Royal Netherlands
Meteorological Institute
Ministry of Infrastructure
and Water Management

Probabilistic wind speed forecasting using parametric and non-parametric statistical post-processing methods

E. Ioannidis, K. Whan and M. Schmeits

KNMI Internal report IR-2018-07

Probabilistic Wind Speed Forecasting using Parametric and non-parametric Statistical Post-Processing Methods

Eleftherios Ioannidis[1, 2], *Kirien Whan*[1] *and Maurice Schmeits*[1]

1. R&D Weather and Climate modeling, The Royal Netherlands Meteorological Institute (KNMI), the Netherlands
2. Laboratory of Meteorology, Department of Physics, University of Ioannina, Greece

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Waterstaat

De Bilt 2018

Abstract

Probabilistic wind speed forecasts are generated in this study using parametric and non-parametric statistical post-processing (also called calibration) methods. The data concern 10 ensemble members of 10m wind speed which are derived from the non-hydrostatic Harmonie MetCoop ensemble prediction system (HarmonieMEPS; run by Met Norway and SMHI (Sweden)) and observations from stations covering Denmark and the surrounding areas for the period December 2016-2017. We have used HarmonieMEPS data, as HarmonieMEPS has been running operationally since autumn 2016, while HarmonieKEