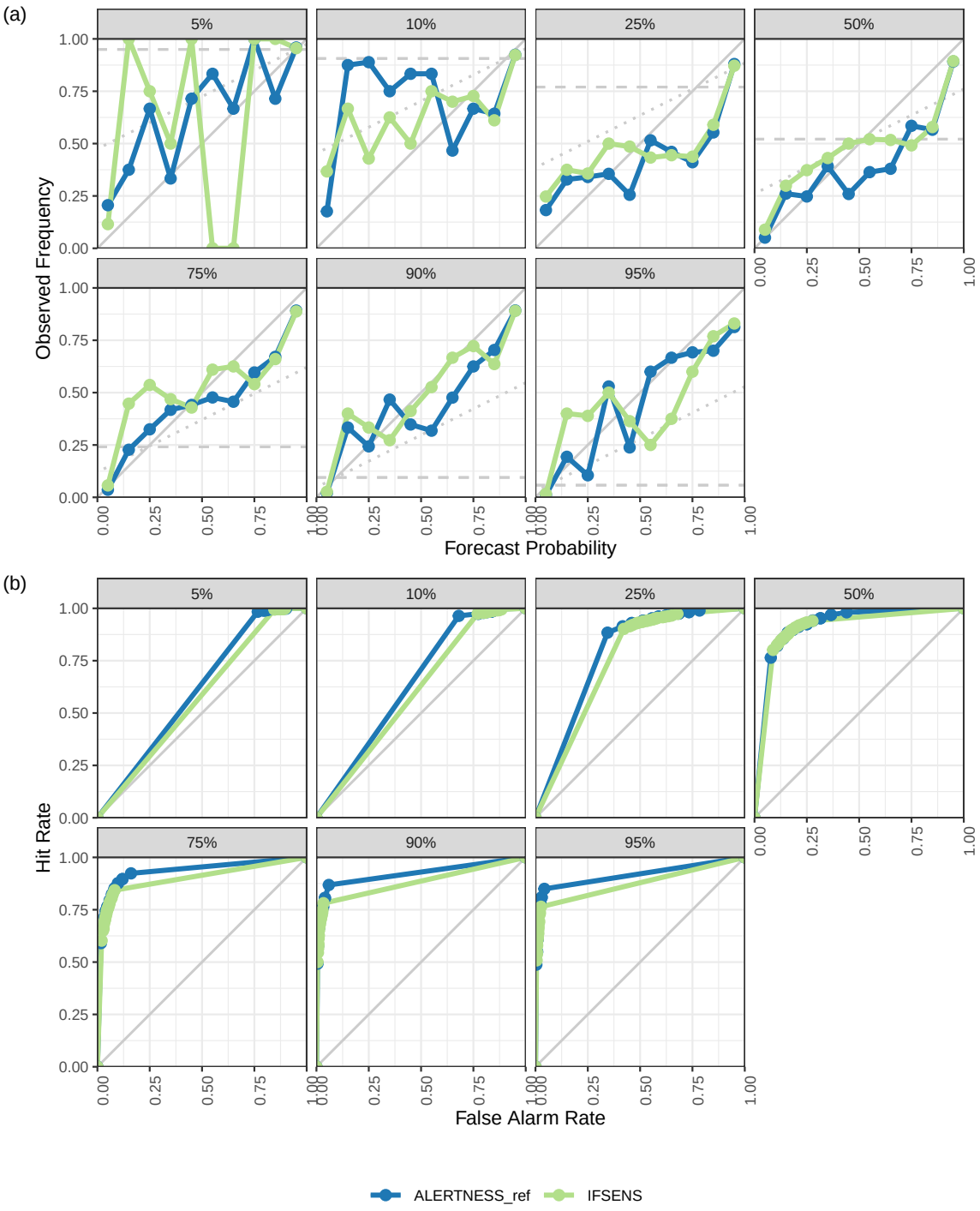

Figure 9: Verification for 2m temperature during SOP 1 at a lead time of 24 hours and for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values for (a) reliability and (b) ROC.

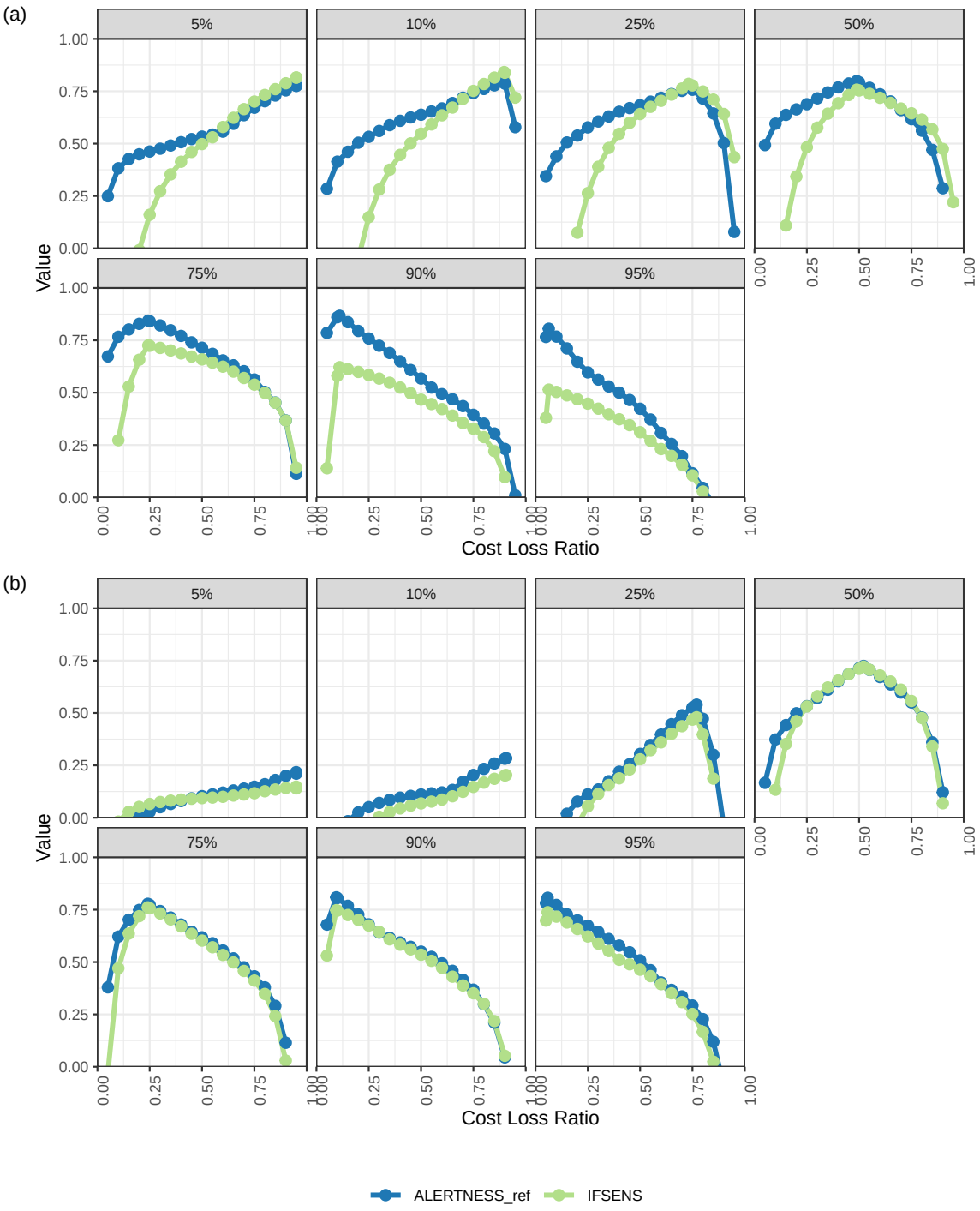

Figure 10: Economic value for 2m temperature forecasts during SOP 1 for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values at lead times of (a) 12 hours (b) 24 hours.

5.1.2 10m Wind Speed

Forecasts for 10m wind speed are verified for all available stations with observations inside the AROME-Arctic domain. These stations are shown in Fig. 11. No adjustments are made for differences between model elevation and station elevation. While the vast majority of stations are over land, there are also a large number of coastal stations and some offshore stations. Further verification against satellite derived winds over the sea will be possible in the future following developments from WP1.

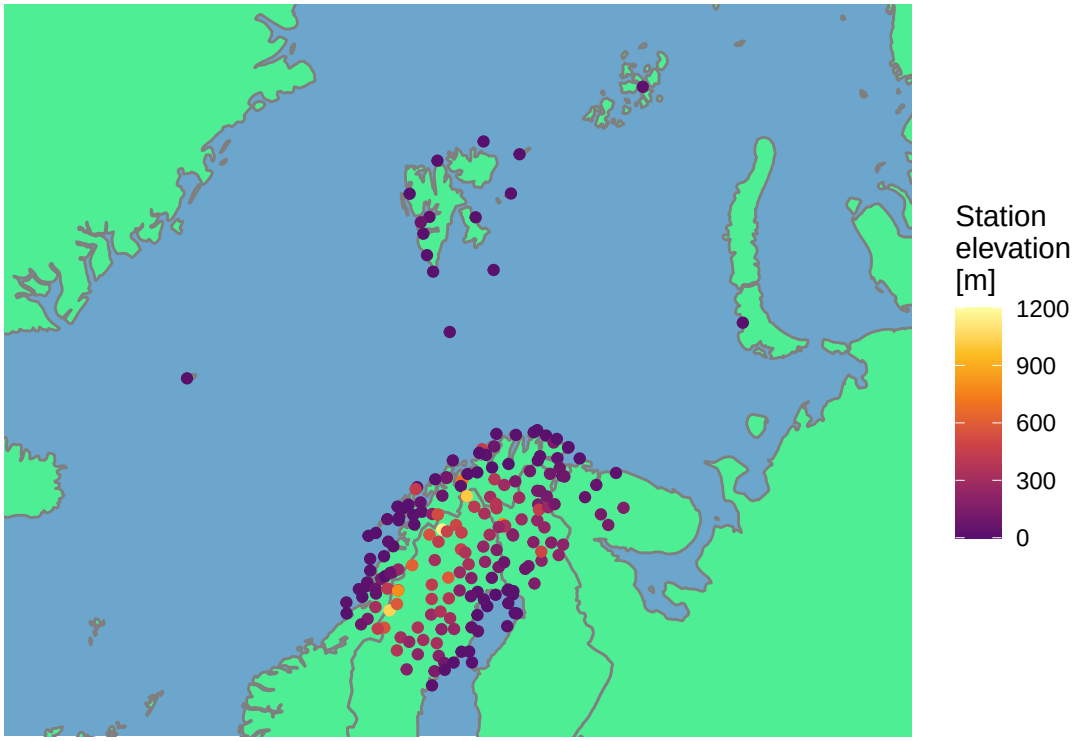

Figure 11: Stations used for 10m wind speed verification.

Summary scores for SOP1 are shown in 12, comparing scores obtained from ALERTNESS_ref with those from IFSENS. Unlike for 2m temperature, the diurnal cycle in the verification is not so pronounced except for the bias of the ensemble mean for IFSENS 12(c). ALERTNESS_ref has lower RMSE (12(a)) and CRPS (12(b)) than IFSENS throughout the 48 hours of the forecasts, though these differences become smaller with increasing lead time. Most striking is that the spread for 10m wind speed is considerably larger for ALERTNESS_ref than for IFSENS (12(a)) throughout the forecast. While the spread for ALERTNESS_ref grows slowly through the forecast, there is a much more pronounced, almost linear, growth in spread for IFSENS. The bias of the ensemble mean (12(c)) suggests that ALERTNESS_ref has a tendency to over forecast 10m wind speeds that becomes slightly weaker during the day, and IFSENS has a strong diurnal cycle in the bias, slightly over forecasting 10m wind speeds during the night time and strongly under forecasting wind speeds during the day time. The rank histogram (12(d)) shows that both ALERTNESS_ref and and IFSENS do not posses sufficient spread, with the normalized observation counts generally being closer to 1 for ALERTNESS_ref.

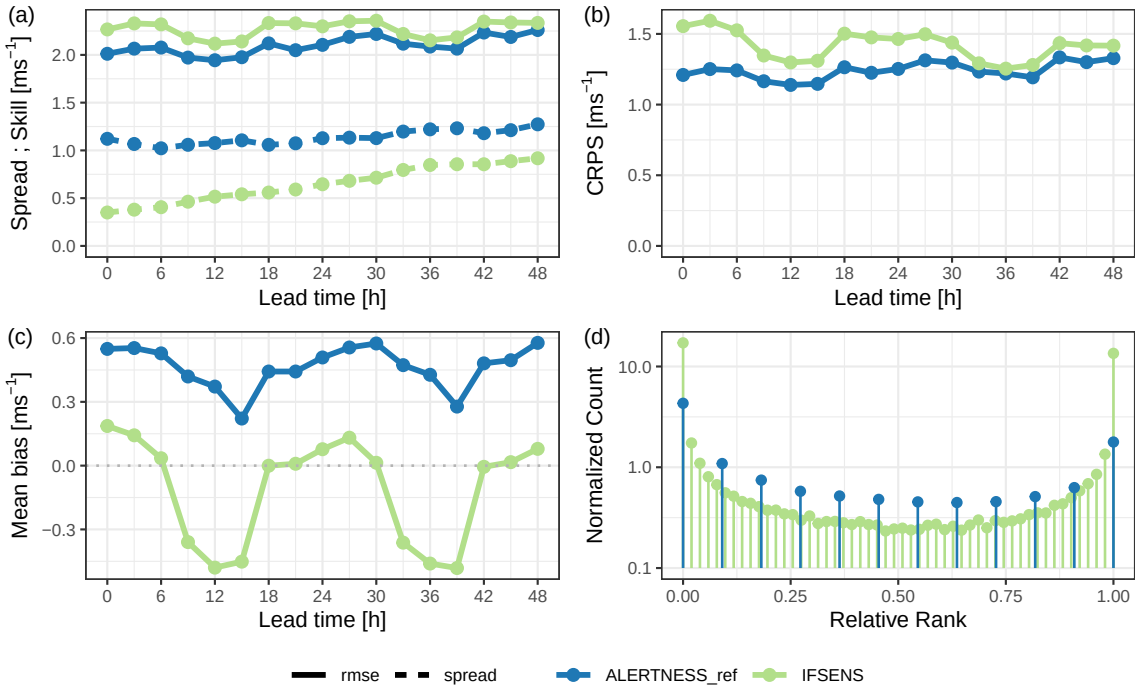

Figure 12: Summary verification scores for 10m wind speed during SOP 1: (a) RMSE and spread, (b) CRPS, (c) Bias of the ensemble mean and (d) Normalized relative rank histogram.

As for 2m temperature, the performance of the models is assessed for different thresholds using percentiles of the wind speed observations available for each lead time. However, only the 50th, 75th, 90th and 95th percentiles are considered. The values of these thresholds are shown in Fig. 13. The obtained thresholds were rounded to the nearest ms^{-1}, which meant that in some cases, for percentiles lower than the 50th, this resulted in the same value for different percentiles. There is an indication of a weak diurnal cycle in the thresholds with slightly stronger 10m wind speeds during the day for the 50th and 75th percentiles, and the opposite for the 95th percentile. This approach using percentiles means that we do not verify any extreme wind speed cases (the 95th percentile is around 10 ms^{-1}), but it also means that there are not a sufficient number of cases to obtain meaningful verification scores for those extreme wind speeds.

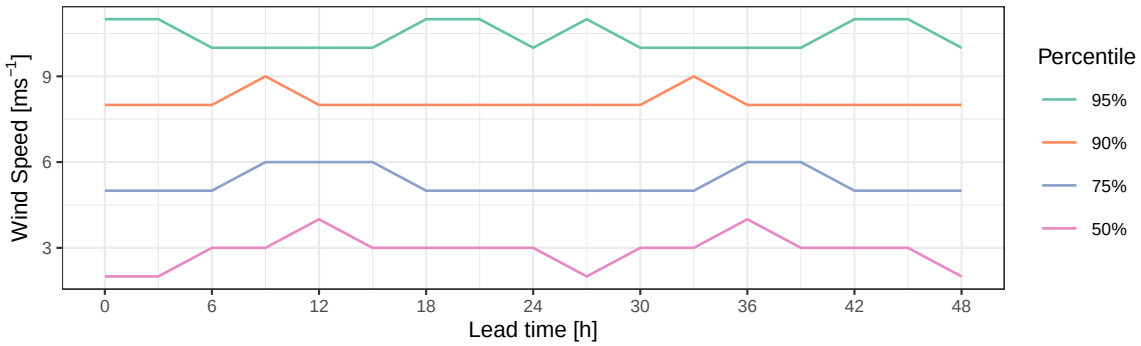

Figure 13: Thresholds used for categorical scores for 10m wind speed during SOP 1 derived from the 50th, 75th, 90th and 95th percentiles of the observed values valid at each lead time.

Fig. 14 shows the Brier Skill Score for each of the thresholds using the sample climatology as reference. For the 50th percentile the Brier Skill Score appears, to some extent, to be influenced by the diurnal cycle in the value of the 50th percentile of 10m wind speed, dropping to close to zero for both ALERTESS_ref and IFSENS for the lowest values of the threshold. For the higher thresholds, both models have lower Brier Skill Scores as the wind speed threshold increases from the 75th to the 95th percentile. IFSENS has a relatively consistent Brier Skill Score throughout the 48 hours of the forecasts, while the Brier Skill Score for ALERTNESS_ref is higher at the beginning of the forecast and drops as the lead time increases. For the 75th percentile the Brier Skill Score for ALERTNESS_ref converges with that for IFSENS at around the 30th hour of the forecast, but for the 95th percentile this convergence occurs at around the 12th hour.

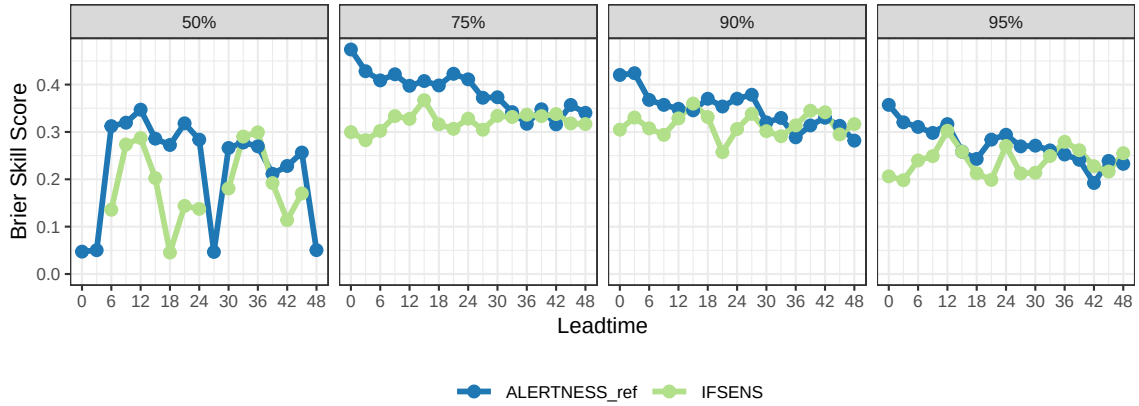

Figure 14: Brier Skill Score for 10m wind speed during SOP 1 for the 50th, 75th, 90th and 95th percentiles of the observed values The sample climatology is used as the reference forecast.

The reliability and ROC for 10m wind speed for the 50th, 75th, 90th and 95th percentiles of observed 10m wind speed at a lead time of 12 hours are shown in Fig. 15. ALERTNESS_ref tends to forecast the lower probabilities for all thresholds reliably, but over forecast the higher probabilities (Fig. 15(a)). IFSENS under forecasts the lower probabilities for the 50th, 75th and 95th percentiles, but for the 90th percentile is reliable up to 40% probability. Overall there is little systematic difference between the reliabilities of ALERTNESS_ref and IFSENS. For the ROC, ALERTNESS_ref has higher hit rates than IFSENS for all percentiles (Fig. 15(b)), with the difference between them least pronounced for the 50th percentile. There is drop in both hit rate and false alarm rates for both models as the percentile of the threshold is increased. Similar results were seen for 24, 36 and 48 hour lead times and are thus not shown here.

The economic value for the same thresholds is shown in Fig. 16 for lead times of 12, 24, 36 and 48 hours. For most percentiles at most lead times, ALERTNESS_ref offers more value than IFSENS to users with a larger range of cost-loss ratios. As lead time increases, the difference in the value curves between ALERTNES_ref and IFSENS becomes smaller as the value offered by ALERTNESS_ref becomes smaller and that offered by IFSENS remains broadly unchanged. This is most notable for users with the lower cost-loss ratios for decisions related to the higher 10m wind speed thresholds, where ALERTNESS_ref offers considerably more value than IFSENS during the first day of the forecast.

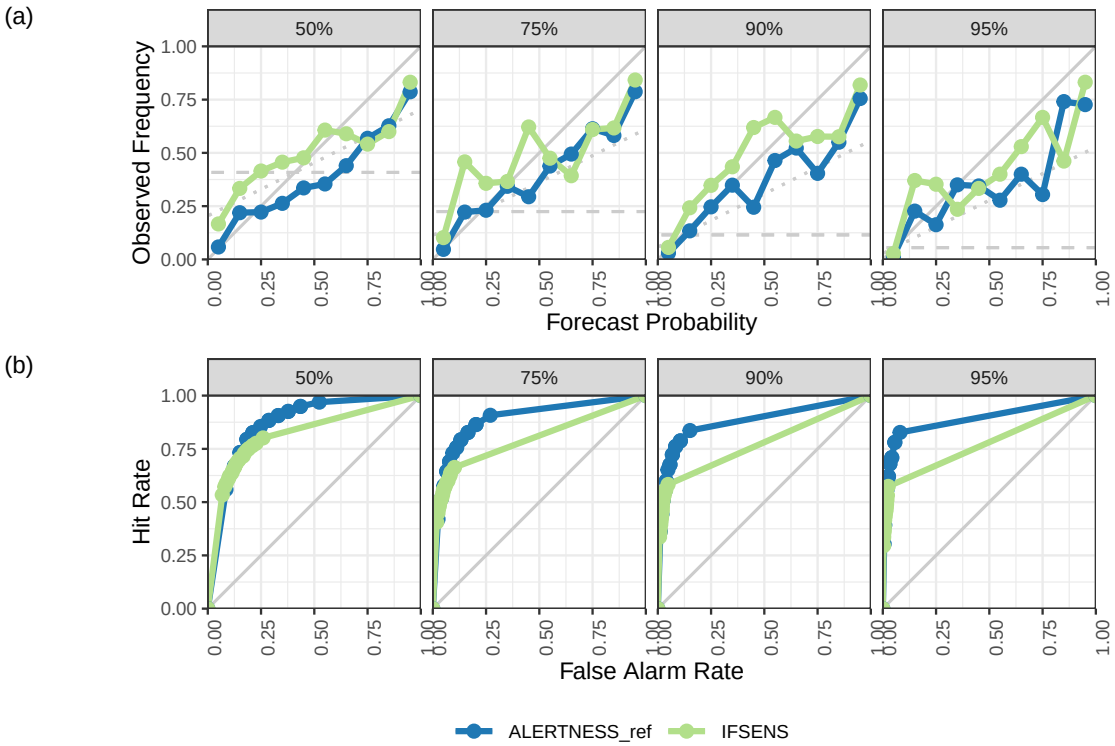

Figure 15: Verification for 10m wind speed during SOP 1 at a lead time of 12 hours and for the 50th, 75th, 90th and 95th percentiles of the observed values for (a) reliability and (b) ROC.

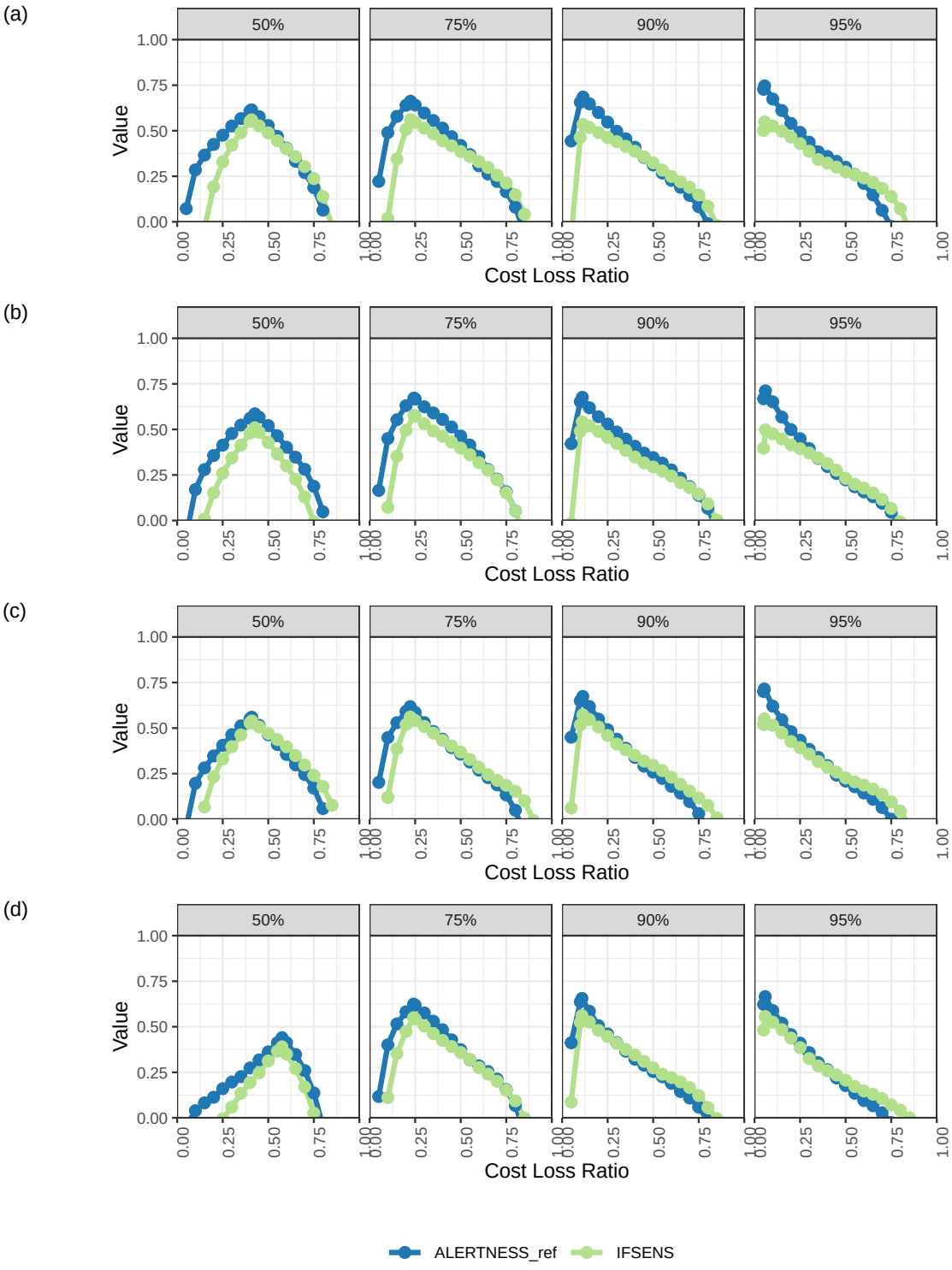

Figure 16: Economic value for 10m wind speed forecasts forecasts during SOP 1 for the 50th, 75th, 90th and 95th percentiles of the observed values at lead times of (a) 12 hours, (b) 24 hours, (c) 36 hours and (d) 48 hours.

5.1.3 2m Relative Humidity

Forecasts for 2m relative humidity are verified for all available stations with observations inside the AROME-Arctic domain. These stations are shown in Fig. 17. No adjustments are made for differences between model elevation and station elevation.

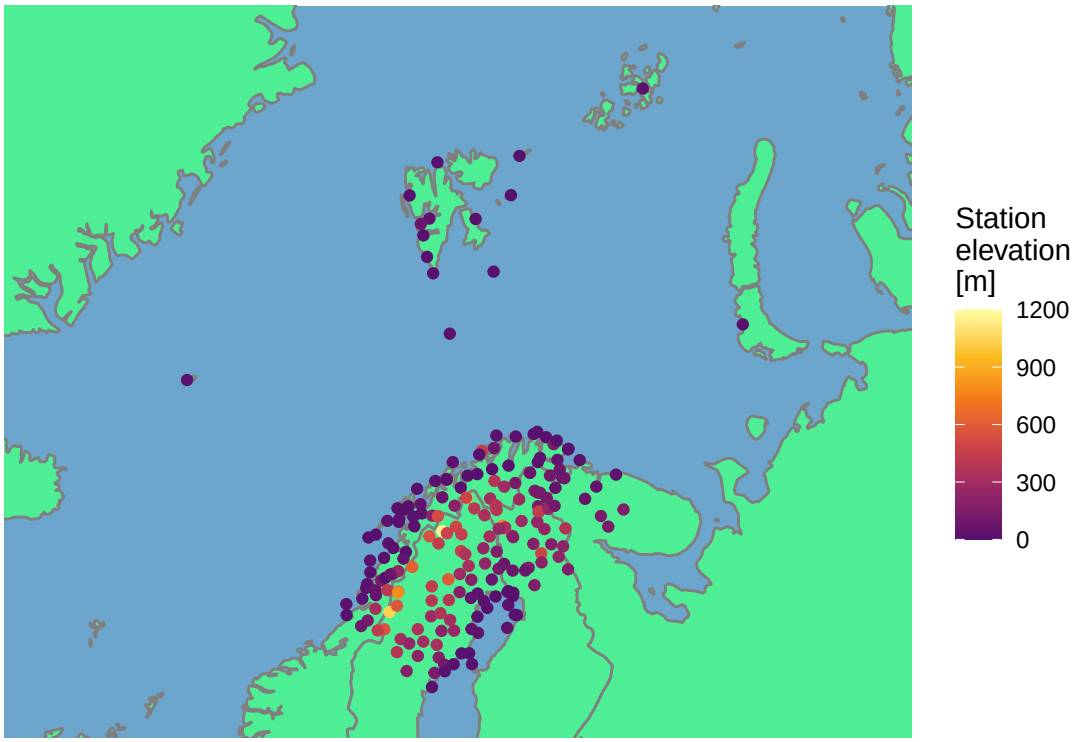

Figure 17: Stations used for 2m relative humidity verification.

Summary scores for SOP1 are shown in 18, comparing scores obtained from ALERTNESS_ref with those from IFSENS. Unlike for both 2m temperature and 10m wind speed, ALERTNESS_ref is not clearly superior to IFSENS. There is a clear diurnal cycle in the RMSE (Fig. 18(a)) and CRPS (Fig. 18(b)), though there is a 3 - 6 hour offset between IFSENS and ALERTNTESS_ref. There is no systematic difference between ALERTNESS_ref and IFSENS in terms of RMSE (Fig. 18(a)) and CRPS (Fig. 18(b)). The spread for ALERTNESS_ref has a clear diurnal cycle and is larger than that for IFSENS, which increases almost linearly throughout the length of the forecast. The ensemble mean of ALERTNESS_ref shows a moist bias throughout the forecast with maxima during the afternoon / early evening hours, whereas IFSENS initially has a dry bias that is quickly removed resulting in a close to zero bias for the rest of the forecast (Fig. 18(c)). The rank histogram suggests that ALERTNESS_ref is slightly better dispersed than IFSENS for 2m relative humidity, though there is a signal of the moist bias in ALERTNESS_ref with more observations having lower ranks (Fig. 18(d)).

The performance of the models is further assessed for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed 2m relative humidity for each lead time. The evolution of the thresholds for these percentiles with lead time is shown in Fig. 19, where a clear diurnal cycle can be seen in the thresholds, with lower

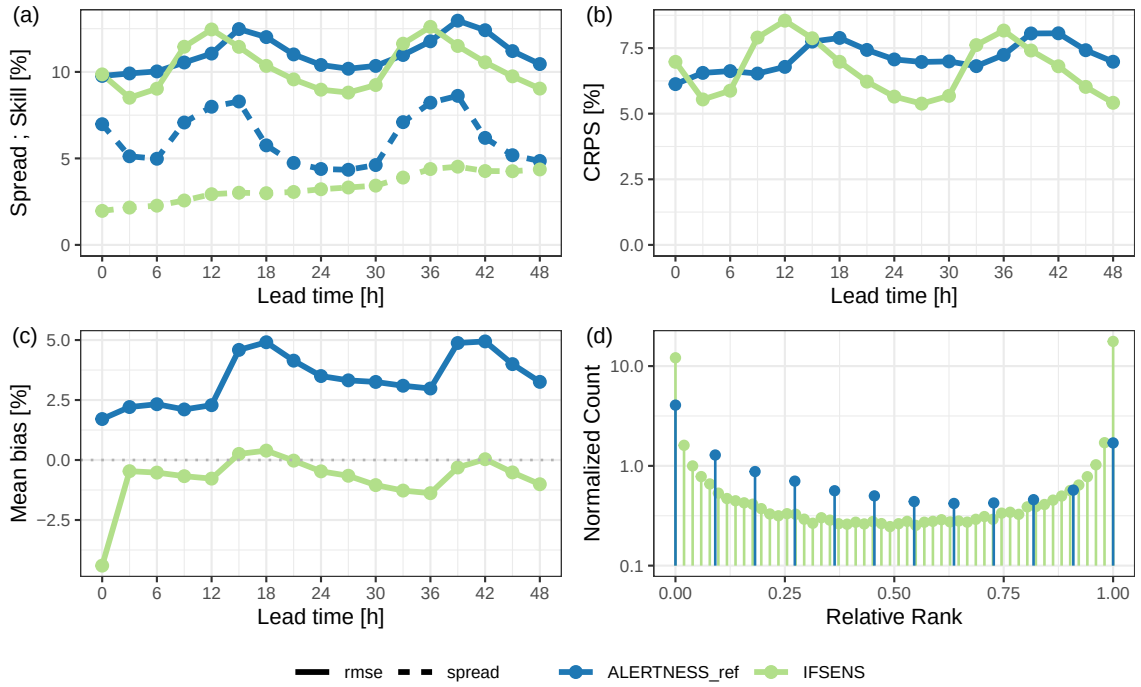

Figure 18: Summary verification scores for 2m relative humidity during SOP 1: (a) RMSE and spread, (b) CRPS, (c) Bias of the ensemble mean and (d) Normalized relative rank histogram.

relative humidity in the day time hours than the night time hours. The diurnal cycle becomes less pronounced for the higher percentiles.

Fig. 20 shows the Brier Skill Score for each of the thresholds using the sample climatology as reference. It is clear that both ALERTNESS_ref and IFSENS score poorly for all percentiles, though day time scores are better than those for the night time. For the the 90th and 95th percentiles, ALERTNESS_ref has no skill compared to the sample climatology, whereas IFSENS is able to maintain a Brier Skill Score of just above zero for most of the forecast. It is only really for the 75th percentile that ALERTNESS_ref is clearly superior to IFSENS in terms of the Brier Skill Score, and that is only between lead times 9 - 15 hours and 33 - 39 hours.

Fig. 21 shows reliability and ROC at 12h. In terms of reliability at 12 hours lead time (Fig. 22(a)), ALERTNESS_ref is slightly more reliable than IFSENS for the lower percentiles, but the reliability curves for both models become flatter and move towards no skill for the 90th and 95th percentiles. The ROC performance suggests that ALERTNESS_ref is superior to IFSENS for all percentiles (Fig. 22(b)) with lower false alarm rates for the lower percentiles and higher hit rates for the higher percentiles. However, the hit rate for ALERTNESS_ref does not exceed 50% for the 95th percentile. At 24 hours lead time (Fig. 22) the reliability curves for both models are fairly flat (Fig. 22(a)) with under forecasted probabilities for the lower percentiles and severely over forecasted probabilities for the higher percentiles. For the 75th and 90th percentiles IFSENS shows some reliability while the curves for ALERTNESS_ref are almost flat. The ROC curves (Fig. 22(b)) suggest very similar abilities to discriminate between events and non events for both ALERTNESS_ref and IFSENS, though for the 75th and higher percentiles, ALETRNESS_ref has a higher

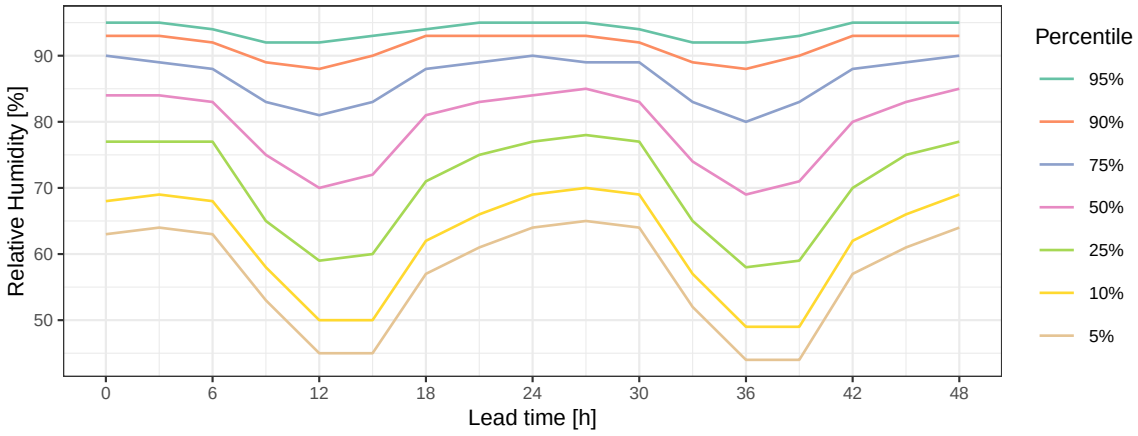

Figure 19: Thresholds used for categorical scores for 2m relative humidity during SOP 1 derived from the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values valid at each lead time.

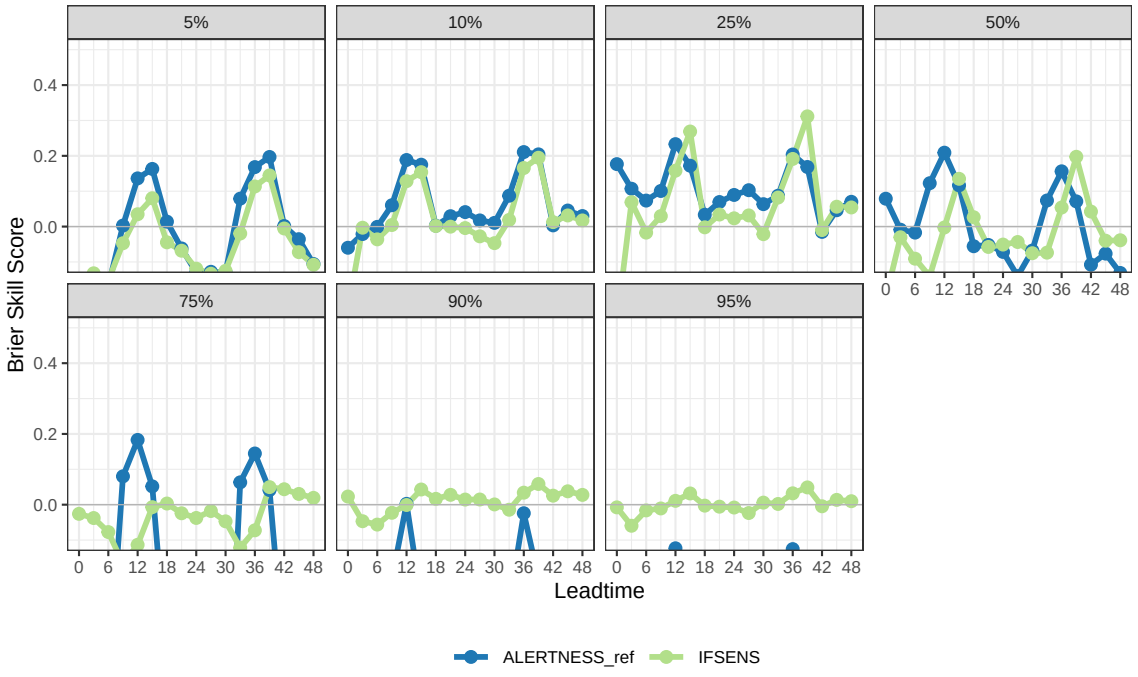

Figure 20: Brier Skill Score for 2m relative humidity during SOP 1 for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values The sample climatology is used as the reference forecast.

rate than IFSENS that is also accompanied with a higher false alarm rate.

In terms of economic value (Fig. 23), ALERTNESS_ref and IFSENS perform mostly similarly for both 12 (Fig. 23(a)) and 24 hour (Fig. 23(b)) lead times. At 12 hours lead time (Fig. 23(a)), ALERTNESS_ref provides marginal value for cost-loss ratios lower than about 30% for the 5th and 10th percentiles, whereas IFSENS provides zero value for those low cost-loss ratios. For the 75th percentile at 12 hours lead time ALERTNESS_ref offers more value for a large range of cost-loss ratios than IFSENS, and for the 90th and especially the 95th percentiles, ALERTNESS_ref provides more value than IFSENS for only the lowest cost-loss ratios. At 24 hours lead time (Fig. 23(b)), IFSENS has similar economic value to that provided at 12 hours lead time. For ALERTNESS_ref, however, the value provided is greatly reduced, especially for thresholds higher than the 75% percentile.

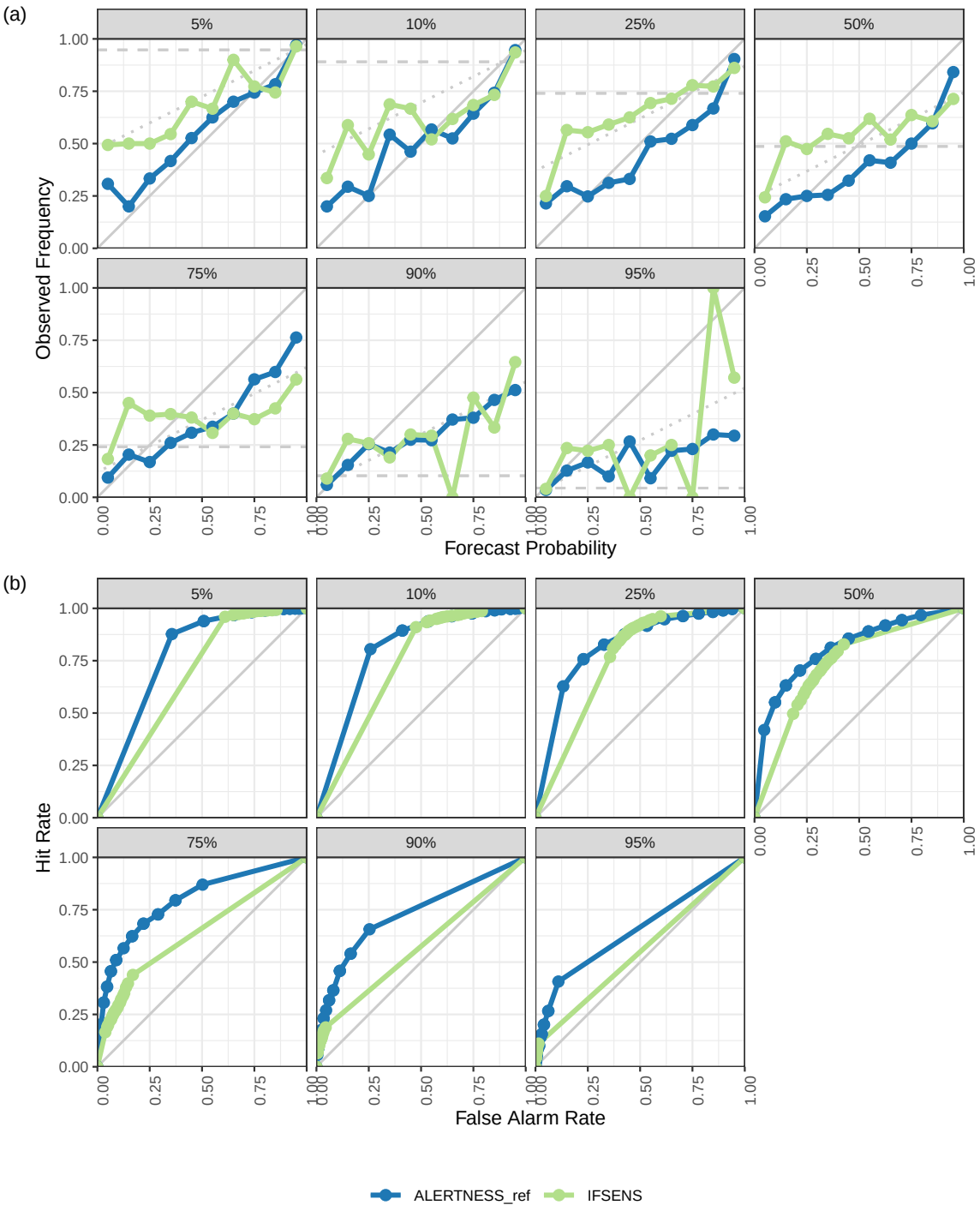

Figure 21: Verification for 2m relative humidity during SOP 1 at a lead time of 12 hours and for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values for (a) reliability and (b) ROC.

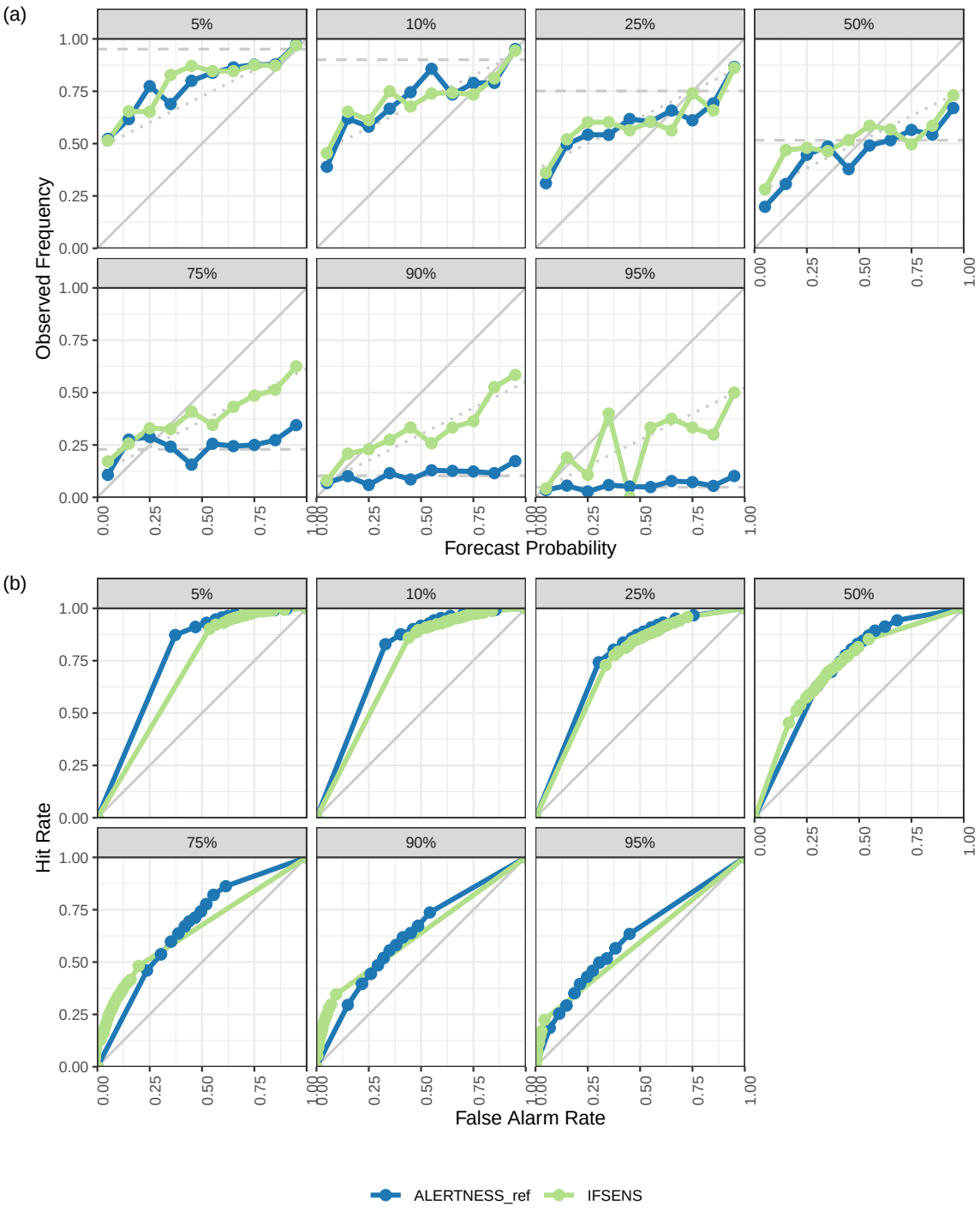

Figure 22: Verification for 2m relative humidity during SOP 1 at a lead time of 24 hours and for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values for (a) reliability and (b) ROC.

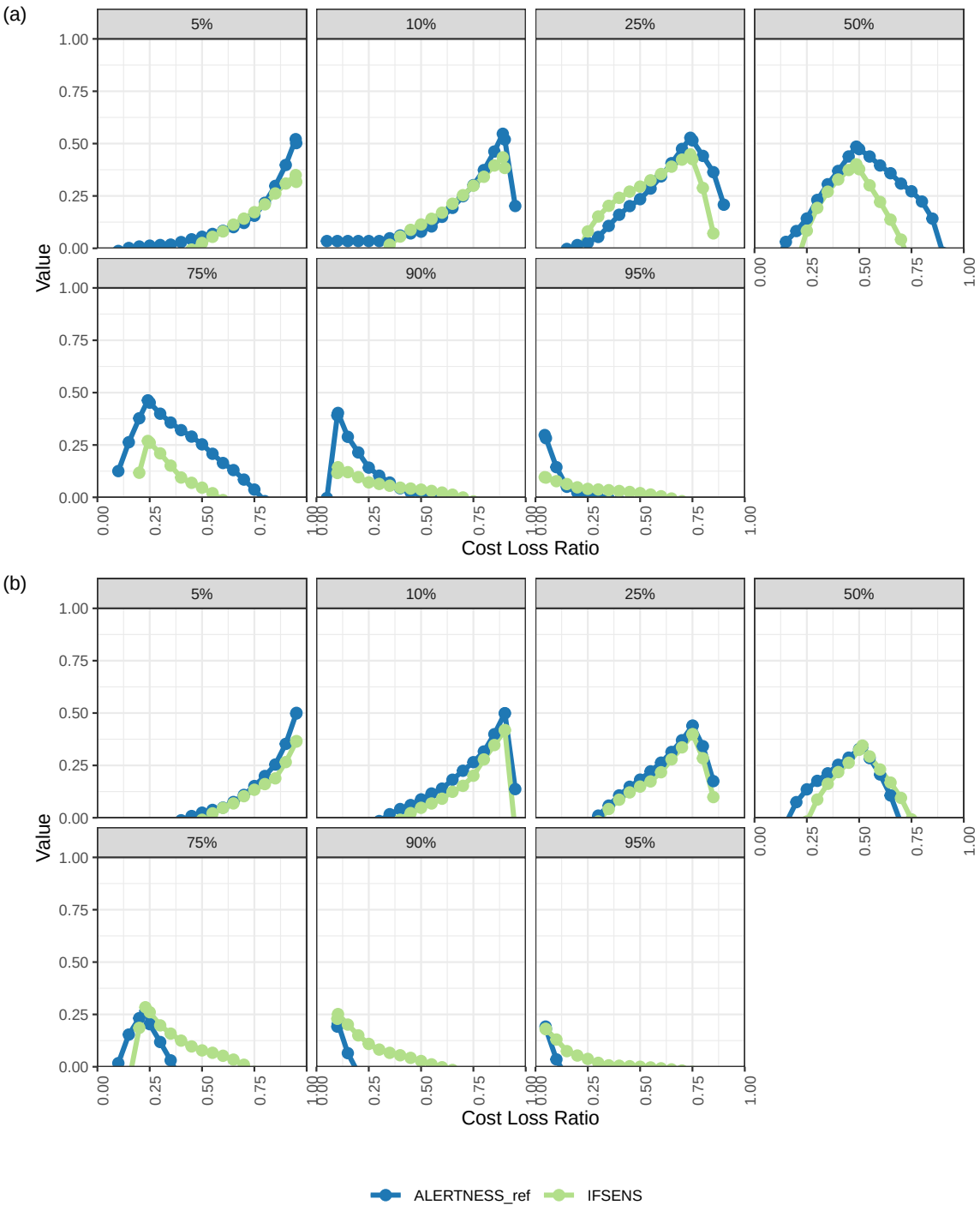

Figure 23: Economic value for 2m relative humidity forecasts during SOP 1 for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values at lead times of (a) 12 hours (b) 24 hours.

5.1.4 12 hour precipitation

Forecasts for 12 hour accumulated precipitation are verified at 06 and 18 UTC each day - this roughly separates the precipitation into day time and night time components. Furthermore, the largest number of stations is available for 12h precipitation at these hours. These stations are shown in Fig. 24. It should be noted that about half of the stations only have observations for 18 UTC (i.e. lead times of 18 and 42 hours).

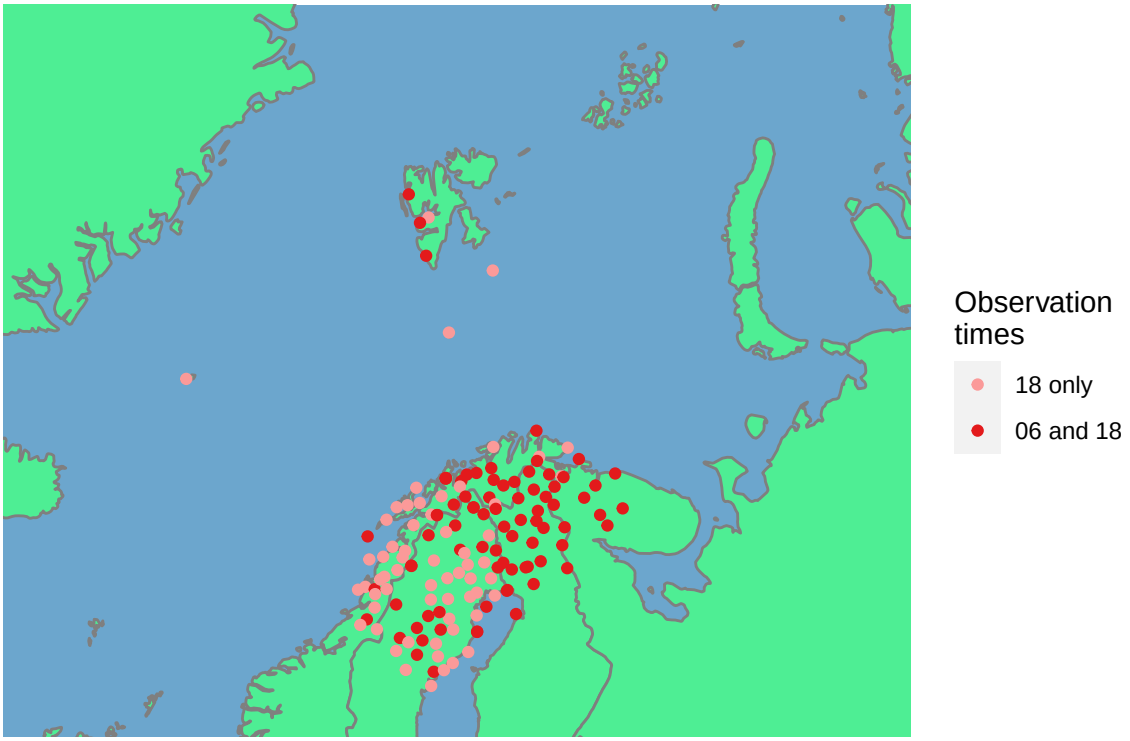

Figure 24: Stations used for 12 hour accumulated precipitation verification.

Summary scores for 12 hour accumulated precipitation at lead times 18, 30 and 42 hours for SOP1 are shown in Fig. 25. The 18 hour lead time can be considered to be 12 hour lead time on the first day of the forecast, the 30 hour lead time can be considered to be the night time precipitation and the 42 hour lead time the day time accumulated precipitation on the second day of the forecast. IFSENS clearly has lower RMSE (Fig. 25(a)) and CRPS (Fig. 25(b)) then ALERTNESS_ref for night time precipitation and day time precipitation on the second day of the forecast. On the first day of the forecast ALERTNESS_ref and IFSENS have similar RMSE and CRPS. The spread for ALERTNESS_ref is larger than that for IFSENS throughout the forecast, though the difference is much smaller on the second day of the forecast (Fig. 25(a)). IFSENS has a positive bias throughout the forecast that is slightly lower for the night time, while ALERTNESS_ref has a small negative bias on the first day of the forecast that becomes a small positive bias on the second day of the forecast, with a large positive bias for the night time precipitation (Fig. 25(c)). The rank histogram suggests that the ALERTTNESS_ref ensemble is slightly better dispersed than IFSENS (Fig. 25(d)).

The performance of the models is further assessed for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed 12h accumulated precipitation, where it is greater than zero, for each lead time. The evolution

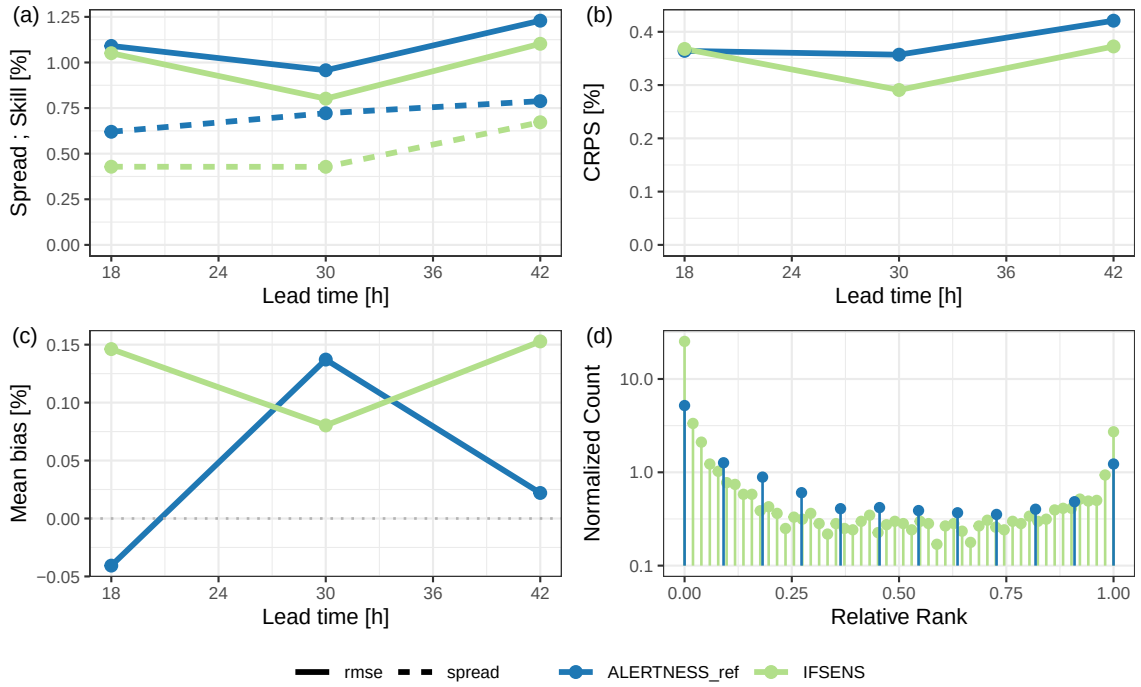

Figure 25: Summary verification scores for 12 hour accumulated precipitation during SOP 1: (a) RMSE and spread, (b) CRPS, (c) Bias of the ensemble mean and (d) Normalized relative rank histogram.

of the thresholds for these percentiles with lead time is shown in Fig. 26. Note that the 5th and 10th percentiles had almost exactly the same values so from hereon in, the 10th percentile is not shown for 12h accumulated precipitation verification. A diurnal cycle, with slightly lower precipitation during the night, is only apparent for the 90th and 95th percentiles.

The Brier Skill Score, using the sample climatology as reference, suggest that in general IFSENS is superior to ALERTNESS_ref for all percentiles (Fig. 27). On the first day of the forecast (18 hours lead time), ALERTNESS_ref and IFSENS have very similar Brier Skill Scores for all percentiles, with ALERTNESS_ref slightly higher than IFSENS for the 95th percentile. For all percentiles, except for the 75th and 90th, the Brier Skill Score is clearly higher for IFSENS for night time precipitation - for the 5th and 25th percentiles in particular there is a large drop in the Brier Skill Score at the 30 hour lead time. For the second day of the forecast, the Brier Skill Score for IFSENS is higher than that for ALERTNESS_ref for all percentiles except for the 95th, where the Brier Skill Score for ALERTNESS_ref is marginally higher

The reliability for the first day of the forecast suggests that ALERTNESS_ref and IFSENS perform broadly similarly (Fig. 28(a)), generally over forecasting the higher probabilities. The ROC curves, however, suggest that IFSENS performs better than ALERTNESS_ref with both higher hit rates and lower false alarm rates (Fig. 28(b)). For the night time precipitation, the reliability for both ALERTNESS_ref and IFSENS is similar (Fig. 29(a)), whereas the ROC suggests that the difference between IFSENS and ALERTNESS_ref is larger for the night time than for the day time ((Fig. 29(a) cf Fig. 28(b)), mostly due to lower hit rates for ALERTNESS_ref. The reliability and ROC performance for day time precipitation on the second day is comparable to that on the first day.

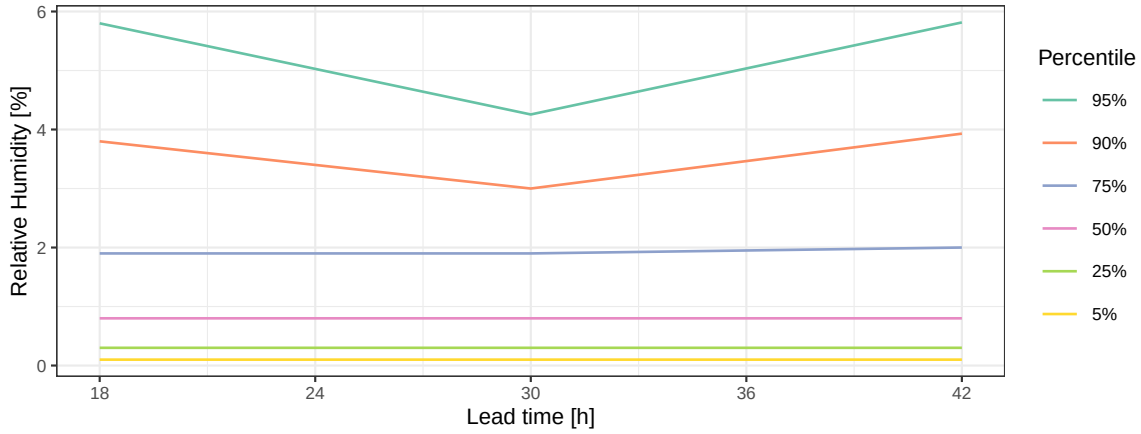

Figure 26: Thresholds used for categorical scores for 12 hour accumulated precipitation during SOP 1 derived from the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values that were greater than zero valid at each lead time.

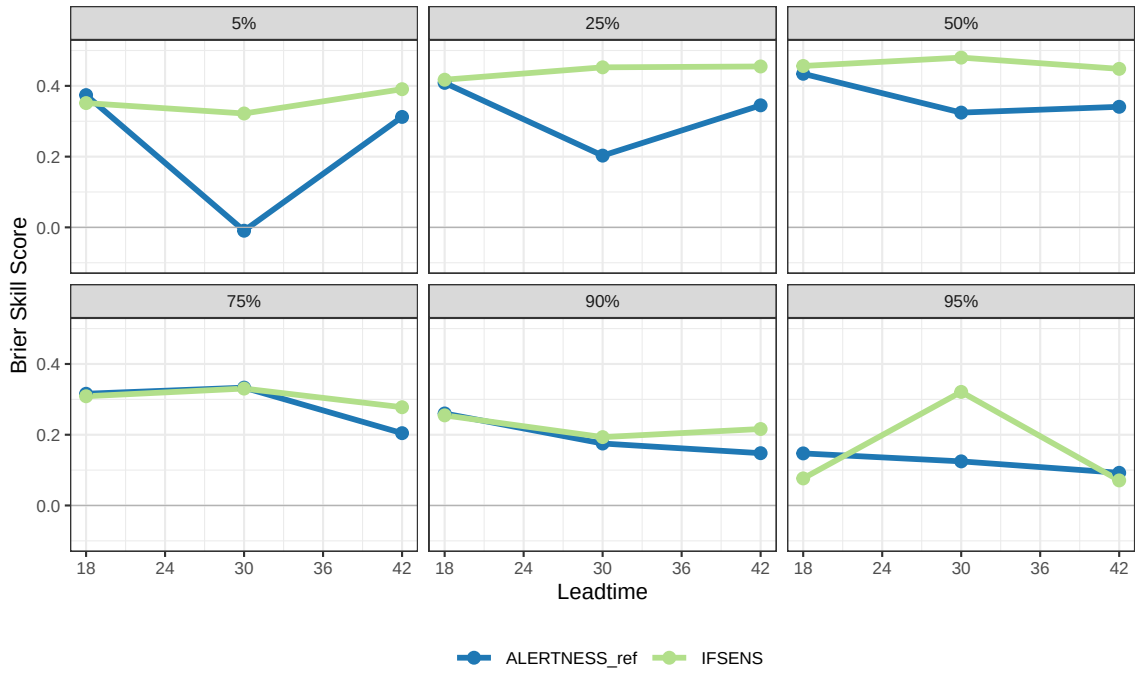

Figure 27: Brier Skill Score for 12 hour accumulated precipitation during SOP 1 for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed 12h accumulated precipitation greater than zero. The sample climatology is used as the reference forecast.

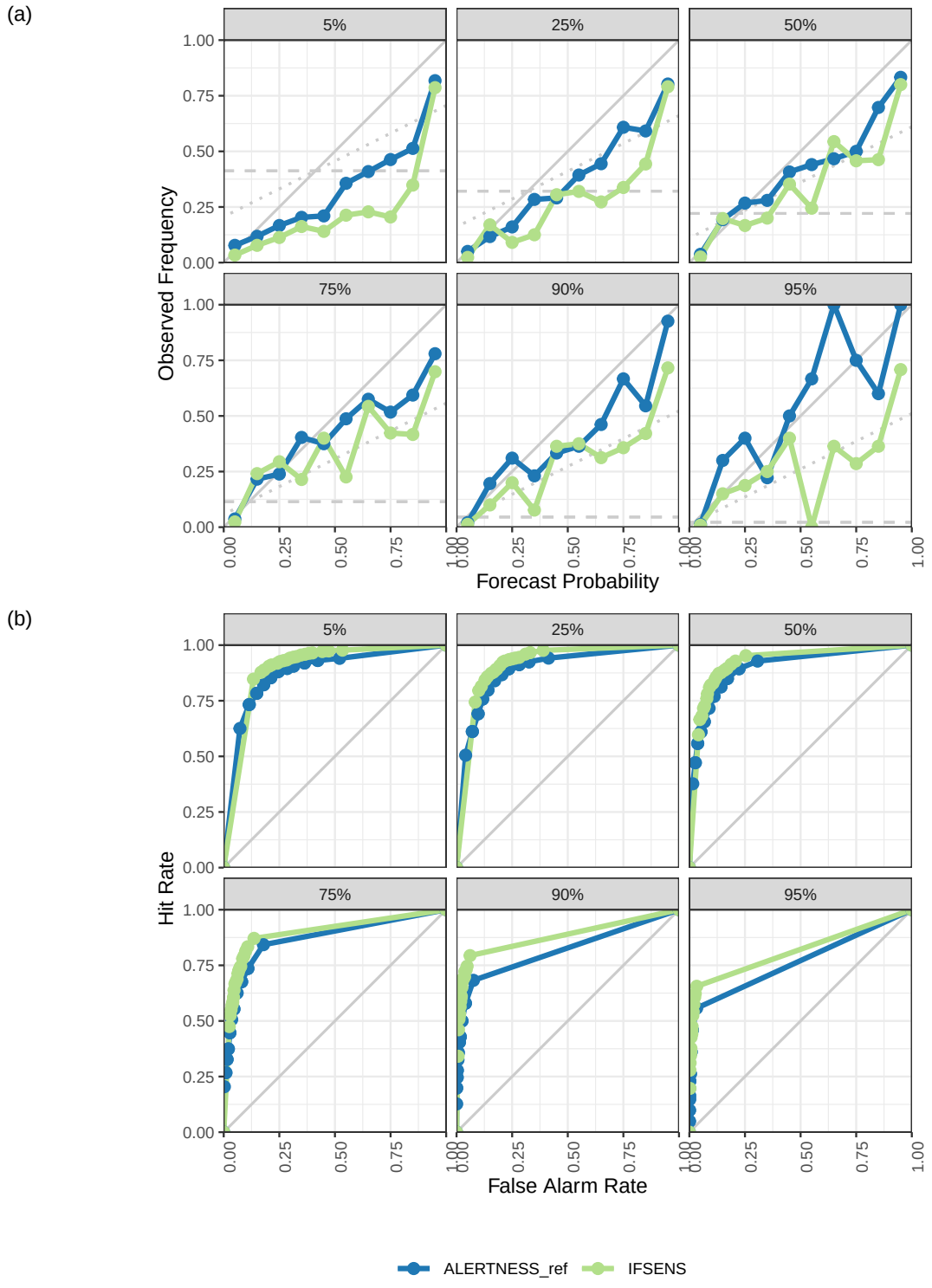

Figure 28: Verification for 12 hour accumulated precipitation during SOP 1 at a lead time of 18 hours and for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed 12 hour accumulated precipitation greater than zero for (a) reliability and (b) ROC.

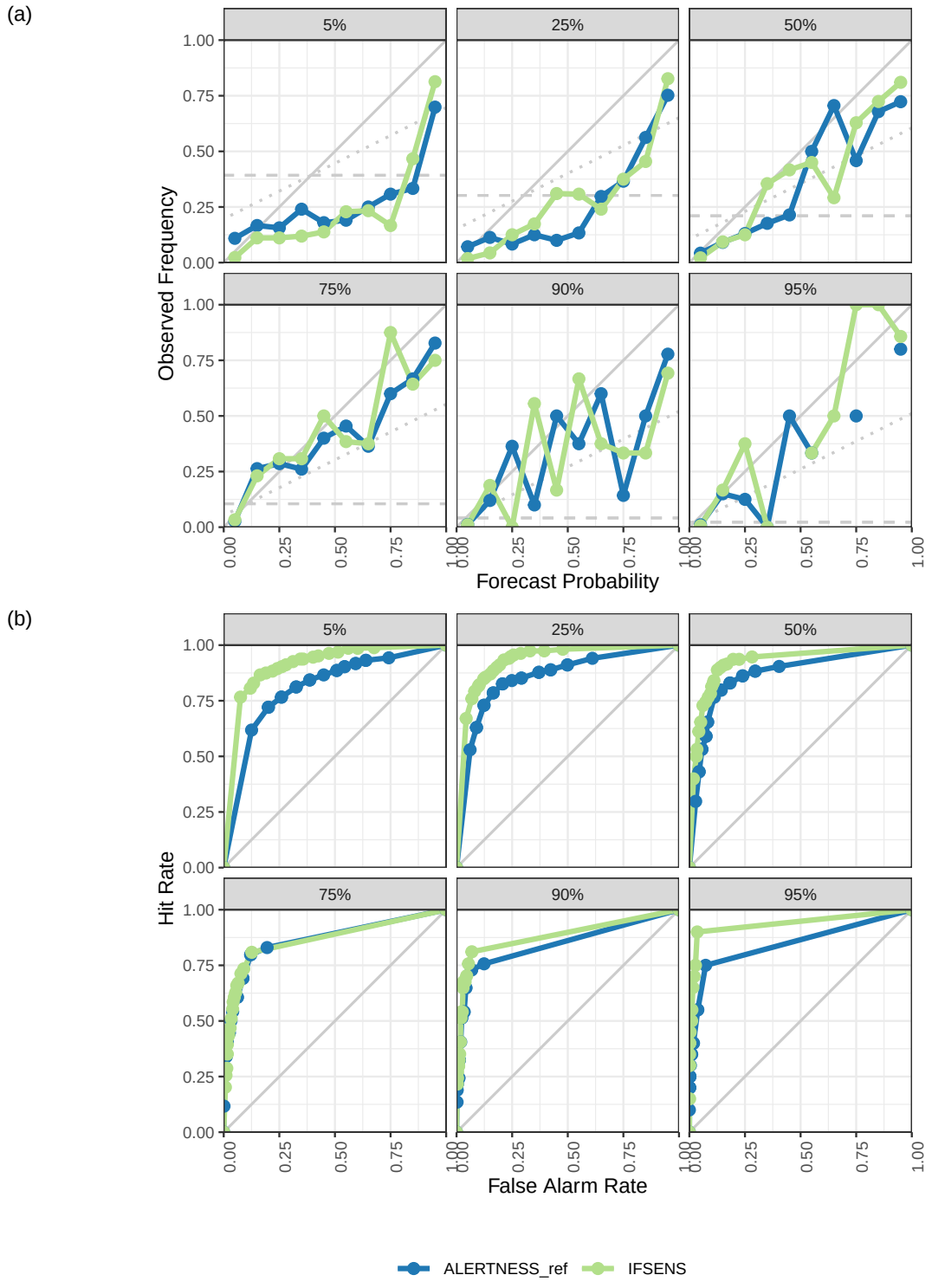

Figure 29: Verification for 12 hour accumulated precipitation during SOP 1 at a lead time of 30 hours and for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed 12 hour accumulated precipitation greater than zero for (a) reliability and (b) ROC.

The economic value (Fig. 30) of IFSENS is better than that for ALERTNESS_ref for all percentiles for both day time (Fig. 30(a)) and night time (Fig. 30(b)) precipitation. The exceptions are for cost-loss ratios greater than about 0.7 on the first day of the forecast for thresholds larger than the 75th percentile. For the night time precipitation, particularly for thresholds lower than the 50th percentile, IFSENS provides more value to users with a wider range of cost-loss ratios (Fig. 30(b)). For the second day of the forecast (not shown), the economic value for both ALERTNESS_ref and IFSENS is similar to that on the first day of the forecast, except that the extra economic value ALERTNESS_ref provides over IFSENS for users with high cost loss ratios for the higher percentiles is no longer apparent.

(a)

(b)

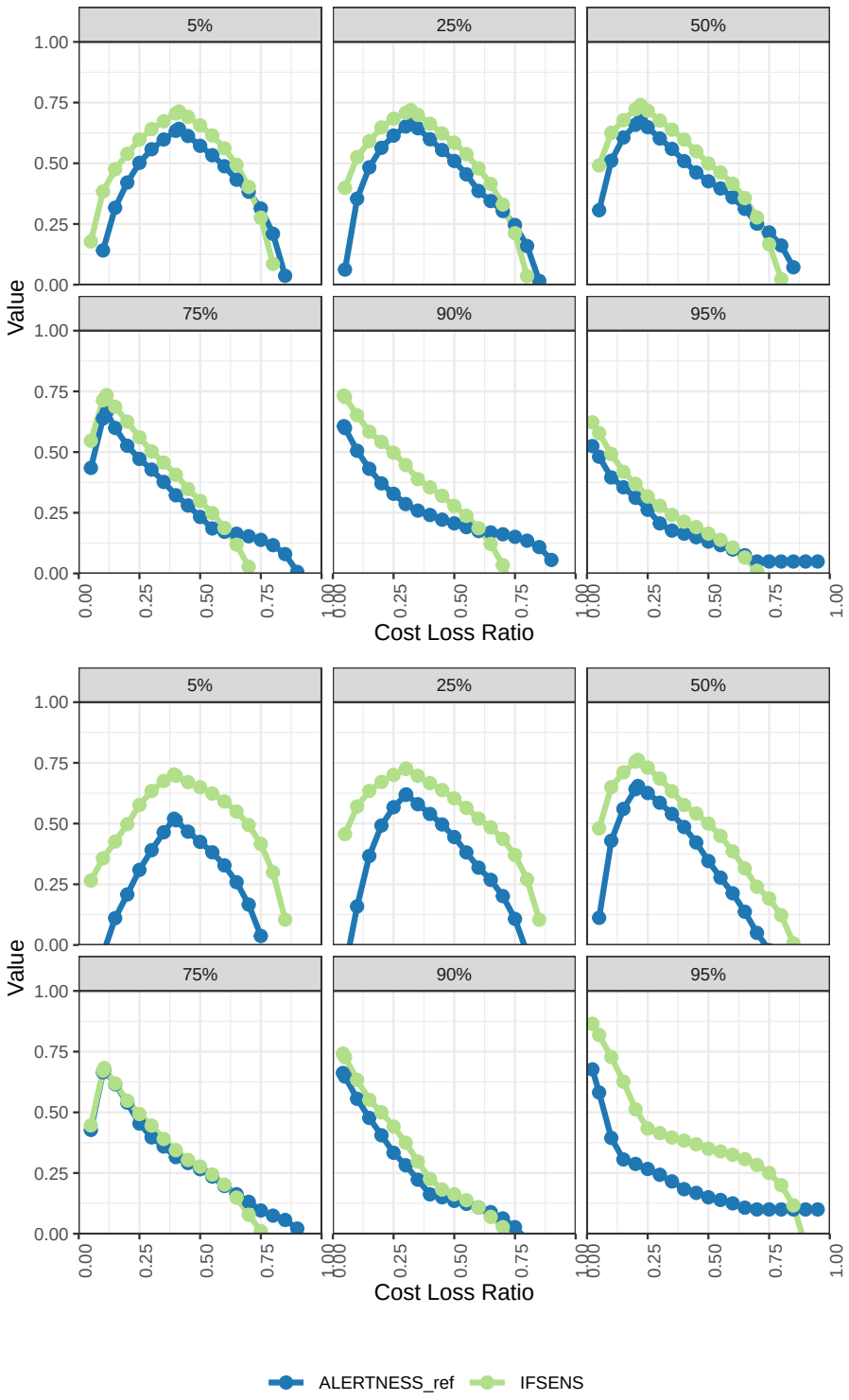

Figure 30: Economic value for 12 hour accumulated precipitation forecasts during SOP 1 for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed 12 hour accumulated precipitation gretaer than zero at lead times of (a) 18 hours (b) 30 hours.

5.2 SOP 2

5.2.1 2m temperature

As for SOP 1, forecasts of 2m temperature for SOP 2 are verified against all available stations within the AROME-Arctic domain (Fig. 4), with the forecasted 2m temperatures adjusted for differences in height between the model elevation and the station elevation using a lapse rate of $6.5°C/km$.

Summary scores for SOP 2 are shown in Fig. 31, comparing scores for 2m temperature from ALERTNESS_ref with those from IFSENS. As in SOP 1, there is a distinct diurnal cycle in the RMSE (Fig. 31(a)), CRPS (Fig. 31(b)) and bias of the ensemble mean (Fig. 31(c)), with the scores better in the day time than the night time, though the magnitude of the errors are slightly smaller for SOP 2 than for SOP 1. ALERTNESS_ref is clearly superior to IFSENS on the first day of the forecast with lower RMSE (Fig. 31(a)) and CRPS (Fig. 31(b)) and close to zero bias in the ensemble mean between lead times of 6 and 18 hours (Fig. 31(c)). Furthermore, the spread for ALERTNESS_ref is larger than for IFSENS throughout the forecast (Fig. 31(a)), and the rank histogram suggests that ALERTNESS_ref is better dispersed than IFSENS, with the relative ranks having a normalized count close to 1 (Fig. 31(d)).

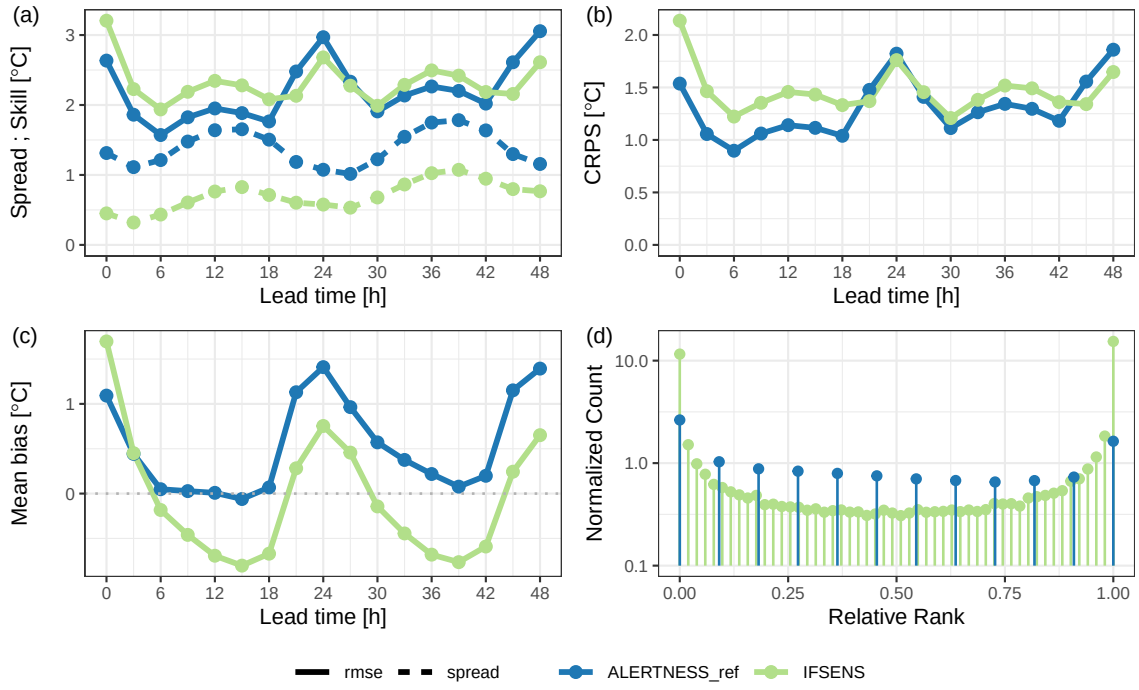

Figure 31: Summary verification scores for 2m temperature during SOP 2: (a) RMSE and spread, (b) CRPS, (c) Bias of the ensemble mean and (d) Normalized relative rank histogram.

Taking the performance of the model for different percentiles of 2m temperature into account resulted in the thresholds shown in Fig. 32. SOP 2 was an extremely warm period and this is reflected in high temperature thresholds, with the 95th percentile being close to $30°C$ during the day and around $21°C$ during the night. The diurnal cycle in the thresholds is much stronger for the higher percentiles than for the lower percentiles.

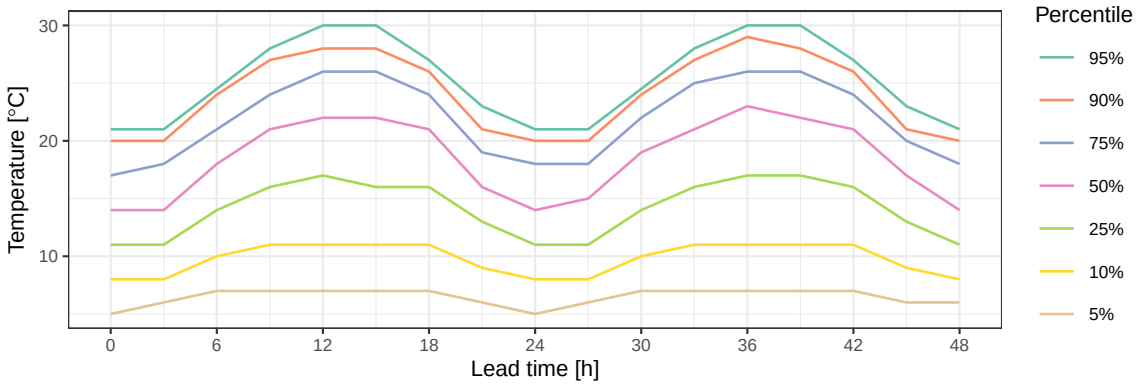

Figure 32: Thresholds used for categorical scores for 2m temperature during SOP 2 derived from the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values valid at each lead time.

The Brier Skill Scores for the thresholds shown in Fig. 32 can be seen in Fig. 33. In general ALERTNESS_ref has slightly higher Brier Skill Scores during the day time for all percentiles, but during the night time the Brier Skill Score for ALERTNESS_ref drops below that obtained from IFSENS for the higher percentiles. In fact, for the 90th and 95th percentiles, the Brier Skill Score for ALERTNESS_ref drops below zero indicating no skill in comparison to the sample climatology.

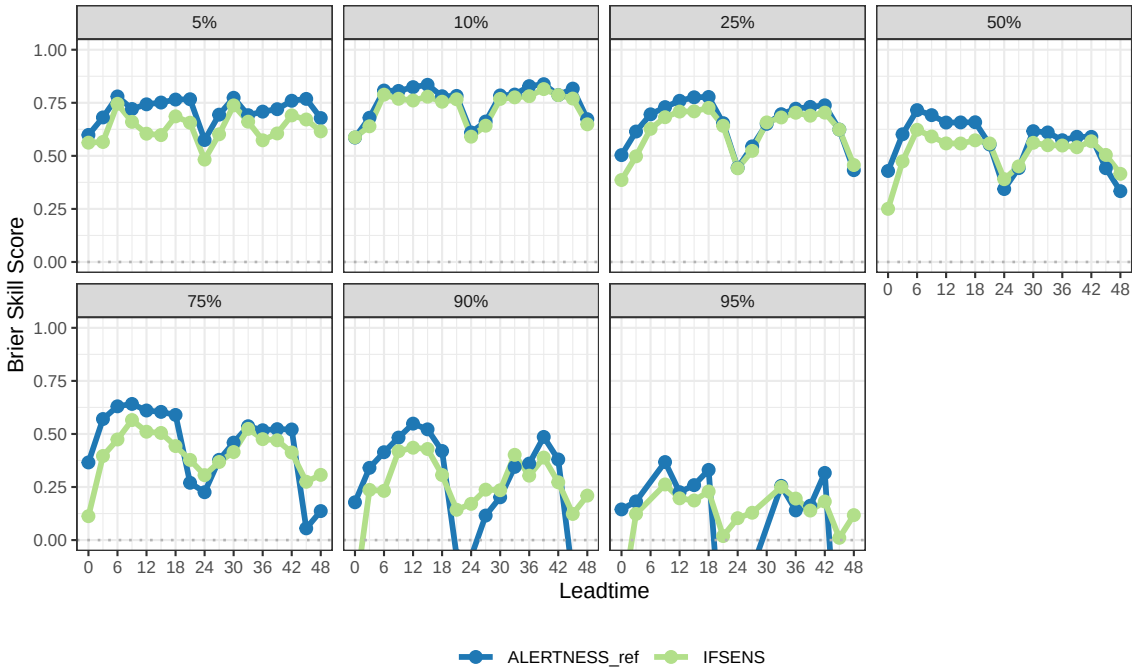

Figure 33: Brier Skill Score for 2m temperature during SOP 2 for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values. The sample climatology is used as the reference forecast.

The reliability for 2m temperature at 12 hours lead time suggests that ALERTNESS_ref and IFSENS perform similarly for the lower percentiles, but for the 75th and 90th percentiles ALERTNESS_ref is more reliable, with IFSENS tending to under forecast the probabilities (Fig. 34(a)). For the 95th percentile ALERT-

NESS_ref tends to over forecast the probabilities while IFSENS continues to under forecast. However for ROC at 12 hours lead time, the superiority of ALERTNESS_ref over IFSENS grows with increasing thresholds for 2m temperature (Fig. 34(b)). This is due to ALERTNESS_ref achieving much higher hit rates than IFSENS whilst maintaining a low false alarm rate. At 24 hours lead time both ALERTNESS_ref and IFSENS have similarly poor reliability, with ALERTNESS_ref especially having a tendency towards over forecasting of probabilities for the higher thresholds (Fig. 35(a)). Like at 12 hours lead time, ALERTNESS_ref performs increasingly better than IFSENS as the thresholds increase, but in this case the higher hit rates are also accompanied by higher false alarm rates (Fig. 35(b)).

The comparison of economic value provided ALERTNESS_ref and IFSENS (Fig. 36) suggests that ALERTNESS_ref is provides more value to users with a wider range of cost-loss ratios for the majority of thresholds of 2m temperature at both 12 hours lead time (Fig. 36(a)) and 24 hours lead time (Fig. 36(b)). The difference is most pronounced for users with lower cost-loss ratios.

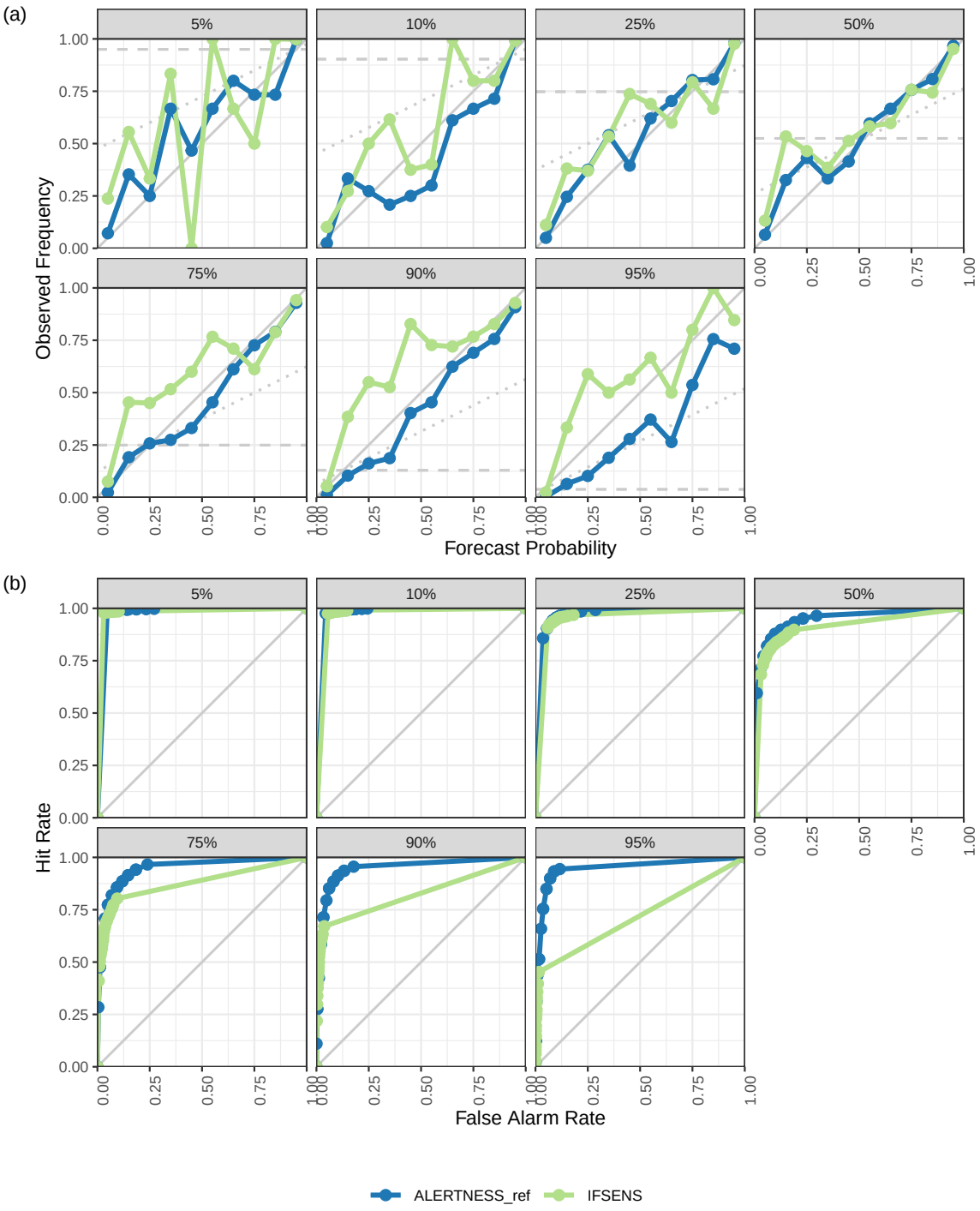

Figure 34: Verification for 2m temperature during SOP 2 at a lead time of 12 hours and for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values for (a) reliability and (b) ROC.

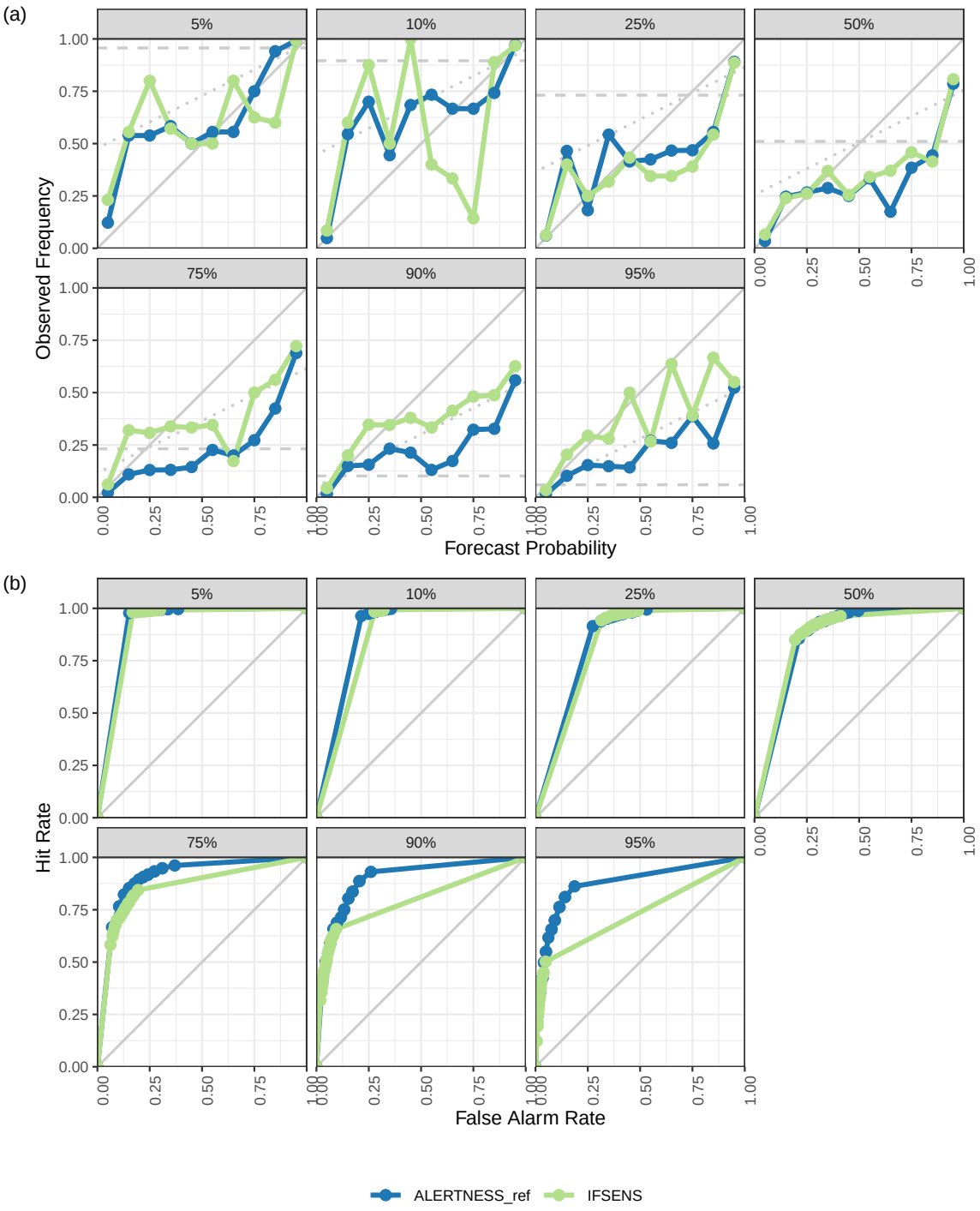

Figure 35: Verification for 2m temperature during SOP 2 at a lead time of 24 hours and for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values for (a) reliability and (b) ROC.

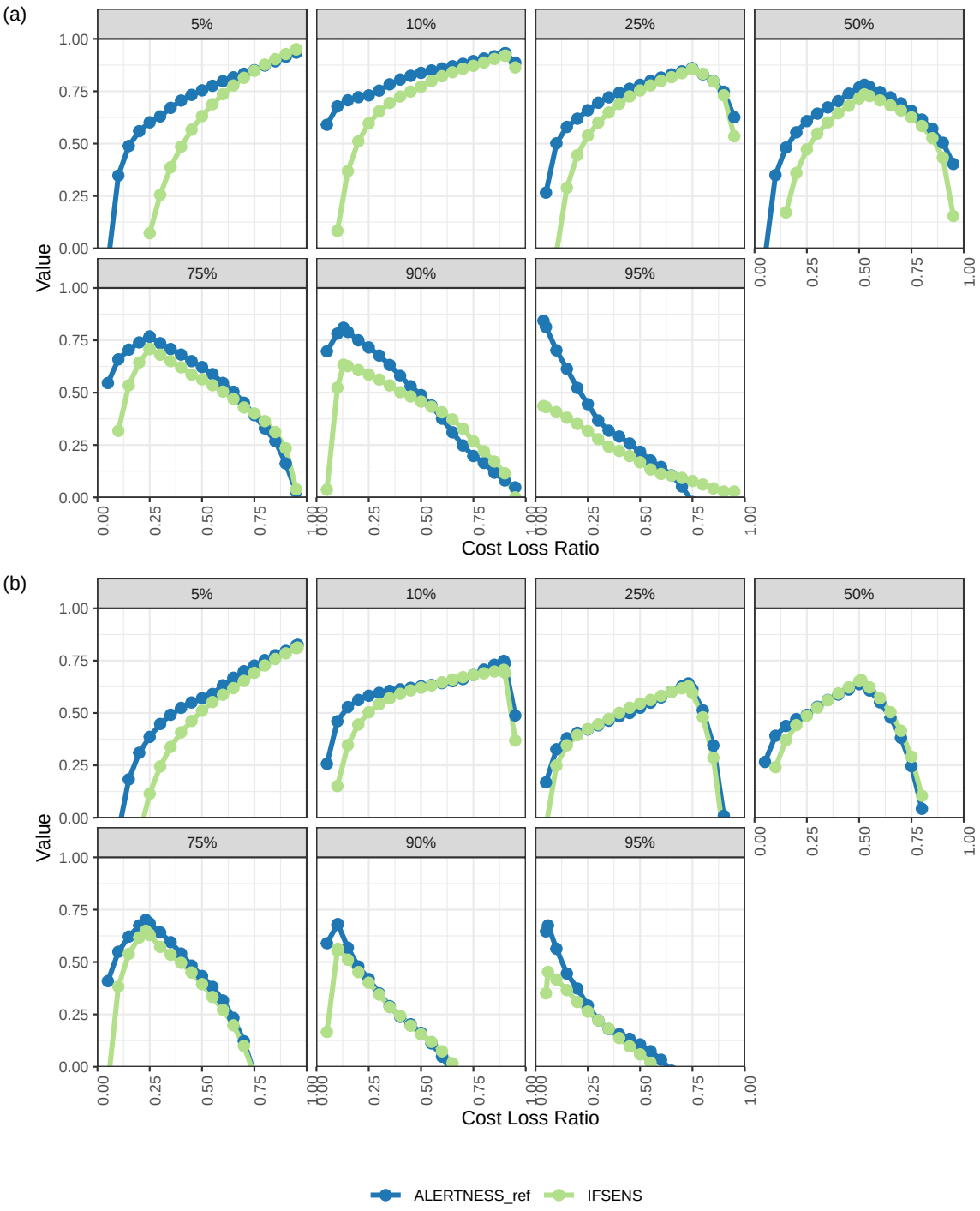

Figure 36: Economic value for 2m temperature forecasts during SOP 2 for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values at lead times of (a) 12 hours (b) 24 hours.

5.2.2 10m Wind Speed

Forecasts for 10m wind speed are verified for all available stations with observations inside the AROME-Arctic domain. These stations are shown in Fig. 11. No adjustments are made for differences between model elevation and station elevation. While the vast majority of stations are over land, there are also a large number of coastal stations and some offshore stations. Further verification against satellite derived winds over the sea will be possible in the future following developments from WP1.

Summary scores for 10m wind speed during SOP 2 are shown in Fig. 37, and are similar to those for SOP 1. The RMSE (Fig. 37(a)) and CRPS (Fig. 37(b)) are lower for ALERTNESS_ref throughout the forecast and the spread for IFSENS grows while that for ALERTNESS_ref remains at the same level if the diurnal cycle is not taken into account (Fig. 37(a)). Furthermore, the rank histogram suggests that ALERTNESS_ref is more evenly dispersed than IFSENS (Fig. 37(d)). Unlike for SOP 1, which suggested a positive bias in the ensemble mean of ALERTNESS_ref, there is a small positive bias during the night time and a negative bias during the day time (Fig. 37(c)). However, the diurnal cycle of the bias of the ensemble mean is stronger for IFSENS than for ALERTNESS_ref.

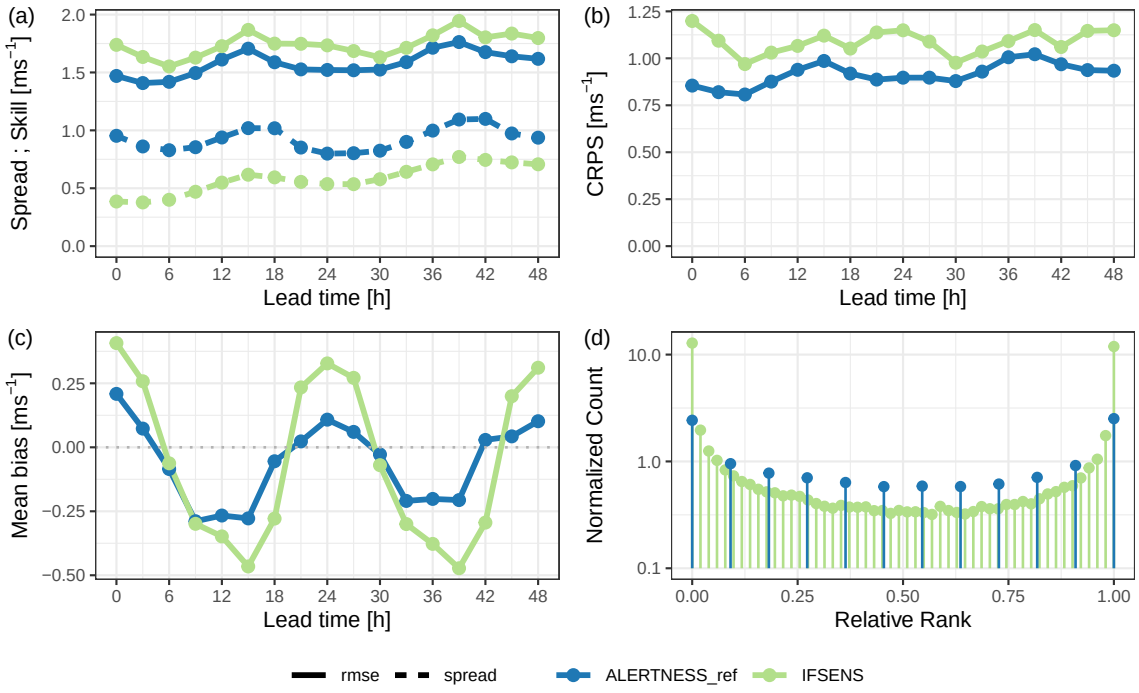

Figure 37: Summary verification scores for 10m wind speed during SOP 2: (a) RMSE and spread, (b) CRPS, (c) Bias of the ensemble mean and (d) Normalized relative rank histogram.

The thresholds obtained from the observed 10m wind speed for the 50th, 75th, 90th and 95th percentiles are shown in Fig. 38. The values obtained from these percentiles result in a maximum of threshold of 8 ms^{-1} for the the 95th percentile, which is lower than the maximum considered during SOP 1 (10.5 ms^{-1}). A weak diurnal cycle in the percentiles is apparent with higher 10m wind speeds in the day time than the night time, and is strongest for the 75th percentile.

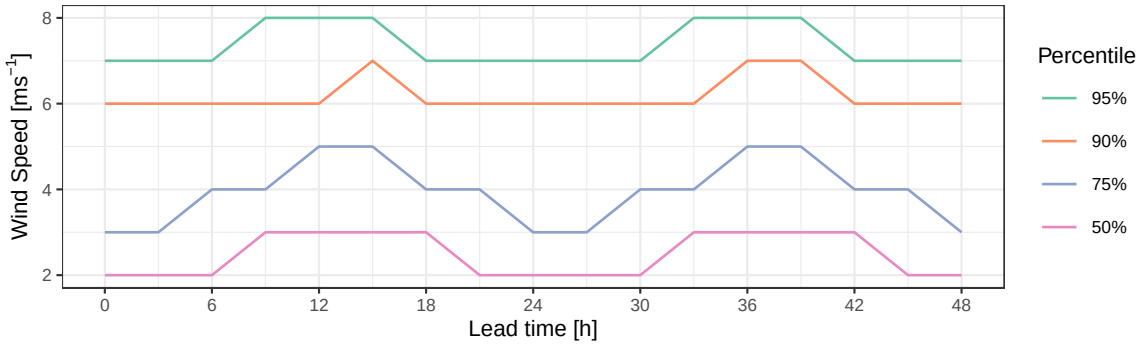

Figure 38: Thresholds used for categorical scores for 10m wind speed during SOP 2 derived from the 50th, 75th, 90th and 95th percentiles of the observed values valid at each lead time.

The Brier Skill Score suggests that ALERTNESS_ref is better than IFSENS throughout the forecast for the 50th and 75th percentiles, but only up to about 15 hours lead time for the 95th percentile (Fig. 39). In fact, for the 50th percentile, IFSENS has a negative Brier Skill Score between lead times 21 and 27 hours inclusive. There is a general increase in the Brier Skill Score for both model from the 50th to the 90th percentile, though there is a small decrease again for the 95th percentile.

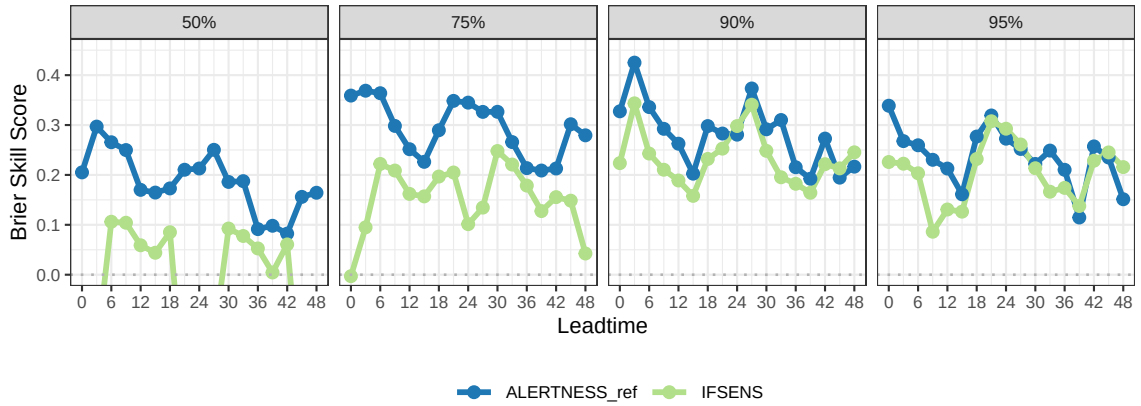

Figure 39: Brier Skill Score for 10m wind speed during SOP 1 for the 50th, 75th, 90th and 95th percentiles of the observed values. The sample climatology is used as the reference forecast.

Both ALERTNESS_ref and IFSENS have similar reliability curves for all percentiles at both 12 hours (Fig. 40(a)) and 24 hours (Fig. 41(a)) lead time, although there is an indication that ALERTNESS_ref is more reliable than IFSENS for the 50th and 75th percentiles when higher probabilities are forecasted. The ROC for 12 hours lead time (Fig. 40(b)) suggests that for the 50th percentile both ALERTNESS_ref and IFSENS perform similarly, but as the percentile increases, the hit rate for IFSENS falls more quickly than for ALERTTNESS_ref. At 24 hours lead time, however, ALERTNESS_ref clearly has better ROC than IFSENS for the 50th percentile as a result of a lower false alarm rate. Although that superiority of ALERTNESS_ref is maintained for all percentiles, mostly due to a higher rate, the difference between the two models doesn't really grow.

The economic value (Fig. 42) suggests that at the 12 hour lead time, ALERTNESS_ref provides slightly

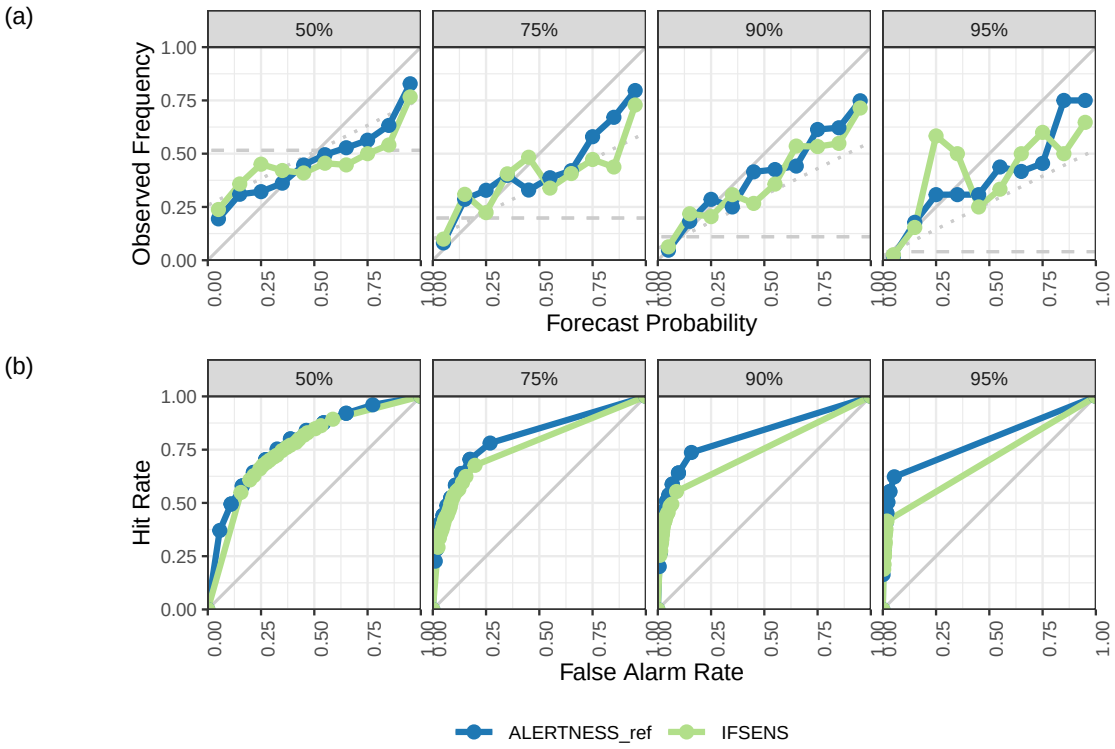

Figure 40: Verification for 10m wind speed during SOP 2 at a lead time of 12 hours and for the 50th, 75th, 90th and 95th percentiles of the observed values for (a) reliability and (b) ROC.

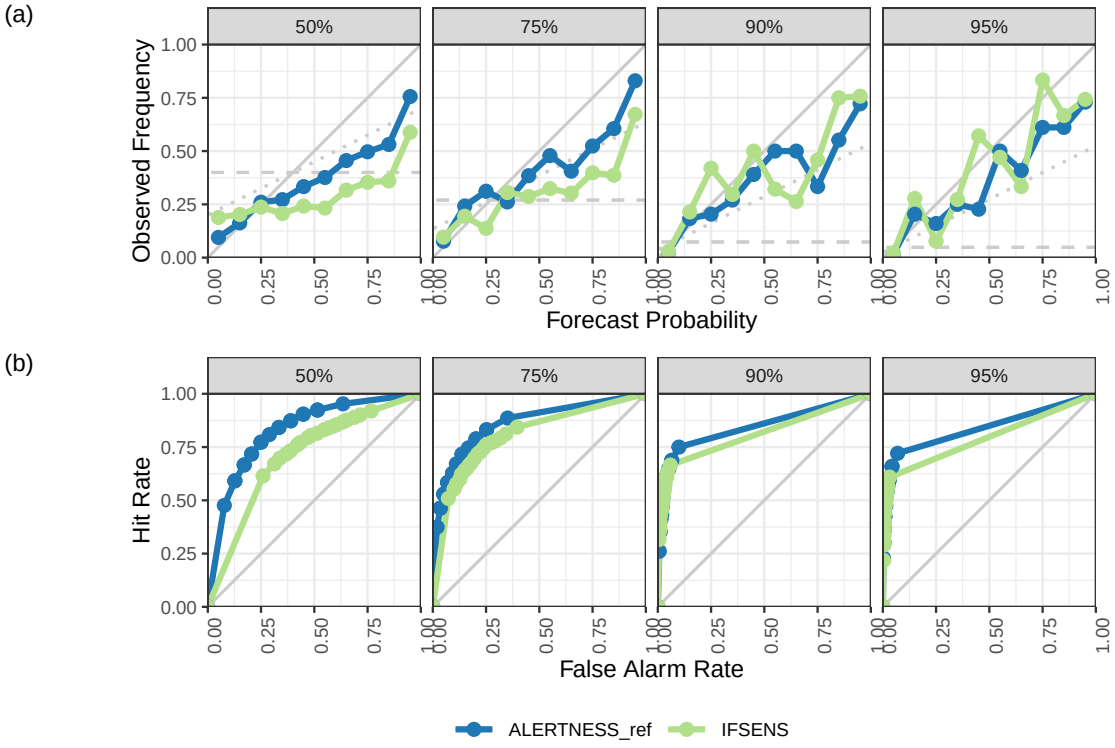

Figure 41: Verification for 10m wind speed during SOP 2 at a lead time of 24 hours and for the 50th, 75th, 90th and 95th percentiles of the observed values for (a) reliability and (b) ROC.

more value than IFSENS (Fig. 42(a)), but at 24 hours lead time (Fig. 42(b)), ALETRNESS_ref clearly provides more value for the 50th percentile, but as the threshold increases, the difference between the two models becomes almost undetectable.

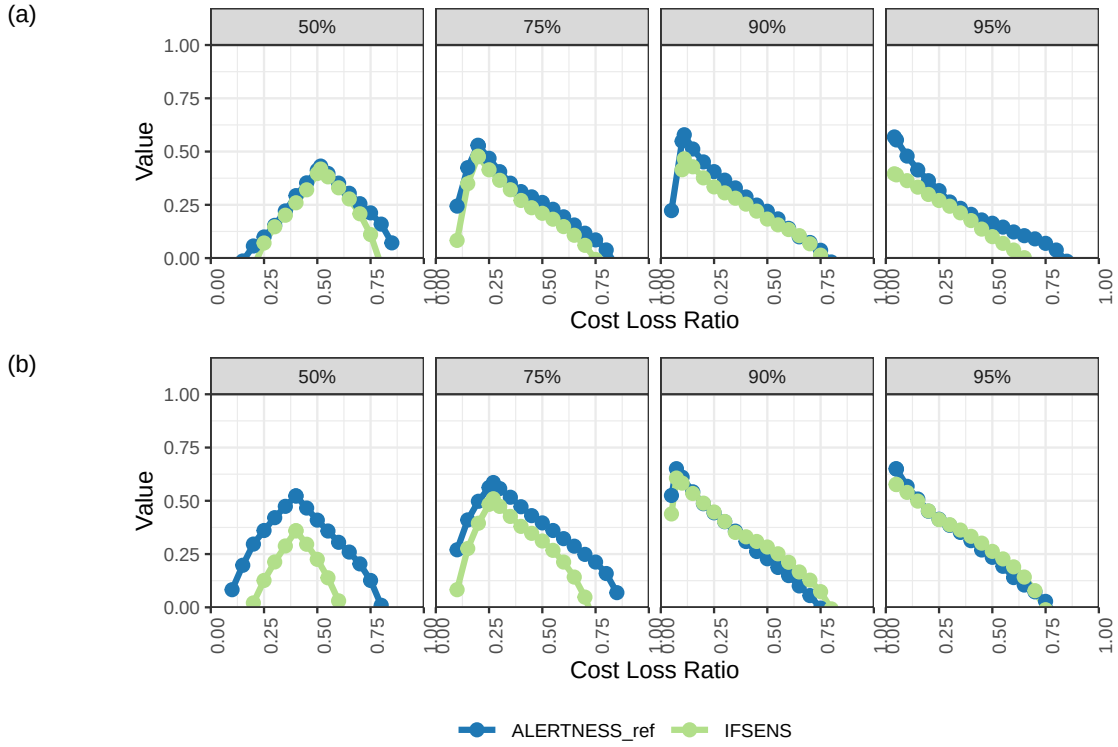

Figure 42: Economic value for 10m wind speed forecasts during SOP 2 for the 50th, 75th, 90th and 95th percentiles of the observed values at lead times of (a) 12 hours (b) 24 hours.

5.2.3 2m Relative Humidity

Forecasts for 2m relative humidity are verified for all available stations with observations inside the AROME-Arctic domain. These stations are shown in Fig. 17. No adjustments are made for differences between model elevation and station elevation.

Summary scores for SOP 2 are shown in Fig. 43, comparing scores obtained from ALERTNESS_ref with those from IFSENS. As for SOP 1, ALERTNESS_ref is not clearly superior to IFSENS. The diurnal cycles in the RMSE (Fig. 43(a)) and CRPS (Fig. 43(b)) are similar to those seen in the 2m temperature (Fig. 31(a), Fig. 31(b)), suggesting that it may be the temperature component of the relative humidity that is driving the diurnal cycle. The RMSE ion 2m relative humidity for ALERTNESS_ref is comparable to that for IFSENS during the day time but is larger during the night time (Fig. 43(a)), with the same seen in the CRPS (Fig. 43(b)). It appears that on the first day of the forecast the spread for ALERTNESS_ref exceeds the RMSE, while on the second day it is roughly equal (Fig. 43(a)), suggesting that the ensemble may be slightly over dispersed, though the rank histogram shows more signs of a slight negative bias with higher normalized counts for higher relative ranks (Fig. 43(d)). The diurnal cycle of the bias of the ensemble mean

for ALERTNESS_ref suggests strong negative biases during the night time, with close to zero bias during the day time (Fig. 43(c)), and it is likely that the night time negative biases are contributing to the negative bias signal in the rank histogram. IFSENS, has a weaker diurnal cycle and the bias of the ensemble mean is mostly positive throughout the forecast.

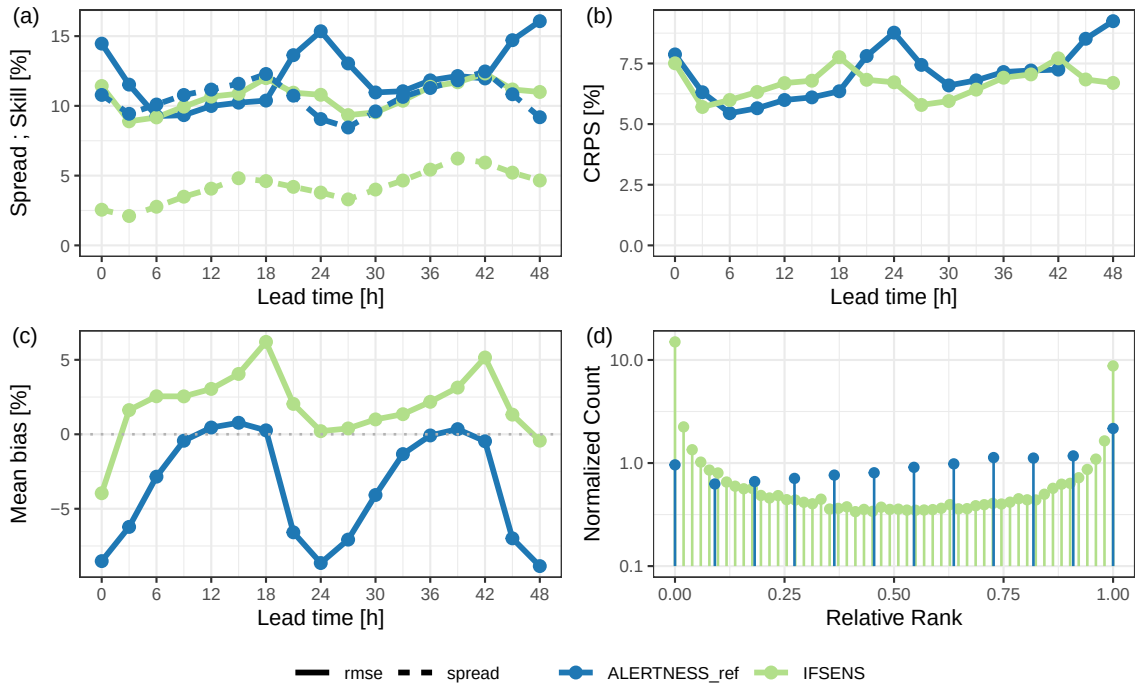

Figure 43: Summary verification scores for 2m relative humidity during SOP 2: (a) RMSE and spread, (b) CRPS, (c) Bias of the ensemble mean and (d) Normalized relative rank histogram.

The performance of the models is further assessed for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed 2m relative humidity for each lead time. The evolution of the thresholds for these percentiles with lead time is shown in Fig. 44, where a clear diurnal cycle can be seen in the thresholds, with lower relative humidity in the day time hours than the night time hours. Like for SOP 1, the diurnal cycle becomes less pronounced for the higher percentiles.

Fig. 45 shows the Brier Skill Score for each of the thresholds using the sample climatology as reference. For the 5th percentile, only IFSENS has skill, and as the percentile increases ALERTNESS_ref begins to have skill during the day time and a diurnal cycle in the score develops for IFSENS. The maximum Brier Skill Score for both models is achieved during the day time for the 50th percentile before beginning to drop. On the first day of the forecast, ALERTNESS_ref has a slightly higher Brier Skill Score than IFSENS for the 25th percentile and above.

The reliability plots for the 12 hour forecasts of 2m relative humidity (Fig. 46(a)) suggest that both models under forecast the probabilities for the lowest percentiles and begin to become reliable for the higher percentiles. The ROC plots suggest a high false alarm rate for the lower percentiles for IFSENS compared with ALERTNESS_ref, and for the 90th and 95th percentiles ALERTNESS_ref tends to have a higher hit rate than IFSENS, although this is accompanied by a slightly higher false alarm rate (Fig. 46(b)). For

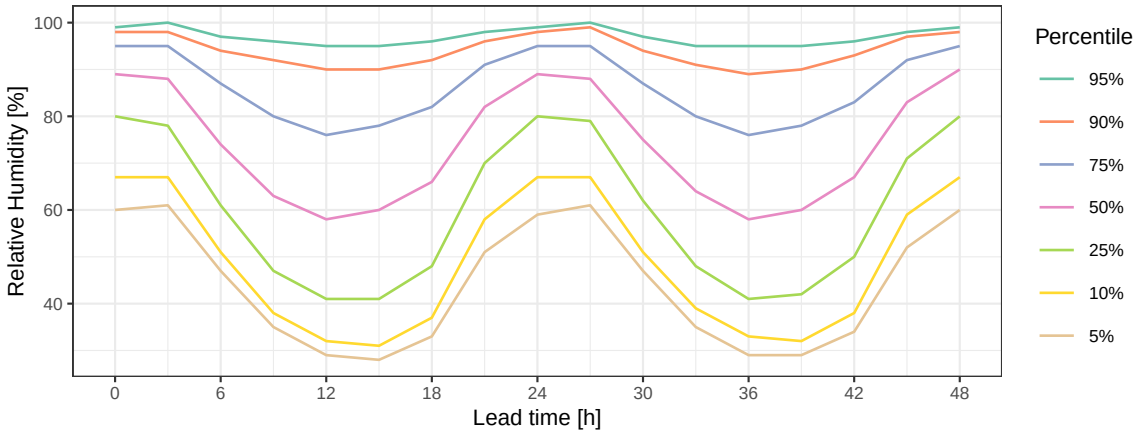

Figure 44: Thresholds used for categorical scores for 2m relative humidity during SOP 2 derived from the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values valid at each lead time.

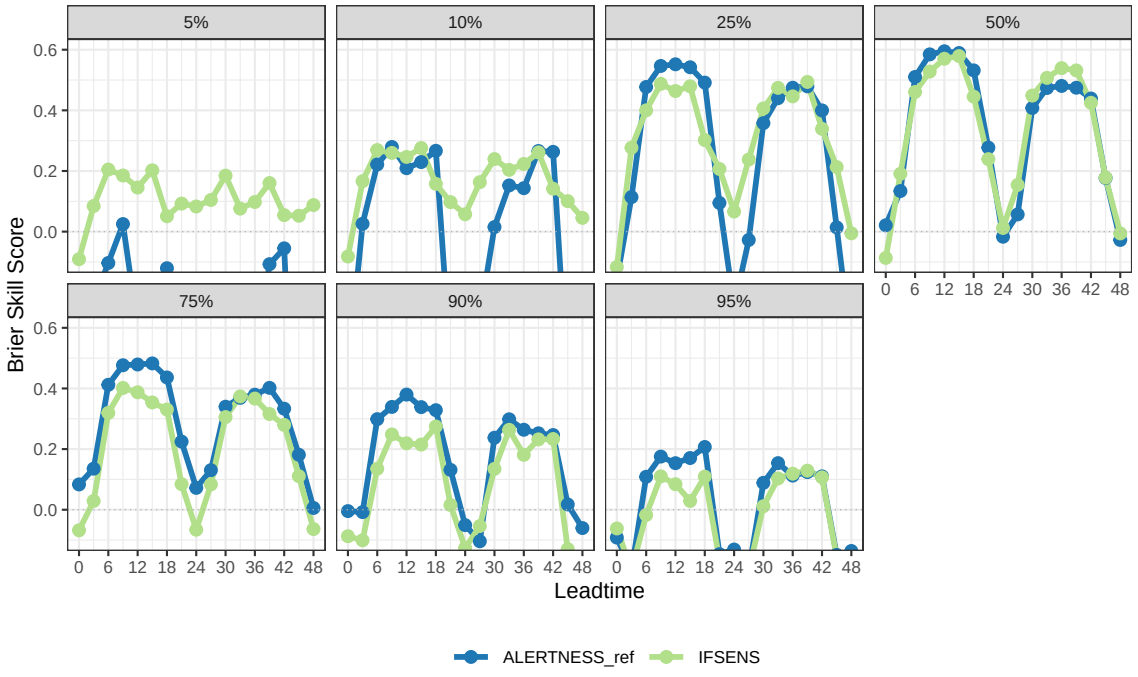

Figure 45: Brier Skill Score for 2m relative humidity during SOP 1 for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values. The sample climatology is used as the reference forecast.

24 hours lead time, both models tend to under forecast the probabilities for the lower percentiles, with ALERTNESS_ref under forecasting more than IFSENS, but for the 90th and 95th percentiles both models over forecast the probabilities (Fig. 47(a)). The ROC at 24 hours lead time is similar to that at 12 hours lead time for the lower percentiles with ALERTNESS_ref having a lower false alarm rate than IFSENS (Fig. 47(b)), but for the 90th and 95th percentiles the ROC curves for both models get quite close to the diagonal, with IFSENS slightly better than ALERTNESS_ref due to a higher (thought still less than 50%) hit rate.

In terms of economic value (Fig. 48), both ALERTNESS_ref and IFSENS perform better at 12 hours lead time (Fig. 48(a)) than at 24 hours lead time (Fig. 48(b)). ALETRNESS_ref provides more value than IFSENS to users with very high cost-loss ratios for all percentiles (Fig. 48(a)). The same is true for 24 hours lead time for percentiles up to the 75th, but for the 90th and 95th percentiles neither model provides for much value excerpt for a small amount for users with cost-loss ratios below about 0.3.

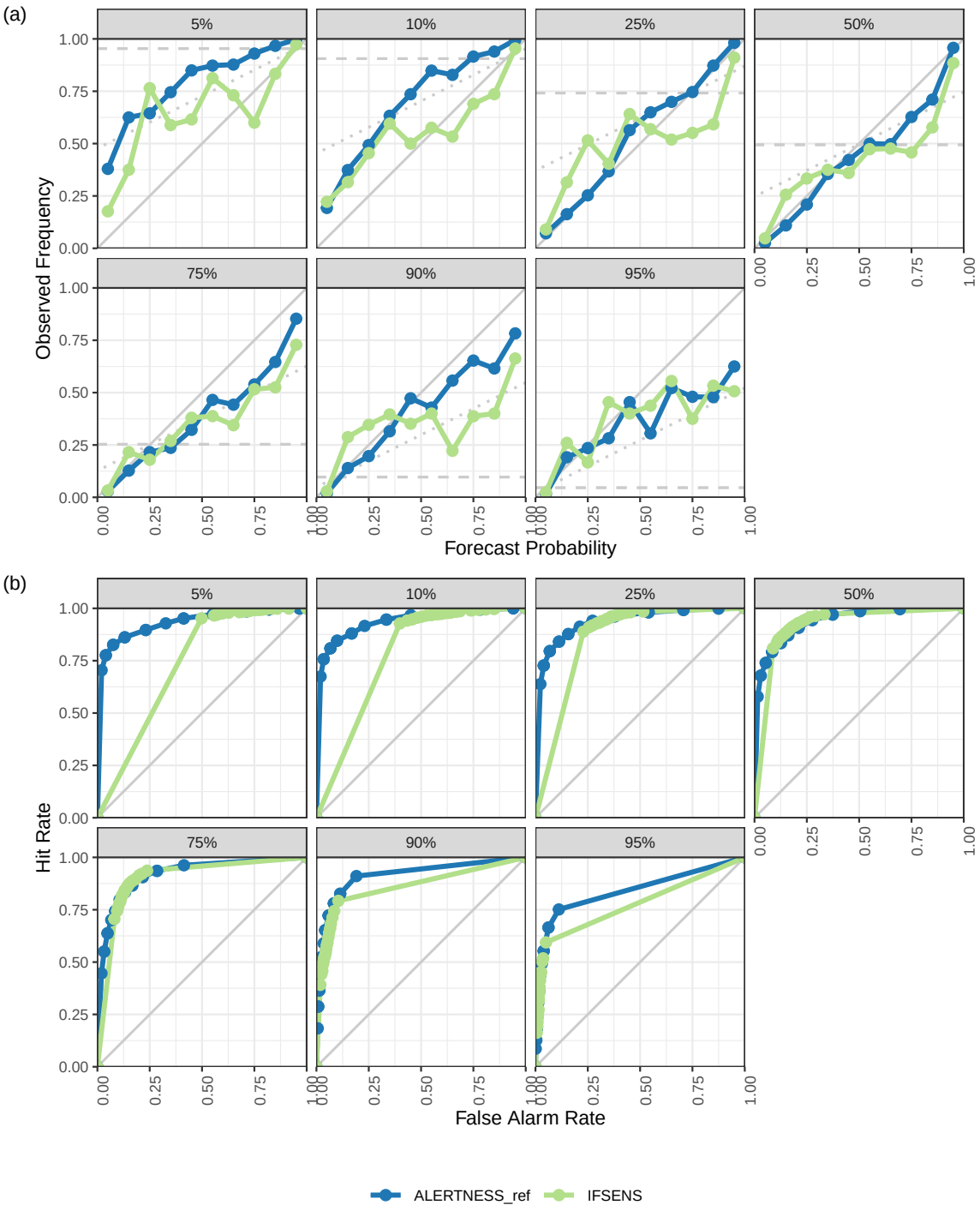

Figure 46: Verification for 2m relative humidity during SOP 2 at a lead time of 12 hours and for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values for (a) reliability and (b) ROC.

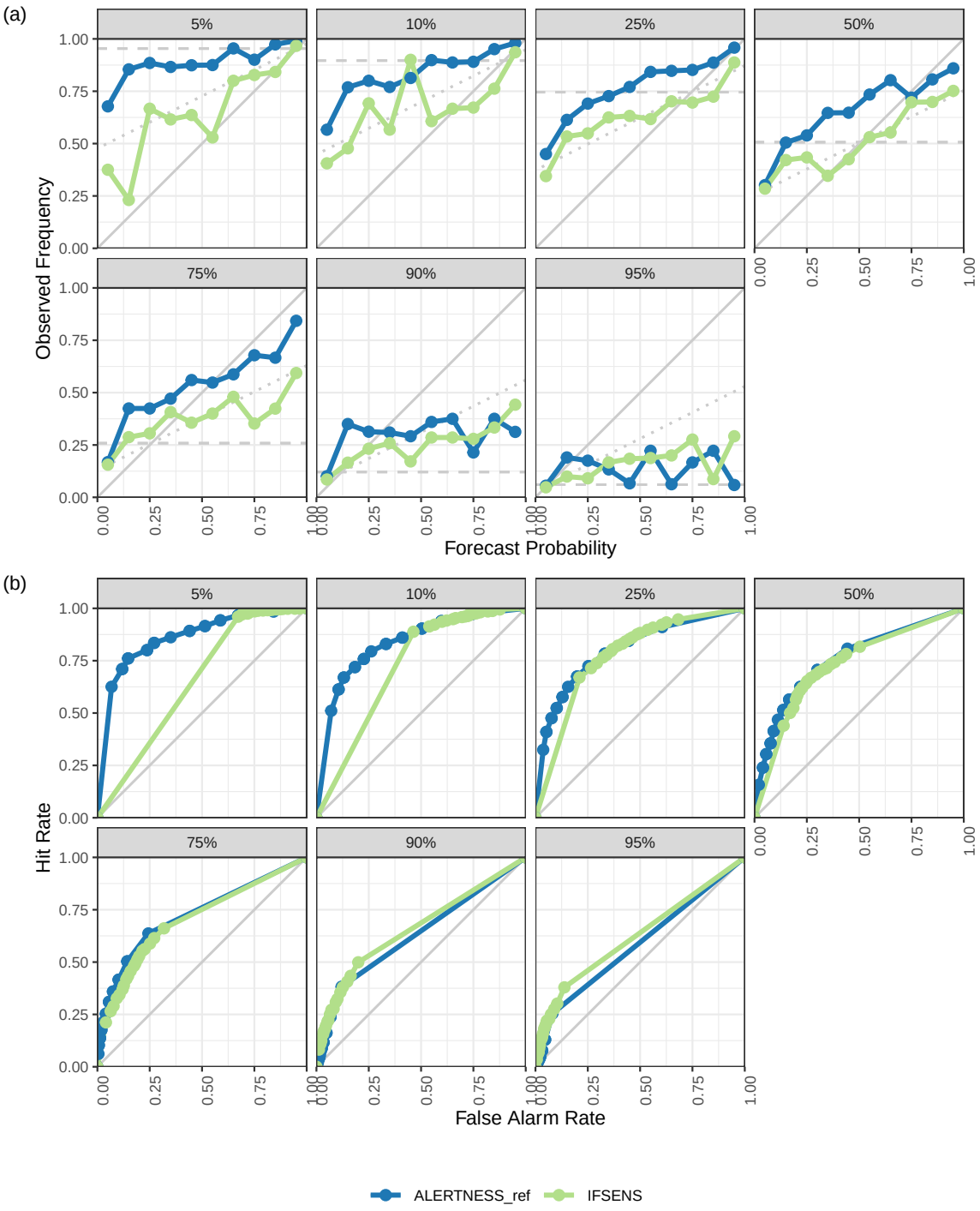

Figure 47: Verification for 2m relative humidity during SOP 2 at a lead time of 24 hours and for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values for (a) reliability and (b) ROC.

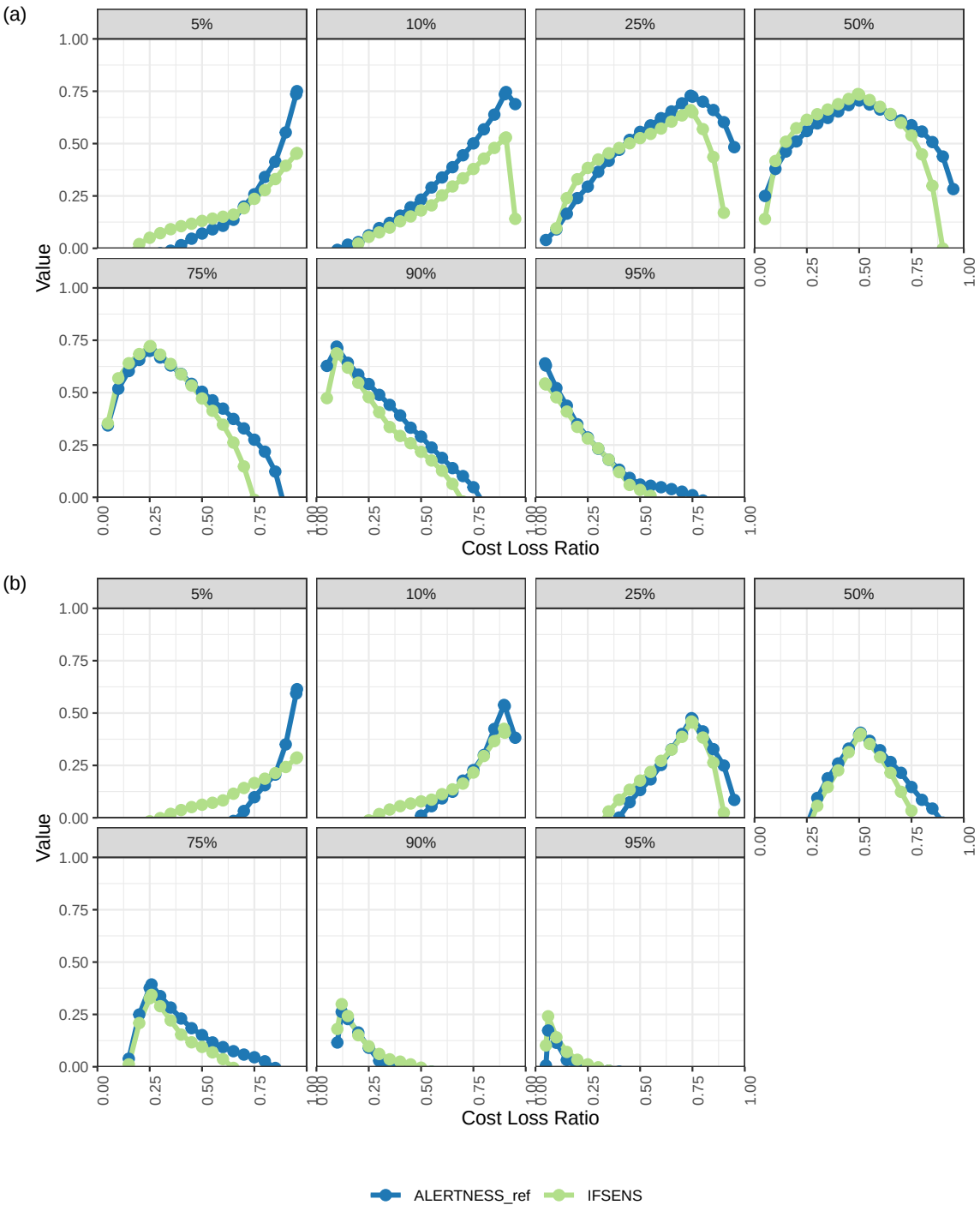

Figure 48: Economic value for 2m relative humidity forecasts during SOP 2 for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values at lead times of (a) 12 hours (b) 24 hours.

5.2.4 12 hour precipitation

Forecasts for 12 hour accumulated precipitation are verified at 06 and 18 UTC each day - this roughly separates the precipitation into day time and night time components. Furthermore, the largest number of stations is available for 12h precipitation at these hours. These stations are shown in Fig. 24. It should be noted that about half of the stations only have observations for 18 UTC (i.e. lead times of 18 and 42 hours).

Summary scores for 12 hour accumulated precipitation at lead times 18, 30 and 42 hours for SOP 2 are shown in Fig. 49. On the first day of the forecast, both ALERTNESS_ref and IFSENS have a similar RMSE, but as the forecast progresses, the RMSE of ALERTNESS_ref increases through the night time precipitation and continues to increase for the day time precipitation on the second day, while the RMSE for IFSENS is roughly the same for the day time precipitation on the second day as the night time precipitation (Fig. 49(a)). For the day time on both the first and second days, the spread for ALERTNESS_ref is almost equal to the RMSE, which a desired feature for an ensemble, but the fact that the RMSE for ALERTNESS_ref is larger than that for IFSENS is less desirable. In terms of the CRPS (Fig. 49(b)), ALERTNESS_ref is lower than IFSENS for day time precipitation on both the first and second days of the forecast, and is about the same for the night time precipitation. The ensemble mean has a negative bias, that becomes increasingly negative throughout the forecast, for ALERTNESS_ref, while IFSENS has a large positive bias for the day time precipitation on the first day of the forecast that is much smaller for the night time precipitation and increases again slightly for the second day (Fig. 49(c)). Although the ensemble spread and RMSE suggest near optimal dispersion for ALERTNESS_ref for daytime precipitation (Fig. 49(a)), the rank histogram, which combines both the day time and night time precipitation data, is dominated by the negative bias signal, while the rank histogram for IFSENS suggests under dispersion (Fig. 49(d))

As for SOP 1, the performance of the models is further assessed for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed 12h accumulated precipitation, where it is greater than zero, for each lead time. SOP 2 was a relatively dry period for most of the region so this results in a fairly small data set with 596, 113 and 683 cases for the 18, 30 and 24 hour lead times respectively. The evolution of the thresholds for these percentiles with lead time is shown in Fig. 50. Note that the 5th and 10th percentiles had almost exactly the same values so from hereon in, the 10th percentile is not shown for 12h accumulated precipitation verification. There is no clear diurnal cycle in the the thresholds.

The Brier Skill Score, using the sample climatology as reference, suggests that day time precipitation (lead times 18 and 42 hours) performance for ALERTNESS_ref is superior to IFSENS for all percentiles up to and including the 75th percentile (Fig. 51). For night time precipitation, ALERTNESS_ref has a higher Brier Skill Score than IFSENS for the 5th and 75th percentiles while the Brier Skill Scores are comparable for the 25th and 75th percentile. For the 95th percentile, neither ALERTNESS_ref nor IFSENS have a positive Brier Skill Score throughout the forecast and for the 90th percentile ALETRNESS_ref has a slightly positive Brier Skill Score for daytime precipitation on both the first second days and IFSENS has a negative Brier Skill Score for both 18 and 30 hour lead times, but a positive Brier Skill Score, that is marginally higher than that for ALERTNESS_ref, for day time precipitation on the second day of the forecast.

The reliability for the first day of the forecast suggests that ALERTNESS_ref is close to perfect reliability

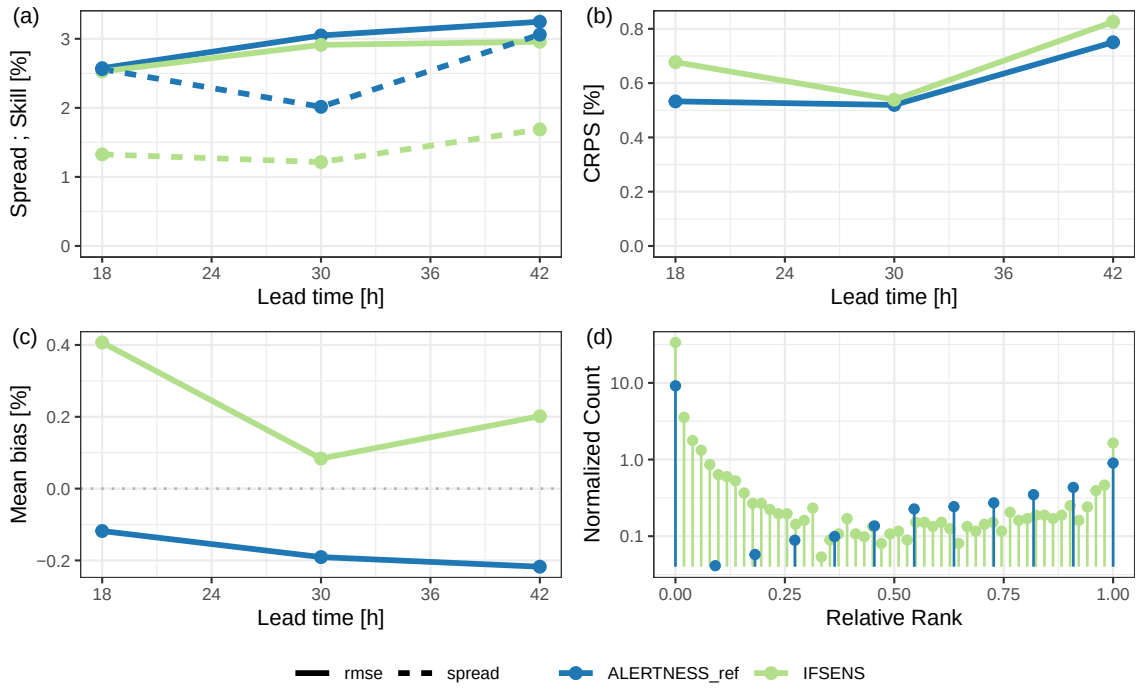

Figure 49: Summary verification scores for 12 hour accumulated precipitation during SOP 2: (a) RMSE and spread, (b) CRPS, (c) Bias of the ensemble mean and (d) Normalized relative rank histogram.

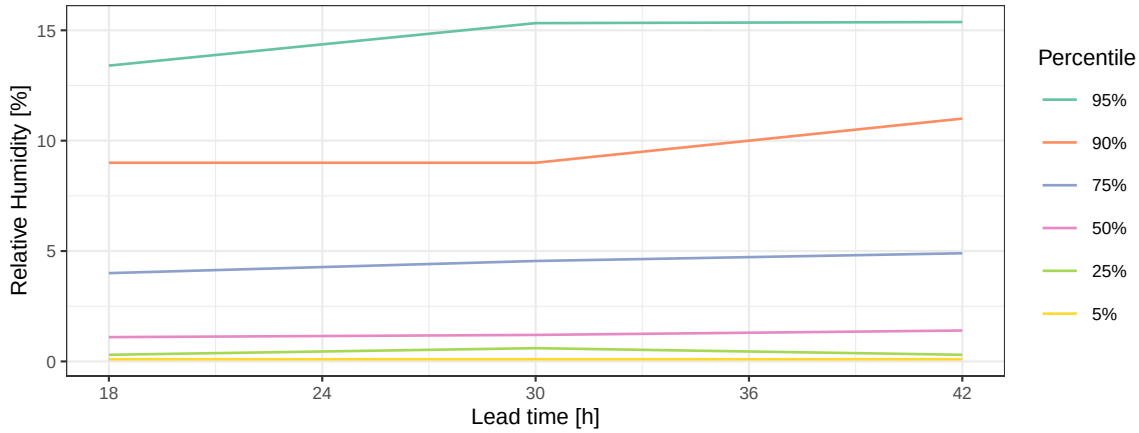

Figure 50: Thresholds used for categorical scores for 12h accumulated precipitation during SOP 2 derived from the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed values that were greater than zero valid at each lead time.

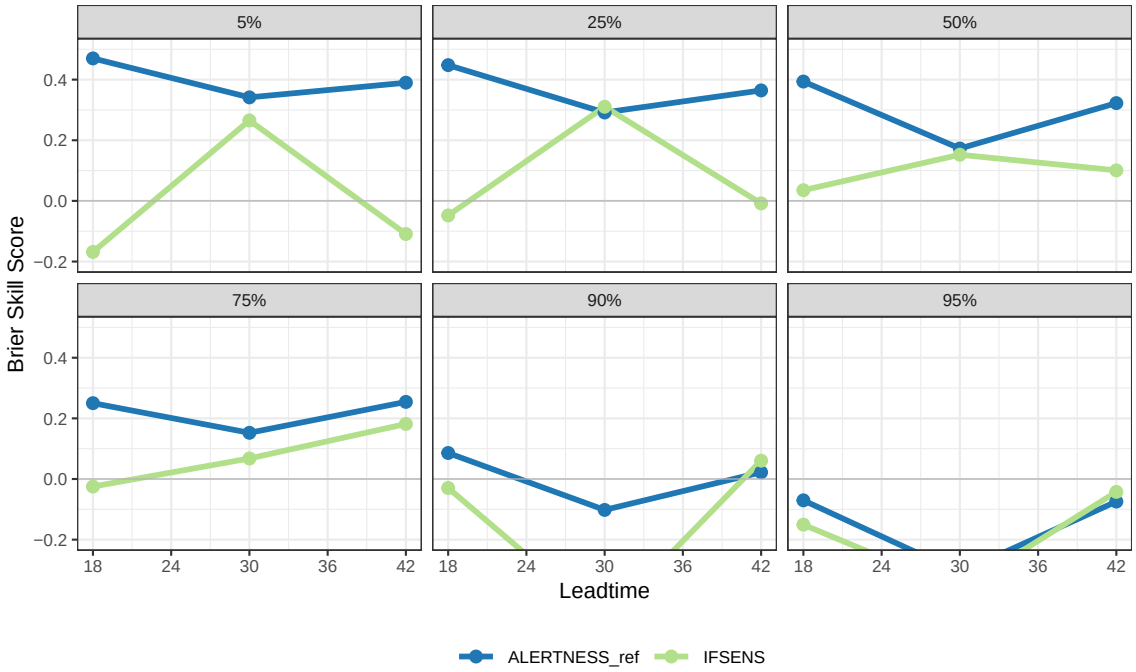

Figure 51: Brier Skill Score for 12 hour accumulated precipitation during SOP 1 for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed 12 hour accumulated precipitation greater than zero. The sample climatology is used as the reference forecast.

for all percentiles up to the 90th percentiles while IFSENS over forecasts the probabilities (Fig. 52(a)), while for the 95th percentile there were too few forecasts to obtain meaningful reliability plots. The ROC curves, however, suggest that IFSENS performs better than ALERTNESS_ref in terms of hit rate up to and including the 75th percentile, though ALERTNESS_ref tends to have a lower false alarm rate (Fig. 52(b)). For the night time precipitation, there were generally too few forecasts with precipitation to obtain meaningful reliability diagrams (Fig. 53(a)), while the ROC suggests that the difference between IFSENS and ALERTNESS_ref is larger for the night time than for the day time ((Fig. 53(a) cf Fig. 52(b)), mostly due to lower hit rates for ALERTNESS_ref. The reliability and ROC performance for day time precipitation on the second day is comparable to that on the first day.

The economic value suggests that ALERTNESS_ref provides more value than IFSENS for users with high cost-loss ratios for percentiles up to and including the 75th for day time precipitation, while IFSENS appears to provide more value for the 90th percentile (Fig. 54(a)). For night time precipitation, IFSENS provides more value than ALERTNESS_ref albeit for roughly the same range of cost-loss ratios (Fig. 54(b)).

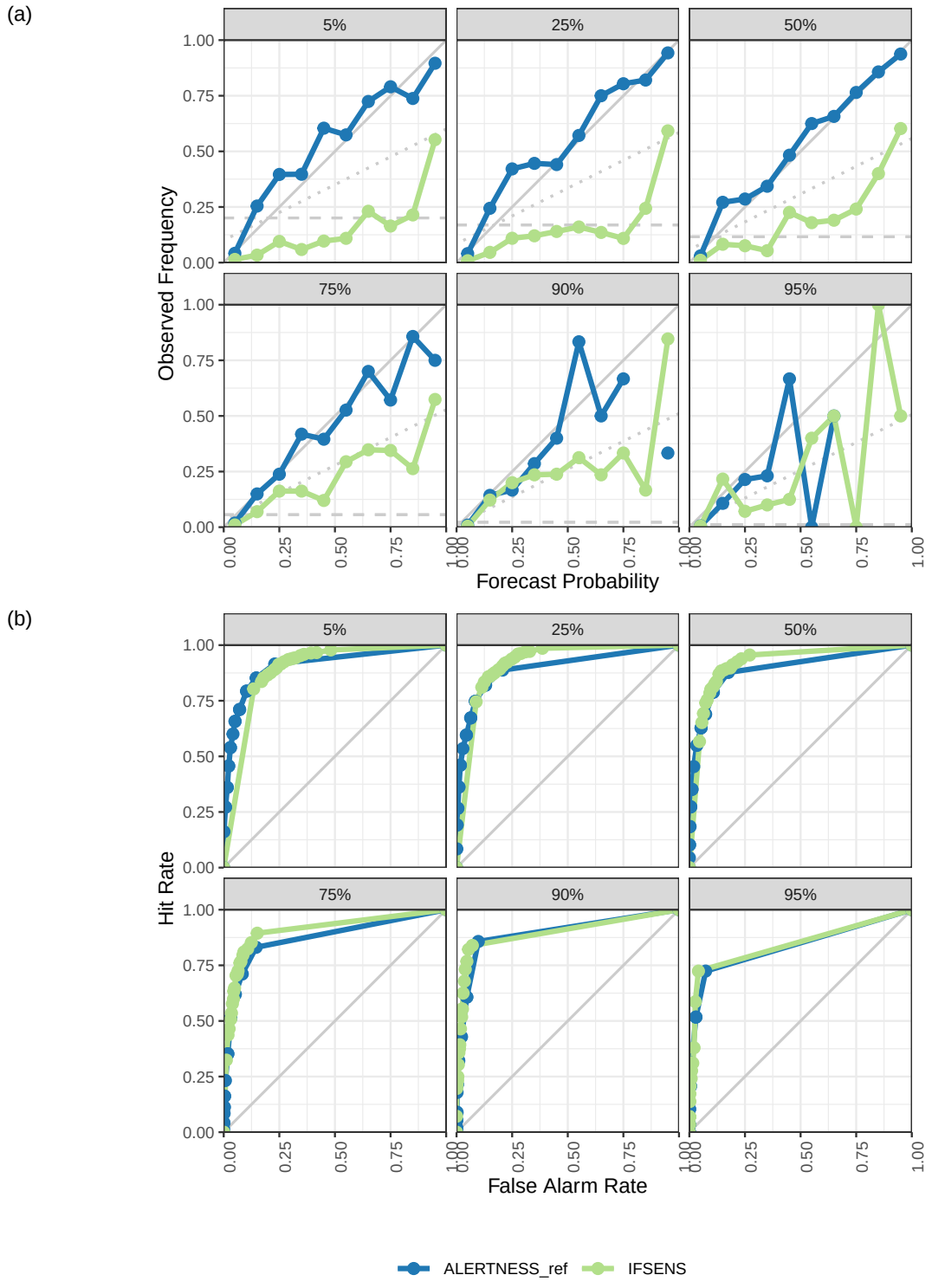

Figure 52: Verification for 12 hour accumulated precipitation during SOP 2 at a lead time of 18 hours and for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed 12 hour accumulated precipitation greater than zero for (a) reliability and (b) ROC.

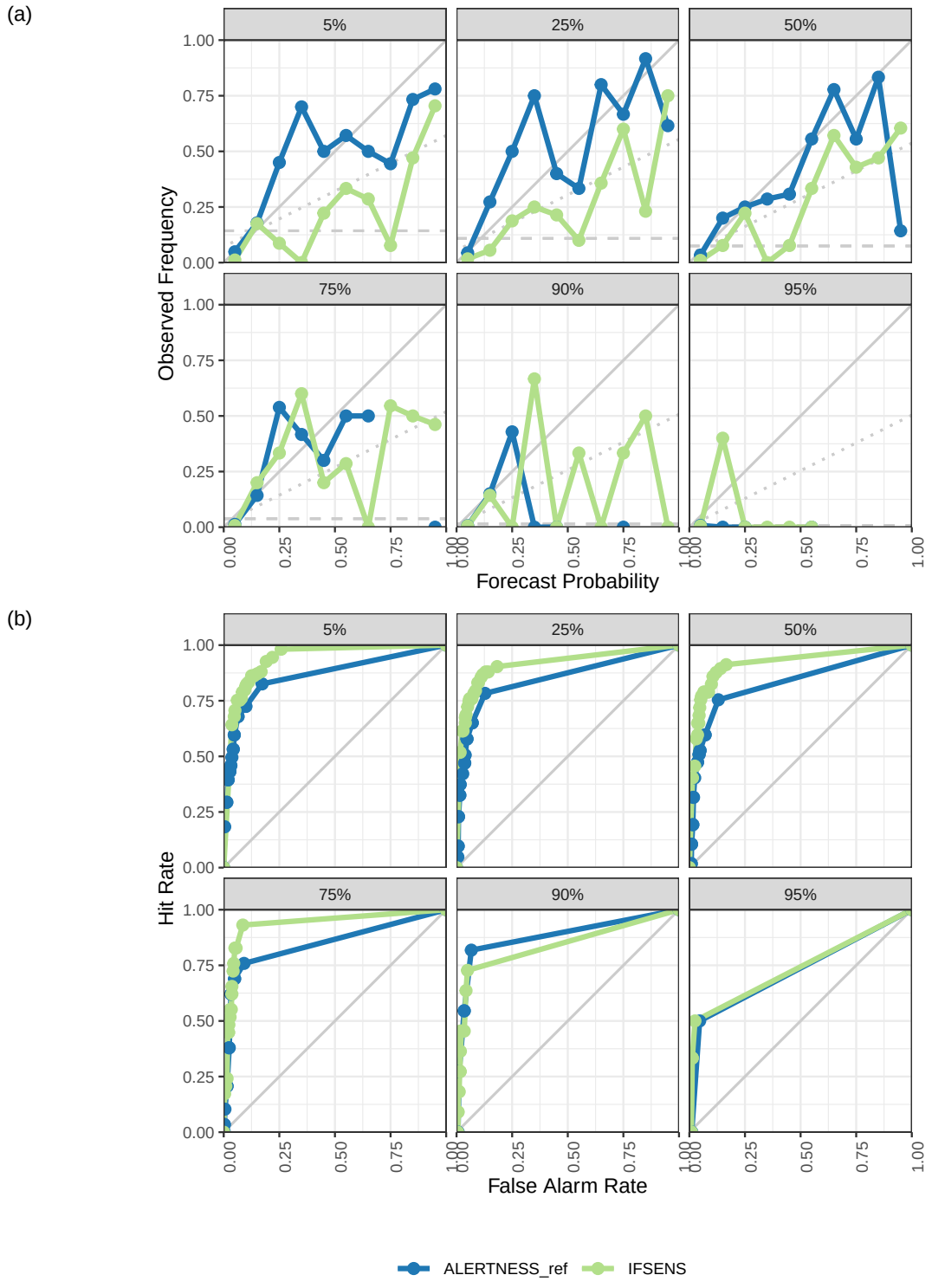

Figure 53: Verification for 12 hour accumulated precipitation during SOP 2 at a lead time of 30 hours and for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed 12 hour accumulated precipitation greater than zero for (a) reliability and (b) ROC.

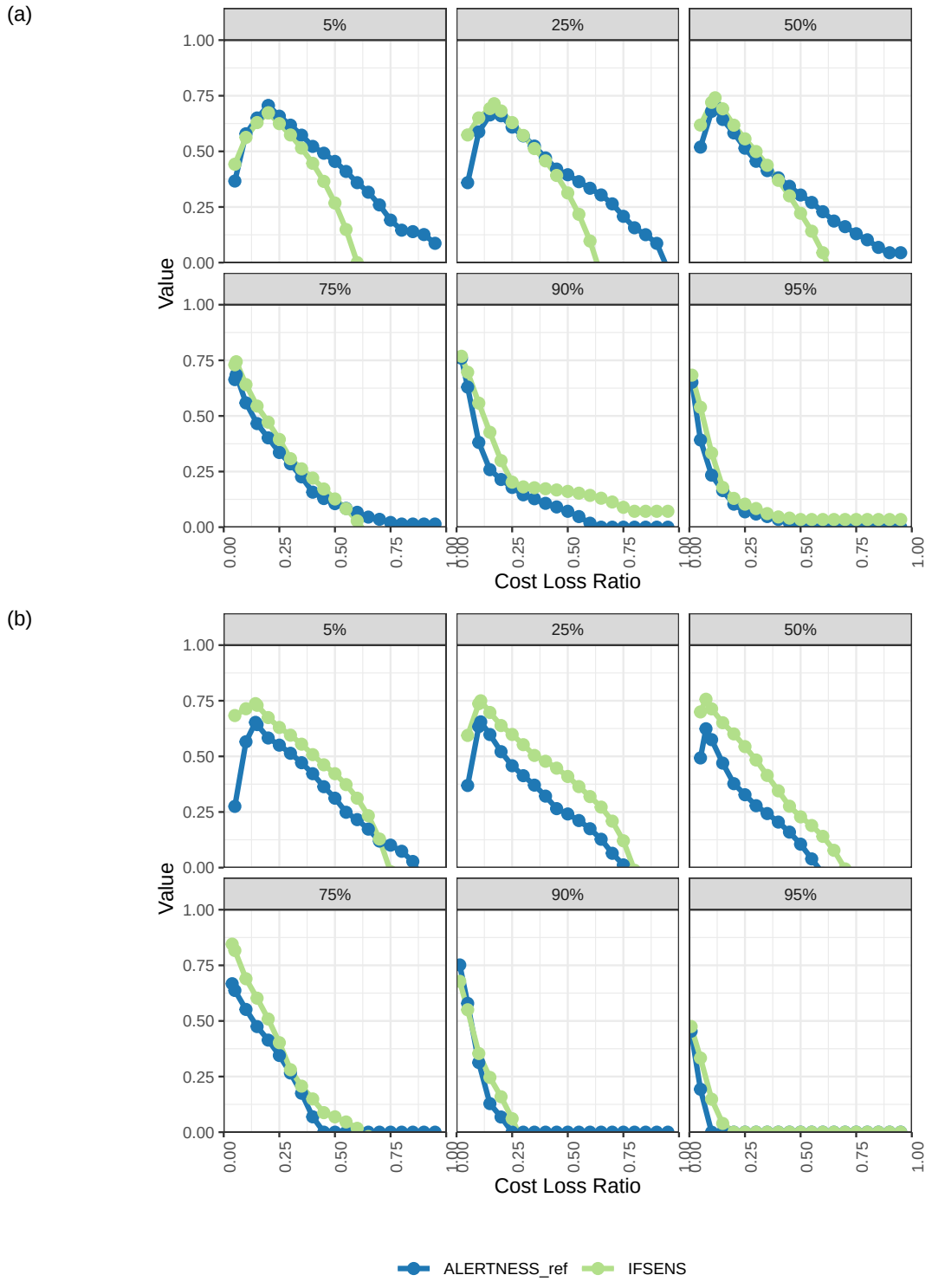

Figure 54: Economic value for 12 hour accumulated precipitation forecasts during SOP 2 for the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the observed 12 hour accumulated precipitation gretaer than zero at lead times of (a) 18 hours (b) 30 hours.

6 Discussion

The main purpose of this work is to provide a benchmark, ALERTNESS_ref, against which future improvements, developed within Work Package 4 of the ALERTNESS project, to the EPS implementation of the AROME-Arctic model can be measured. As a starting point the aim was to be at least comparable to the IFSENS global EPS from ECMWF, which is the only EPS that is able to provide operational forecasts for the region of interest in ALERTNESS. The performance of ALERTNESS_ref relative to IFSENS is measured using well tested verification scores for ensembles computed by comparison with available observations at surface weather stations. The domain used for verification is that of the operational AROME-Arctic deterministic model that is dominated by ocean surfaces. Unfortunately very few conventional observations are available over these ocean areas, so by having nearly all the stations used in the verification over land (Fig. 4, Fig. 11, Fig. 17, Fig. 24), it is not possible to gain a full insight into the performances of ALERTNESS_ref and IFSENS over the whole AROME-Arctic domain. It is hoped that developments in ALERTNESS Work Package 1 will enable ensemble performance to be verified over ocean and sea ice areas using new observation sources such as those derived from satellite radiances.

The periods chosen for which to do the verification were intended to be representative of winter (SOP 1) and summer (SOP 2). During SOP 1 temperatures were typically below $0°C$ and the 5th percentile dropped to as low as $-25°C$ during the night (Fig. 6), while the summer was characterized by the 95th percentile temperature being as high as $30°C$ during the day (Fig, 32). Therefore, there were some quite extreme temperatures against which to test the performance of the models. For other parameters, however, the weather was somewhat benign with 95th percentile wind speeds of 10 ms^{-1} during SOP 1 (Fig. 13) and 8 ms^{-1} during SOP 2 (Fig. 38), and 95th percentile 12 hour accumulated precipitation of up to 6 *mm* during SOP 1 (Fig. 26) and up to 15 *mm* during SOP 2 (Fig. 50). The verification of precipitation is further complicated by the fact that the observation statistics are dominated by zeros leaving only a relatively small number of cases from which to compute categorical scores.

A key finding is that, for all parameters, ALERTNESS_ref is better dispersed than IFSENS. Normalized rank histograms have normalized counts closer to 1 for ALERTNESS_ref than for IFSENS, and the standard deviation of the ensemble members with respect to the ensemble mean, the usual method for quantifying the spread of an ensemble, is higher. This is despite ALERTNESS_ref only have ten perturbed members, as opposed to the 50 in IFSENS. A reason for this may be that IFSENS is designed for medium range forecasts and therefore doesn't look to maximize spread in the short range. Indeed one of the perturbation methods used in IFSENS is to use singular vectors to perturb the model in a way that maximizes the error growth at 48 hours lead time. Furthermore, IFSENS uses perturbations to the tendencies resulting from the model's physics parameterizations to estimate the model error as the forecast progresses (SPPT: Stochastically Perturbed Parameterization Tendencies). Indeed for most parameters the spread of IFSENS is seen to grow steadily throughout the 48 forecast hours that are considered here while the spread for ALERTNESS_ref tends to remain roughly the same (if the diurnal cycle is ignored) from 0 to 48 hours (e.g. Fig. 12).

In terms of the parameters investigated, it is clear that ALERTNESS_ref is superior to IFSENS for 10m wind speed, and this superiority exists for both SOP 1 and SOP 2 for all thresholds. However, it was difficult to

objectively compare the models for extreme winds since the number of observed extreme wind events was small. For 2m temperature, ALERTNESS_ref is generally better than IFSENS, though during the night time there is an increase in RMSE that is not reflected in the spread - this is the case for both models. To further investigate the cause of the increased RMSE during the night, the computation of the RMSE and the bias of the ensemble mean is split into groups of observed 2m temperature at $2.5°C$ intervals for ALERTNESS_ref and weighted by the number of cases for that interval. Fig. 55 shows the weighted RMSE at each lead time during SOP 2 and it is clear that for the night time hours, the RMSE is dominated by errors for 2m temperatures in the range $7.5°C - 20°C$, with the bias showing a strong warm bias for the same temperature range (Fig. 56) for these hours. The bias also suggests that while the mid range temperatures are over estimated, the warmer temperatures tend to be under estimated, although the latter is the case for all lead times. A similar signal is seen for SOP 1 (not shown) with positive bias for the mid-range temperatures at night and a small negative bias for the warmer temperatures.

A reasonable question to ask at this point is whether this warm bias results from a systematic bias in the model as a whole, or the construction of the ensemble. The bias of the individual members at 24 hours lead time is therefore inspected with respect to that of the control member for the same temperature ranges. In this case no weighting is applied, but the number of cases for each temperature range impacts the significance, so only those temperature range with more than 100 cases are shown. Fig. 57 shows the difference in bias for each member compared to the control member and there is a clear signal that for the colder temperatures all of the members are warmer than the control member and for the warmer temperatures the majority of the members are cooler than the control member, while in the middle of the temperature range there is a near even distribution of members. In order to test if the signal for colder temperatures is robust, the same analysis is done for SOP 1 (Fig. 58). In this case it is clear that, for temperatures colder than $-2.5°C$, all of the perturbed members are warmer than the control member. This suggests that, although the control member tends to have a positive bias for 2m temperature at 24 hours lead time ($\sim 6°C$ for the $(-22.5, -20]°C$, but $\sim 0°C$ for the $(-12.5, -10]°C$ range), the perturbations are leading to an increased bias in the ensemble mean, especially for the extreme cold temperatures. This positive bias for the perturbed members during SOP 1 is actually replicated for all temperatures throughout the whole ALERTNESS_ref forecast for SOP 1, while this is not the case for ALERTNESS_ref for SOP 2, or at all for IFSENS (Fig. 59).

The results for 2m relative humidity are troublesome in that the night time performance is especially poor for higher relative humidities. A similar analysis to that for 2m temperature, with the verification done for 5% relative humidity bands, was done. A concern for SOP 2 was that there was a large increase in the errors in ALERTNESS_ref during the night time that was not present for IFSENS and was not reflected in the ensemble spread. Fig. 60 shows the RMSE, weighted by the number of cases, at all lead times for SOP 2 for 2m relative humidity in ALERTNESS_ref. It is clear that at the 0, 24 and 48 hour lead times, the RMSE is largest for the highest humidity bands, with the bias in the ensemble mean (Fig. 61) suggesting that there is a sizable negative bias for these high humidities. By looking at the biases of the perturbed members relative to the control member at 24 hours lead time (Fig. 62), it is clear that as the 2m relative humidity becomes larger, all of the perturbed members have a more negative bias than the control member. For SOP 1, the largest errors in ALERTNESS_ref are seen at the 15 and 39 hour lead times, so the biases of the perturbed members relative to the control member are shown in Fig. 63. Here, the more negative biases in the perturbed

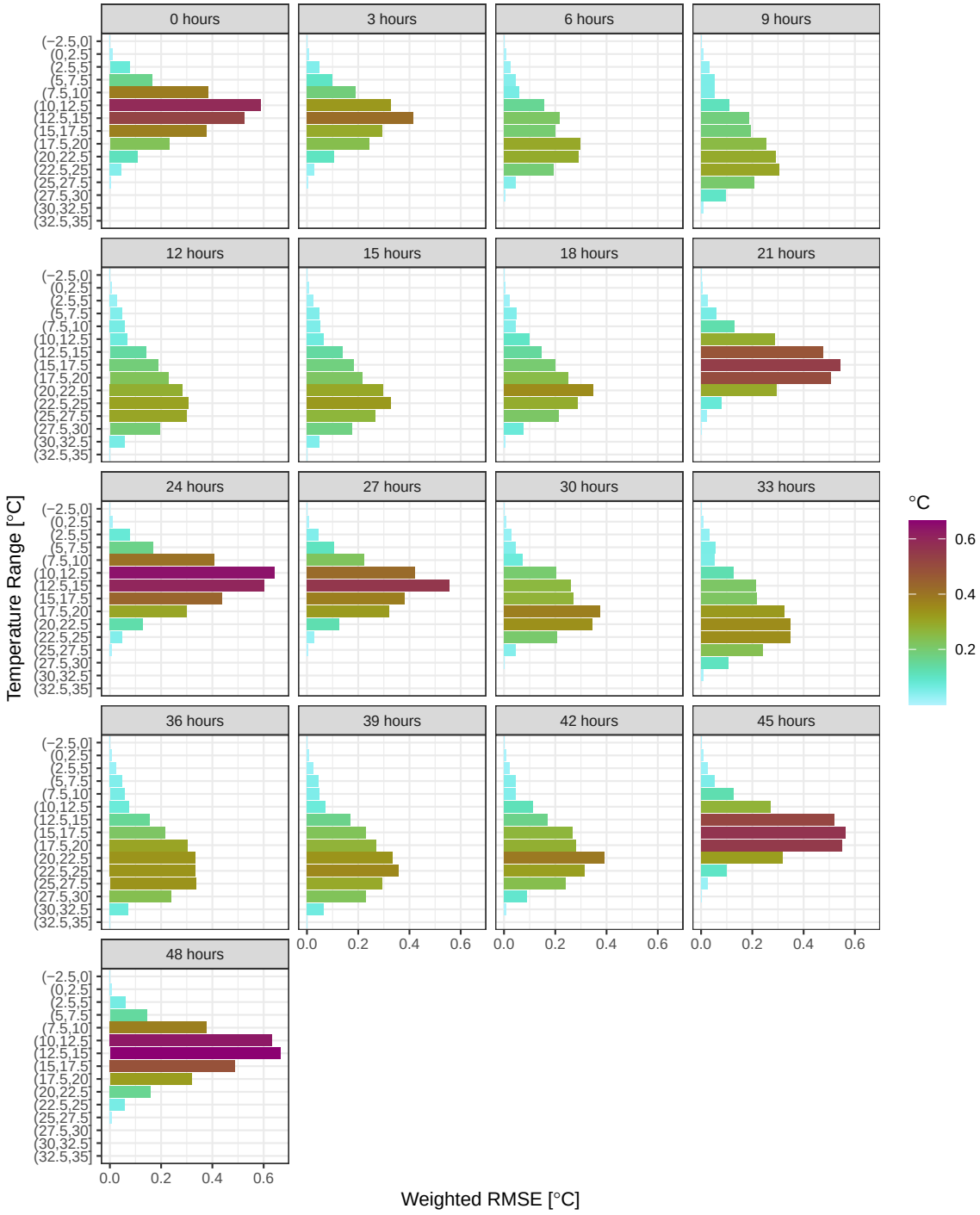

Figure 55: RMSE for 2m temperature for different temperature ranges weighted by the number of cases for the temperature range during SOP 2 for ALERTNESS_ref at all lead times.

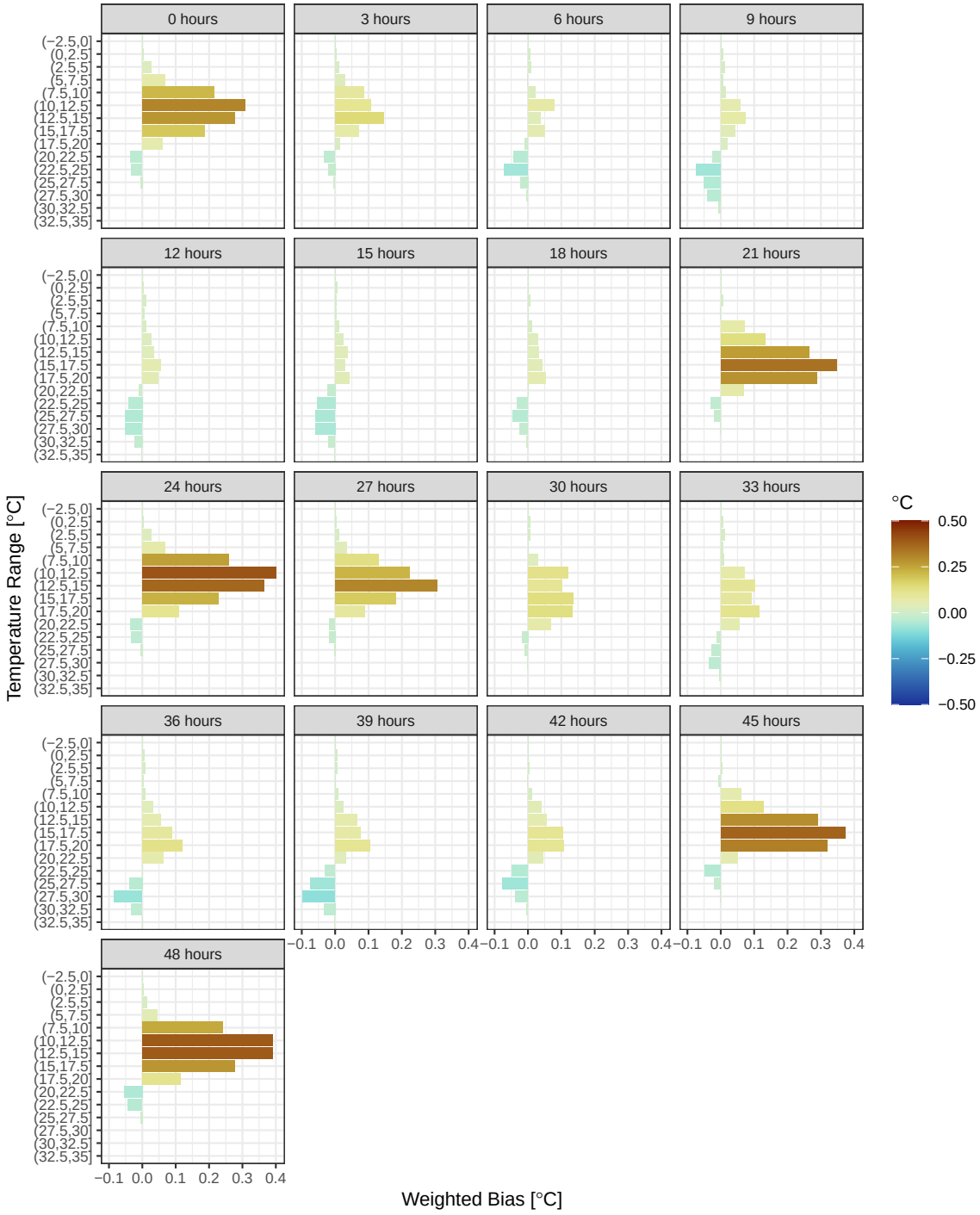

Figure 56: Bias of the ensemble mean for 2m temperature for different temperature ranges weighted by the number of cases for the temperature range during SOP 2 for ALERTNESS_ref at all lead times.

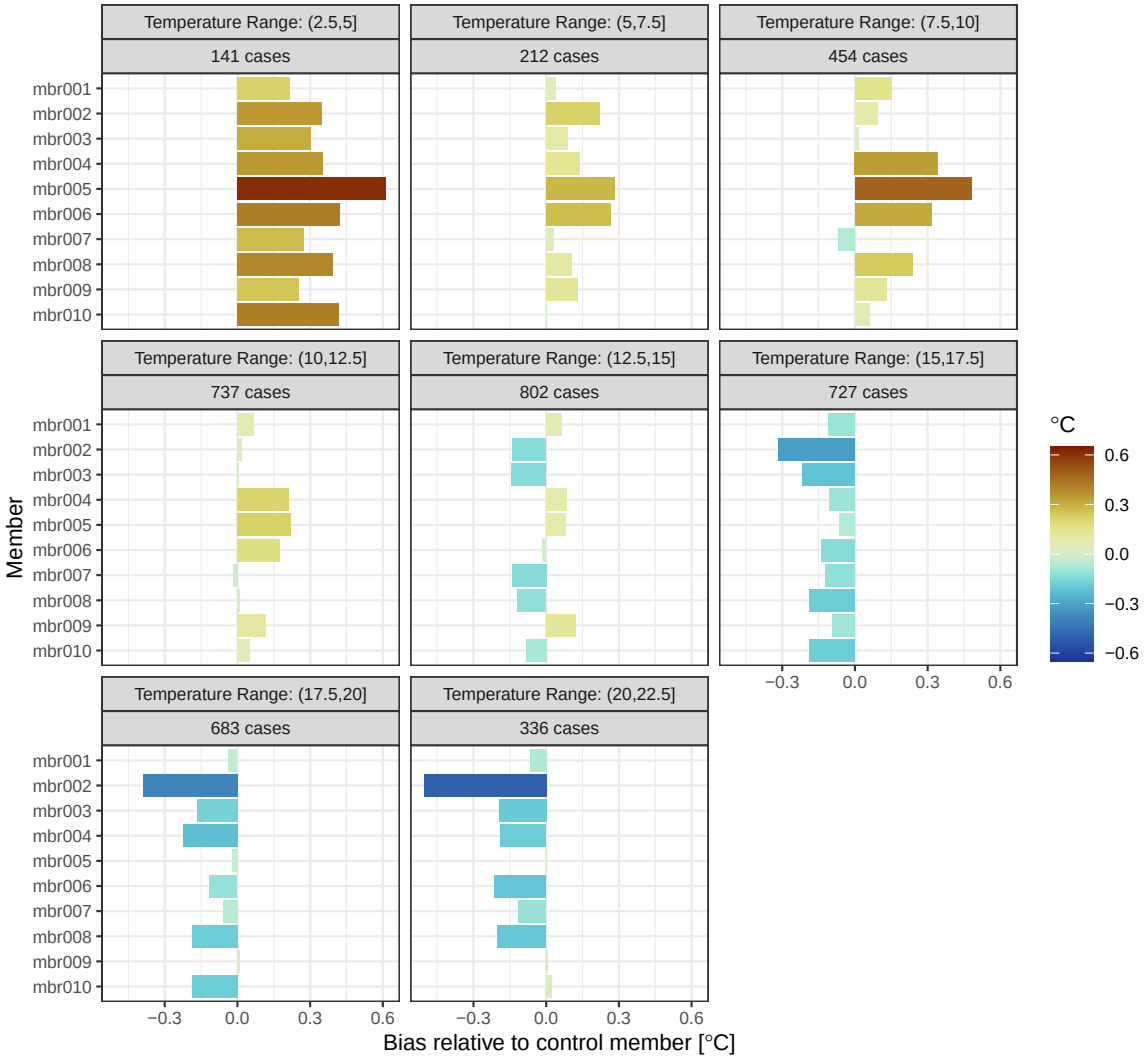

Figure 57: Bias for each ensemble member for 2m temperature for different temperature ranges relative to the control member (mbr000) during SOP 2 for ALERTNESS_ref at 24 hours lead time.

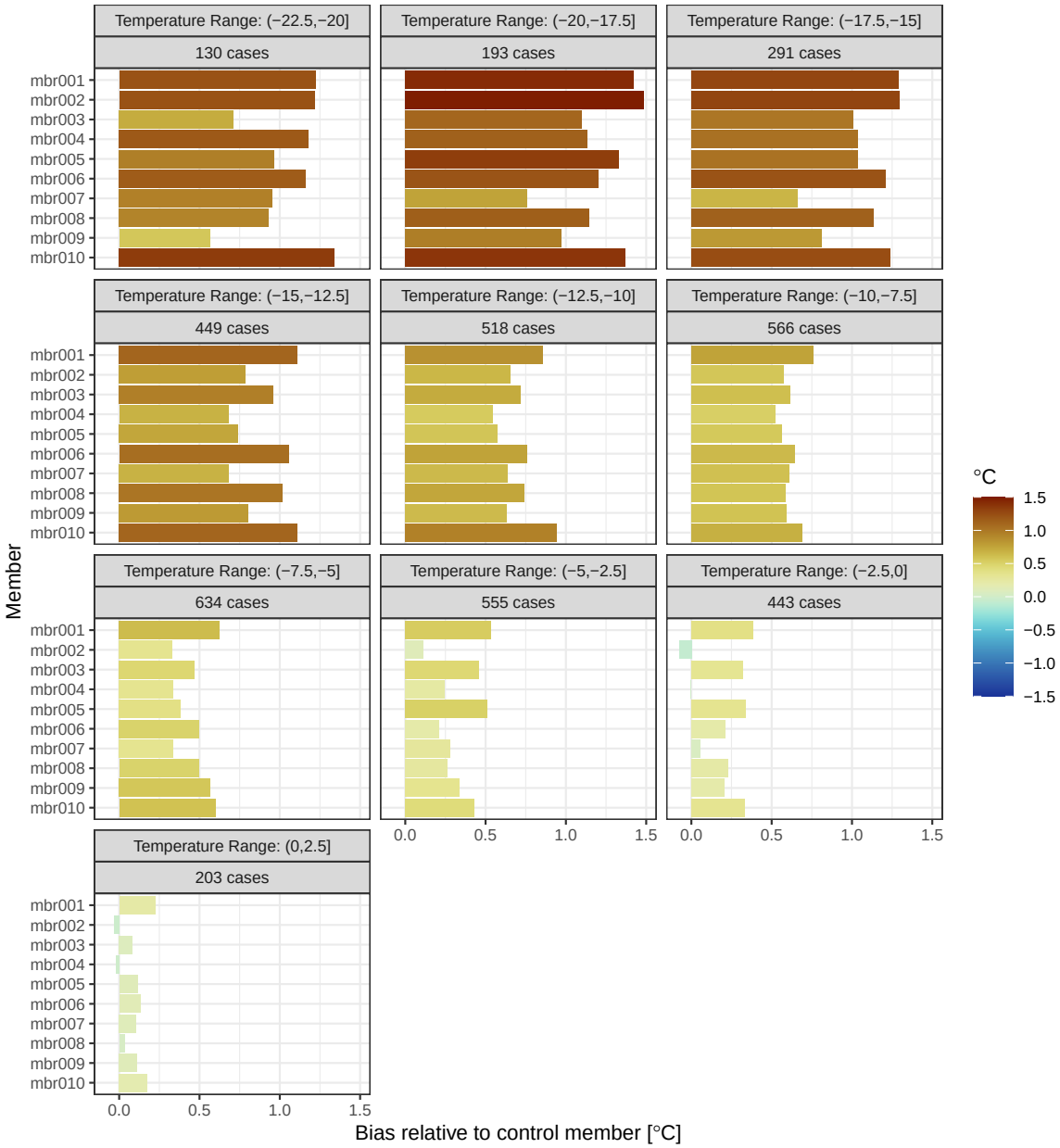

Figure 58: Bias for each ensemble member for 2m temperature for different temperature ranges relative to the control member (mbr000) during SOP 1 for ALERTNESS_ref at 24 hours lead time.

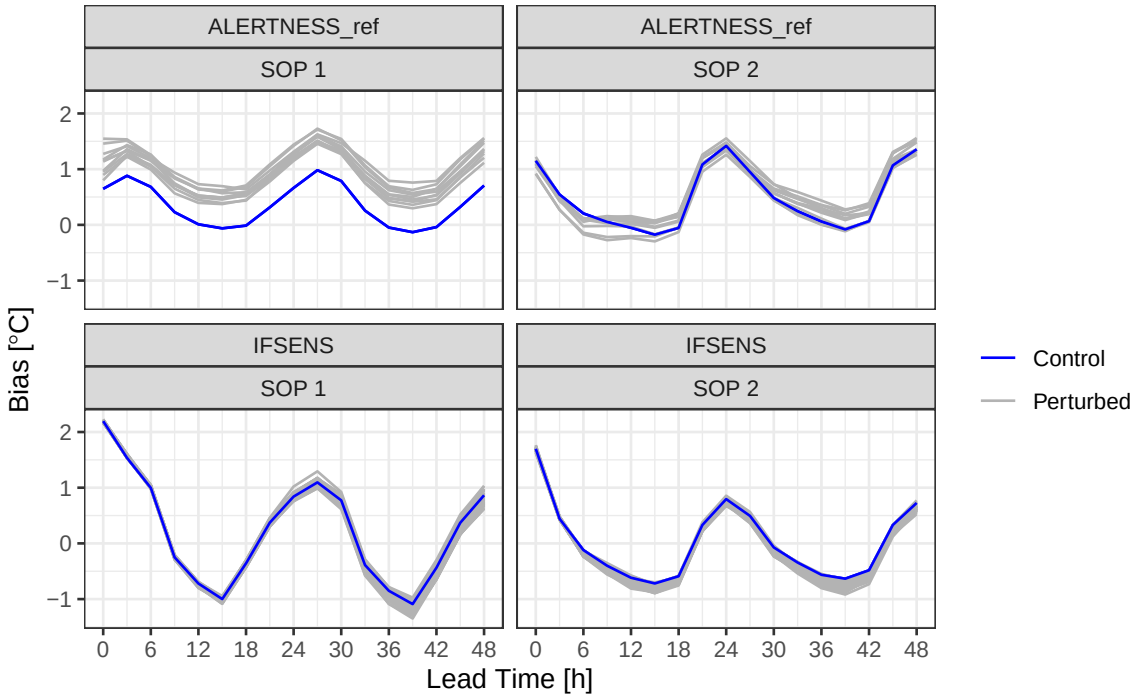

Figure 59: Bias for the control and each of the perturbed ensemble members for 2m temperature.

members compared to the control member are larger, especially for the cases with low relative humidity.

In the case of 2m relative humidity, there appears to be a systematic negative bias in the perturbed members, and an inspection of the biases for each member (Fig. 64) confirms this to be the case. Since relative humidity is related to both temperature and moisture, and there is a systematic positive bias in the perturbed members of ALERTNESS_ref during SOP 1, which would act to reduce the relative humidity, it would be useful to know whether there is also a negative bias in the moisture also contributing to the negative bias in the the 2m relative humidity. Forecasts of specific humidity were therefore also verified for each ensemble member of ALERTNESS_ref and the biases of the individual members are shown in Fig. 65. For SOP 1, the perturbed members have a more positive bias than the control member between lead times of 18 and 33 hours, suggesting that at this time the negative bias of 2m relative humidity in the perturbed members compared with the control member is being driven by the warm bias in 2m temperature. For SOP 2, however, all of the perturbed members have a negative bias in the specific humidity when compared with the control member and there is no systematic bias in the 2m temperature, suggesting that the negative bias in the perturbed members for the relative humidity is driven by the moisture.

It should be noted that the perturbed members being “drier” in terms if relative humidity is a known problem for the EPS implementation of the Harmonie AROME model and scientists in the HIRLAM consortium or actively working to understand the causes. At the time of writing, both the surface data assimilation and the surface perturbations in soil moisture are thought to be contributing.

It is unclear whether these biases in 2m relative humidity are reflected in the precipitation. However, there are key differences between the relative performances of ALERTNESS_ref and IFSENS during SOP 1 and SOP

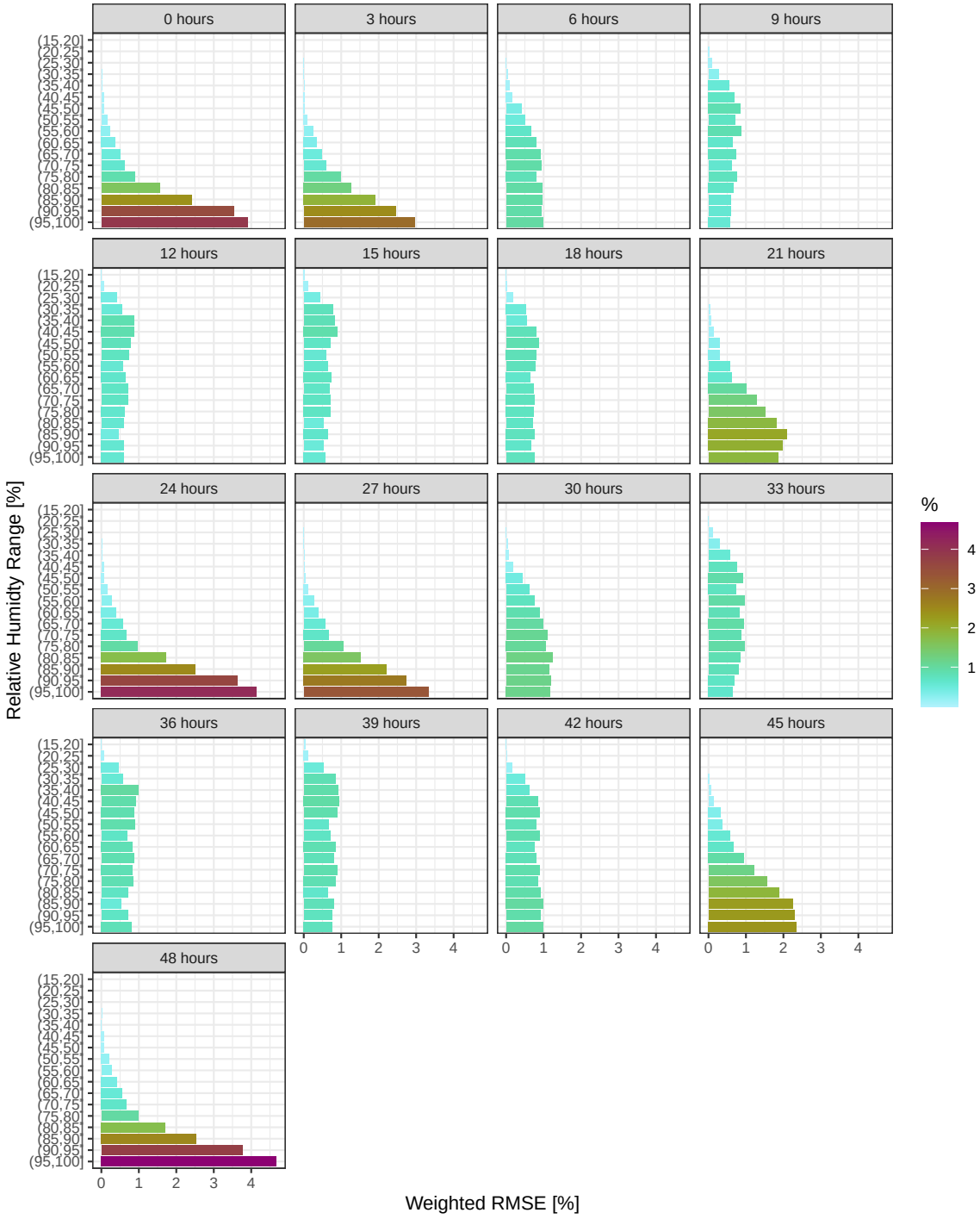

Figure 60: RMSE for 2m relative humidity for different relative humidity ranges weighted by the number of cases for the temperature range during SOP 2 for ALERTNESS_ref at all lead times.

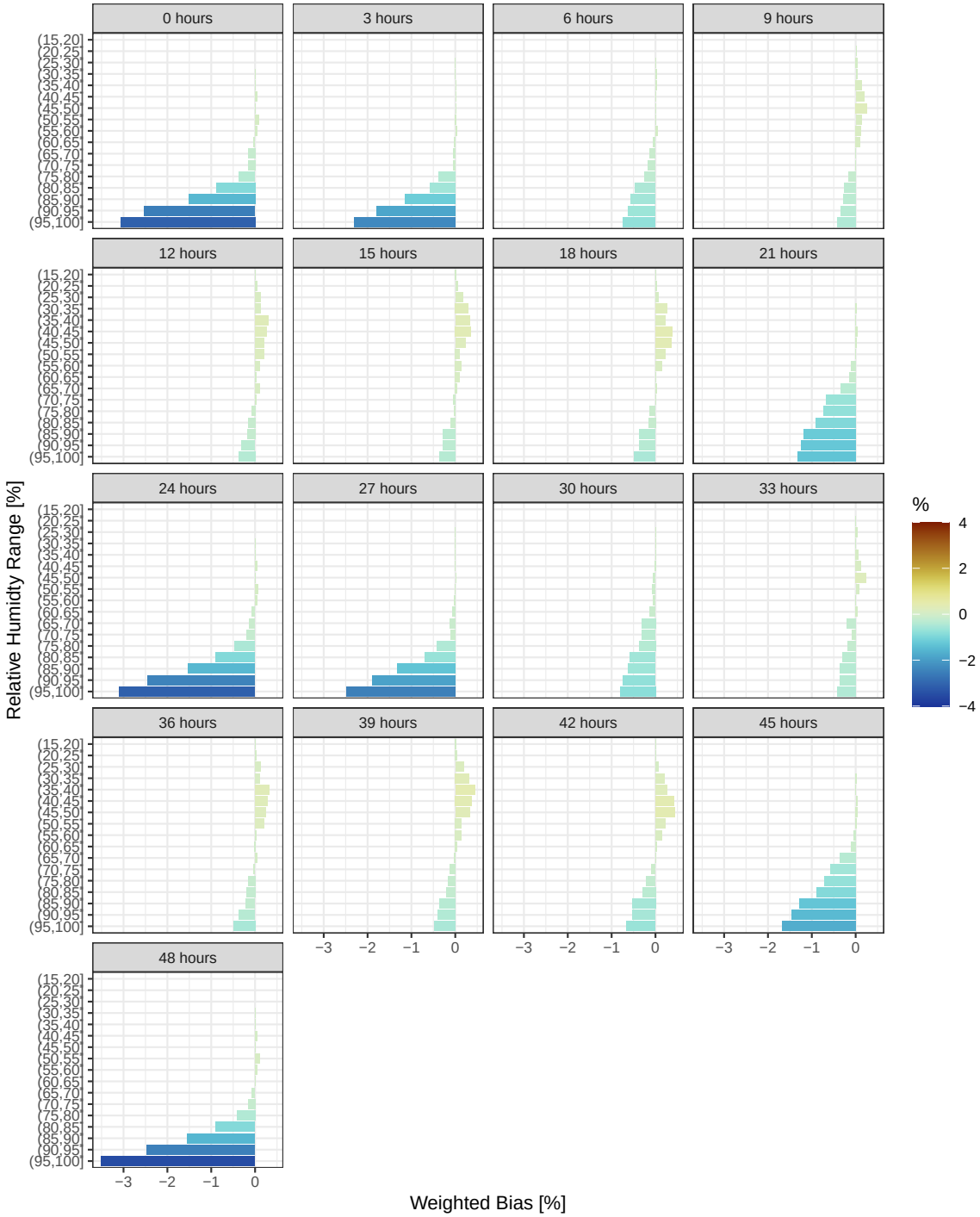

Figure 61: Bias of the ensemble mean for 2m relative humidity for different relative humidity ranges weighted by the number of cases for the temperature range during SOP 2 for ALERTNESS_ref at all lead times.

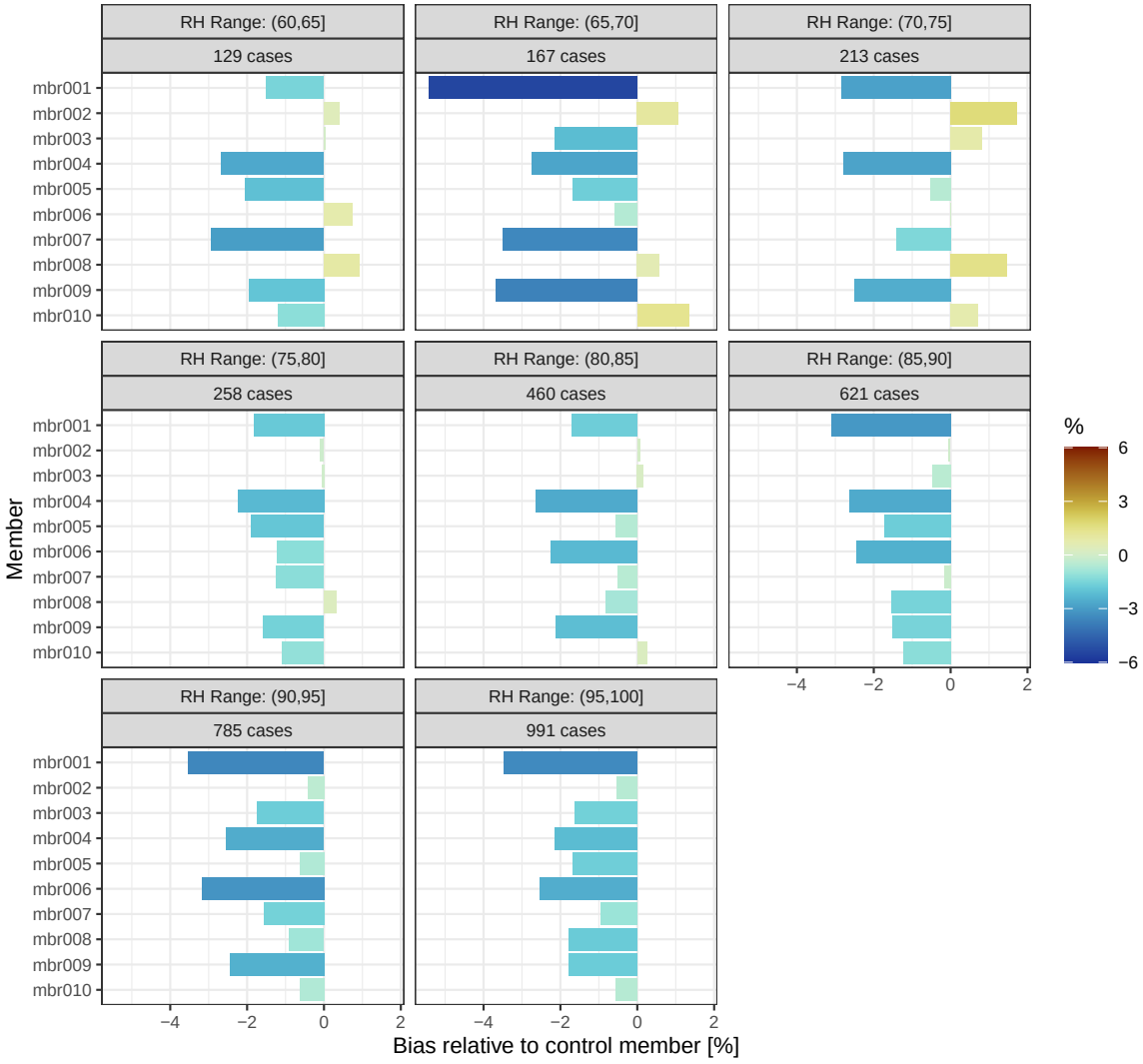

Figure 62: Bias for each ensemble member for 2m relative humidity for different relative humidity ranges relative to the control member (mbr000) during SOP 2 for ALERTNESS_ref at 24 hours lead time.

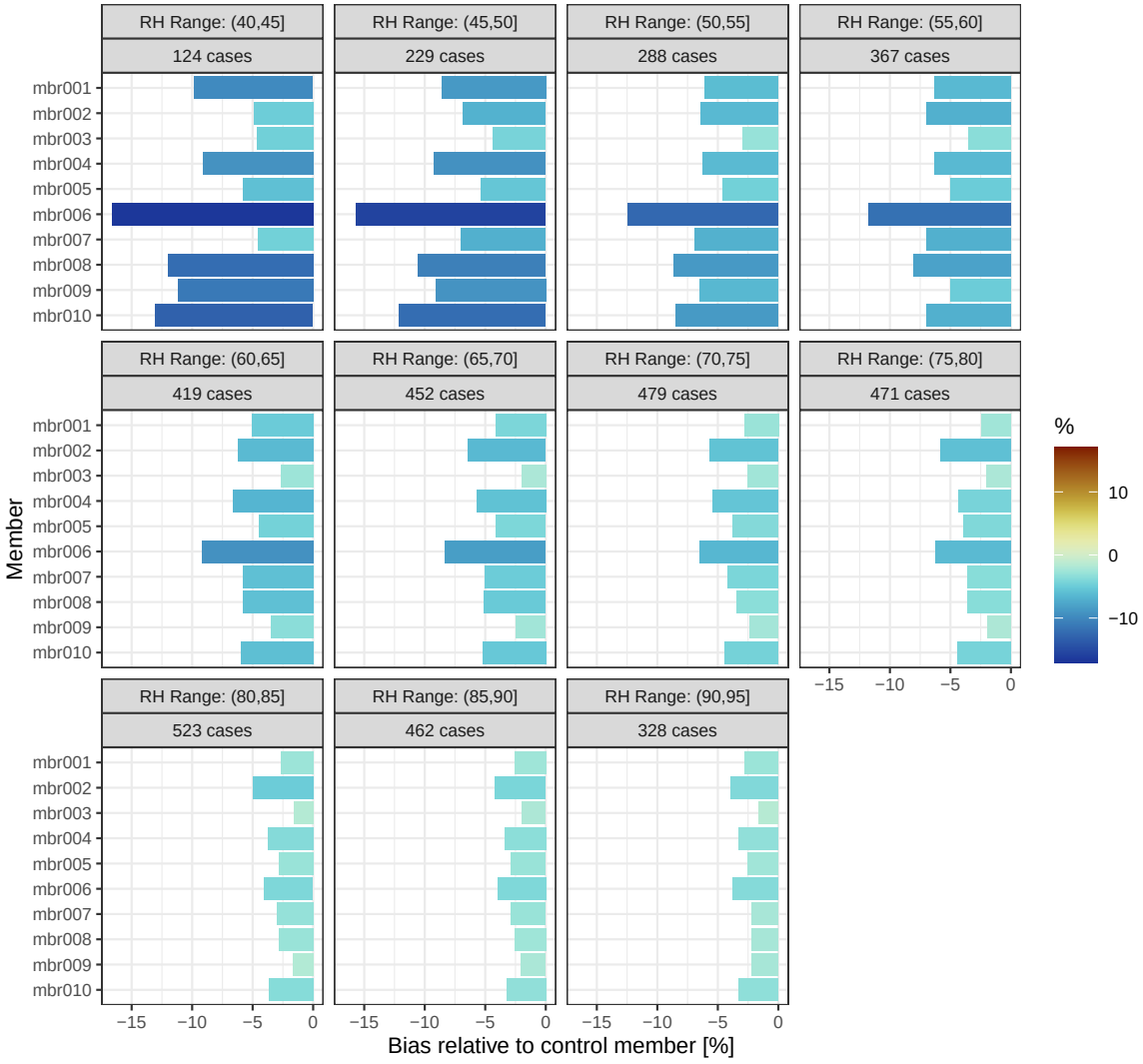

Figure 63: Bias for each ensemble member for 2m relative humidity for different relative humidity ranges relative to the control member (mbr000) during SOP 1 for ALERTNESS_ref at 15 hours lead time.

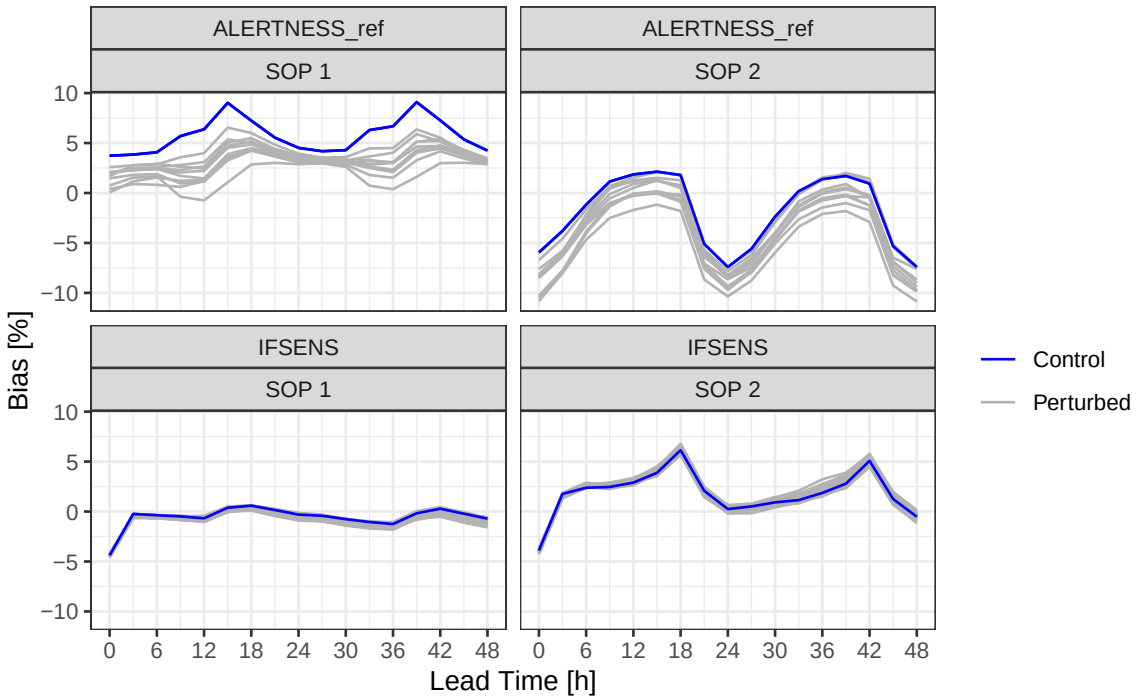

Figure 64: Bias for the control and each of the perturbed ensemble members for 2m relative humidity.

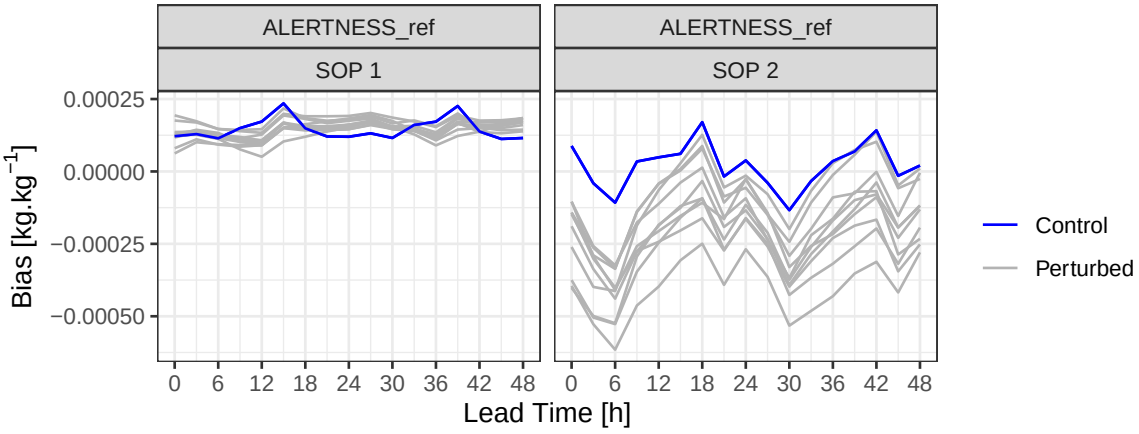

Figure 65: Bias for the control and each of the perturbed ensemble members for 2m specific humidity.

2. During SOP 2, daytime scores for ALERTNESS_ref are typically better than those for IFSENS, whereas for the night time ALERTNESS_ref is either worse, or comparable to IFSENS. However, during SOP 1 IFSENS generally had better verification scores than ALERTNESS_ref. This may be explained, to some extent, by the different processes that determine rainfall. During the day time in the summer (SOP 2), some of the precipitation is likely to be convective, a process that is explicitly permitted by ALETRNESS_ref due to the higher spatial resolution, but parameterized in IFSENS. It is likely this that gives ALERTNESS_ref the better day time scores in SOP 2. During the winter (SOP 1) and at night, the precipitation is likely to be larger in both spatial and temporal scales, which IFSENS is able to forecast well. It is often the case that, when comparing high spatial resolution models with models with a lower spatial resolution, the higher spatial resolution models are doubly penalized by near misses. That is, where precipitation is forecast close to a station, but not at the location of the station it is counted as a missed event at the location of the station and a false alarm at the location at which it occurred. This may not be the case for the lower resolution model as the pixels are large enough to get the "correct" forecast. By using 12 hour precipitation accumulations this double effect is mitigated to some extent as there are less likely to be small scale features. However, the fairest comparison between ALERTNESS_ref and IFSENS would be to use a spatial verification technique. Unfortunately, for this region there is currently no good quality spatial data set of precipitation observations available. Work in Work Package 1 may identify a suitable method to use in the future.

Another aspect to note is that, for 2m temperature and 2m specific / relative humidity in particular, the spread for ALERTNESS_ref does not increase with time at the same rate as the RMSE. For SOP 1, a linear fit to the spread for 2m temperature for ALERTNESS_ref does not increase at all with lead time while the RMSE does increase, and for 2m relative humidity the spread becomes smaller with increasing lead time (Fig. 66). For SOP 2, the results are similar although excessive spread for the 2m specific humidity becomes less excessive as lead time increases (Fig. 67). For IFSENS, there is a clear increase in spread with lead time for all parameters. A key difference between ALERTNESS_ref and IFSENS is that ALERTNESS_ref takes no account of the model uncertainty whereas IFSENS uses SPPT to estimate the model uncertainty. It is possible that it is this modelling of the model uncertainty that is helping the spread to increase as the forecast progresses, and that the planned introduction of SPP (Stochastically Perturbed Parameterizations) to the Harmonie AROME EPS for ALERTNESS will help the spread to increase with increasing lead time.

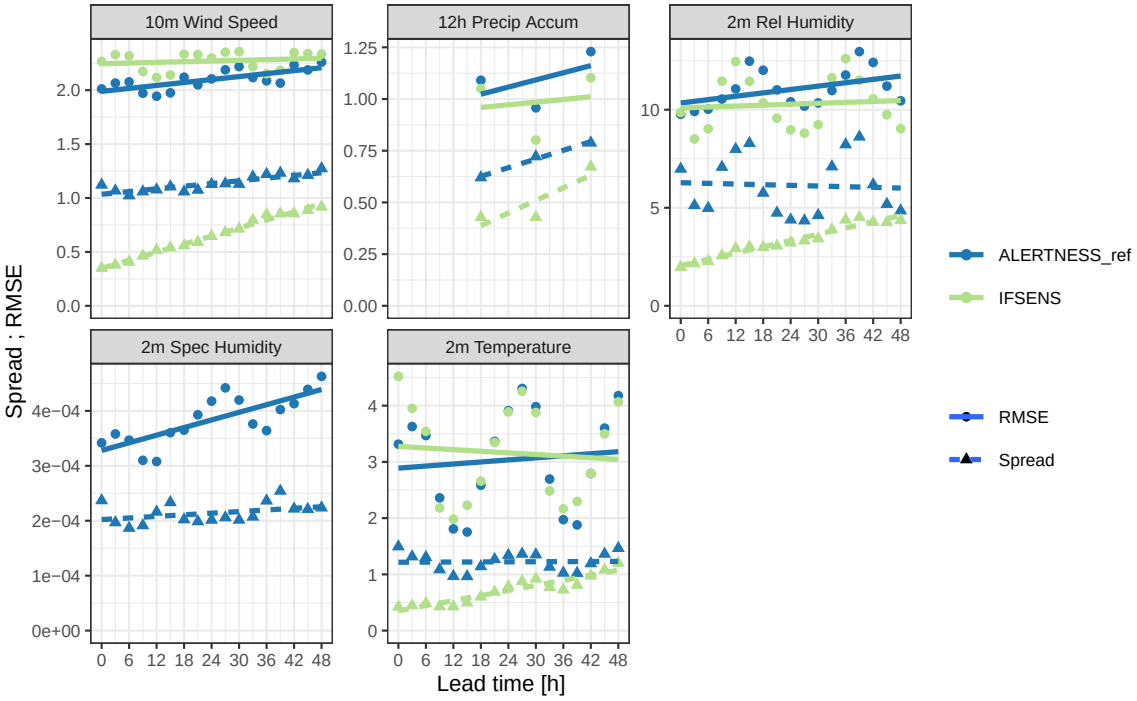

Figure 66: Spread and RMSE for all parameters with a linear model fit for SOP 1 (note that 2m specific humidity data were not available for IFSENS).

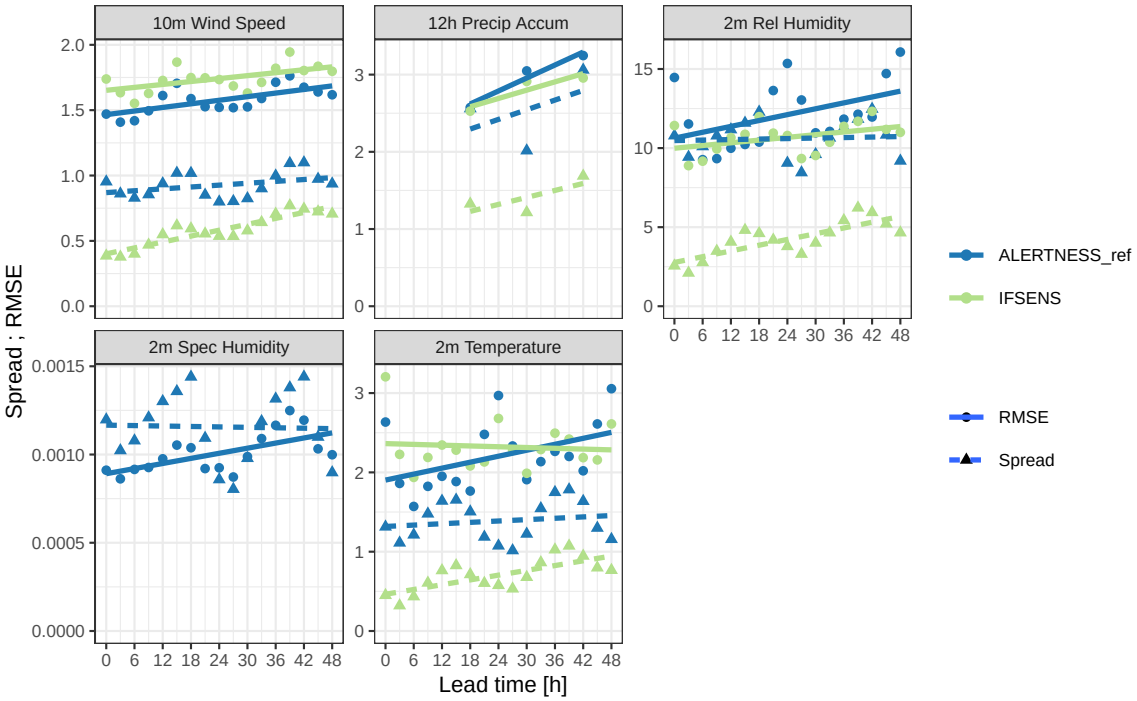

Figure 67: Spread and RMSE for all parameters with a linear model fit for SOP 2 (note that 2m specific humidity data were not available for IFSENS).

7 Conclusions and future plans

The aim of ALERTNESS Task 4.1 was to provide a reference set of EPS data (referred to as ALERTNESS_ref) against which future developments in ALERTNESS could be measured. Furthermore, the applicability of the Harmonie AROME EPS model for the Arctic region was to be assessed by comparing with IFSENS, the only operationally available EPS for the region. For most parameters ALERTNESS_ref was superior to, or equal to IFSENS in terms of verification scores over land. The main issue appears to be a systematic dry bias of the perturbed members compared with the control and a systematic warm bias for very cold temperatures. Both of these are known issues that are being addressed in the HIRLAM consortium. We consider the results reported herein to be a good starting point from which to further develop an EPS for the Arctic within the ALERTNESS project.

However, a major question mark remains about the performance of ALETRNESS_ref over the large ocean areas in the Arctic. Ongoing research in ALERTNESS work package 1 will identify methods and observations against which ALERTNESS_ref could be verified. The most straightforward of these is wind over the ocean as derived from the ASCAT satellite. However, it is believed that both the sea surface temperature and sea ice concentrations are major contributors to the uncertainty in weather forecasts for the Arctic, and ALERTNESS task 4.2 is addressing that by considering novel strategies for sea surface temperature and sea ice perturbations.

While the results for ALERTNESS_ref suggest relatively large uncertainties in the model initial conditions, these uncertainties may be better constrained by the introduction of an ensemble of data assimilation (EDA). EDA will be implemented into the EPS for the Arctic and its impact on the forecast assessed in ALERTNESS task 4.3. Furthermore, the error growth does not appear to be properly modelled in ALERTNESS_ref and the introduction of SPP to estimate the model uncertainty in ALERTNESS task 4.4 may go some way towards achieving a better growth in the forecast spread for some parameters.

8 Collaboration with other projects

Results from SOP 1 have been used in the APPLICATE project to assess the impact of removing perturbations to the sea surface temperature. This experiment revealed a problem in the way that sea surface and land temperature perturbations are treated in the model and is currently being investigated by model developers in HIRLAM.

The results from SOP 1 have also been shared with the Nansen Legacy project, and they are using these data to force an ocean and sea ice model for the period. The results from that work will then be used in ALERTNESS to provide an ensemble of sea surface temperature and sea ice boundary conditions to force an AROME-Arctic ensemble.

All results from ALERTNESS are also shared with the HIRLAM consortium with the aim reporting issues to the model developers so that issues with the model may be improved in future iterations.

9 Data availability

Data from the model runs are available in NetCDF format on the Norwegian Meteorological Institute thredds server at https://thredds.met.no/thredds/catalog/alertness/catalog.html.

10 Acknowledgements

This study was supported by the Norwegian Research Council Project 280573 'Advanced models and weather prediction in the Arctic: enhanced capacity from observations and polar process representations (ALERTNESS).' This work is a contribution to the Year of Polar Prediction (YOPP), a flagship activity of the Polar Prediction Project (PPP), initiated by the World Weather Research Programme (WWRP) of the World Meteorological Organization (WMO).

11 References

Batrak, Y., E. Kourzeneva, and M. Homleid, 2018: Implementation of a simple thermodynamic sea ice scheme, sice version 1.0-38h1, within the aladin–hirlam numerical weather prediction system version 38h1. *Geoscientific Model Development*, **11**, 3347–3368, https://doi.org/10.5194/gmd-11-3347-2018.

Bengtsson, L. and Coauthors, 2017: The harmonie–arome model configuration in the aladin–hirlam nwp system. *Monthly Weather Review*, **145**, 1919–1935, https://doi.org/10.1175/MWR-D-16-0417.1.

Bouttier, F., B. Vié, O. Nuissier, and L. Raynaud, 2012: Impact of stochastic physics in a convection-permitting ensemble. *Monthly Weather Review*, **140**, 3706–3721, https://doi.org/10.1175/MWR-D-12-00031.1.

——, ——, ——, and B. Ménétrier, 2016: Sensitivity of the arome ensemble to initial and surface perturbations during hymex. *Quarterly Journal of the Royal Meteorological Society*, **142**, 390–403, https://doi.org/10.1002/qj.2622.

Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare, 2008: The mogreps short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **134**, 703–722, https://doi.org/10.1002/qj.234.

Buizza, R., and T. N. Palmer, 1995: The singular-vector structure of the atmospheric global circulation. *Journal of the Atmospheric Sciences*, **52**, 1434–1456, https://doi.org/10.1175/1520-0469(1995)052%3C1434:TSVSOT%3E2.0.CO;2.

——, M. Milleer, and ——, 1999: Stochastic representation of model uncertainties in the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **125**, 2887–2908, https://doi.org/10.1002/qj.49712556006.

Buizza, R., M. Leutbecher, and L. Isaksen, 2008: Potential use of an ensemble of analyses in the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **134**, 2051–2066, https://doi.org/10.1002/qj.346.

Frogner, I.-L. and Coauthors, 2019: HarmonEPS—the harmonie ensemble prediction system. *Weather and Forecasting*, **34**, 1909–1937, https://doi.org/10.1175/WAF-D-19-0030.1.

Hacker, J. P. and Coauthors, 2011: The u.s. Air forceweather agency's mesoscale ensemble: Scientific description and performance results. *Tellus A: Dynamic Meteorology and Oceanography*, **63**, 625–641, https://doi.org/10.1111/j.1600-0870.2010.00497.x.

Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant, 2017: The met office convective-scale ensemble, mogreps-uk. *Quarterly Journal of the Royal Meteorological Society*, **143**, 2846–2861, https://doi.org/10.1002/qj.3135.

Keller, J. D., L. Kornblueh, A. Hense, and A. Rhodin, 2008: Towards a gme ensemble forecasting system: Ensemble initialization using the breeding technique. *Meteorologische Zeitschrift*, **17**, 707–718, https://doi.org/10.1127/0941-2948/2008/0333.

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

Liu, Y., and P. J. Minnett, 2016: Sampling errors in satellite-derived infrared sea-surface temperatures. Part i: Global and regional modis fields. *Remote Sensing of Environment*, **177**, 48–64, https://doi.org/https://doi.org/10.1016/j.rse.2016.02.026.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, **20**, 130–141, https://doi.org/10.1175/1520-0469(1963)020%3C0130:DNF%3E2.0.CO;2.

Marsigli, C., A. Montani, and T. Paccagnella, 2014: Provision of boundary conditions for a convection-permitting ensemble: Comparison of two different approaches. *Nonlinear Processes in Geophysics*, **21**, 393–403, https://doi.org/10.5194/npg-21-393-2014.

Masson, V. and Coauthors, 2013: The surfexv7.2 land and ocean surface platform for coupled or offline simulation of earth surface variables and fluxes. *Geoscientific Model Development*, **6**, 929–960, https://doi.org/10.5194/gmd-6-929-2013.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ecmwf ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, **122**, 73–119, https://doi.org/10.1002/qj.49712252905.

——, ——, C. Marsigli, A. Montani, F. Nerozzi, and T. Paccagnella, 2001: A strategy for high-resolution ensemble prediction. I: Definition of representative members and global-model experiments. *Quarterly Journal of the Royal Meteorological Society*, **127**, 2069–2094, https://doi.org/10.1002/qj.49712757612.

Müller, M., Y. Batrak, J. Kristiansen, M. A. Ø. Køltzow, G. Noer, and A. Korosov, 2017: Characteristics of a convective-scale weather forecasting system for the european arctic. *Monthly Weather Review*, **145**,

4771–4787, https://doi.org/10.1175/MWR-D-17-0194.1.

Ollinaho, P. and Coauthors, 2017: Towards process-level representation of model uncertainties: Stochastically perturbed parametrizations in the ecmwf ensemble. *Quarterly Journal of the Royal Meteorological Society*, **143**, 408–422, https://doi.org/10.1002/qj.2931.

Seo, H., and J. Yang, 2013: Dynamical response of the arctic atmospheric boundary layer process to uncertainties in sea-ice concentration. *Journal of Geophysical Research: Atmospheres*, **118**, 12, 383–312, 402, https://doi.org/10.1002/2013JD020312.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at nmc: The generation of perturbations. *Bulletin of the American Meteorological Society*, **74**, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074%3C2317:EFANTG%3E2.0.CO;2.

Wilks, D. S., 2011: *Statistical methods in the atmospheric sciences*. Elsevier Academic Press,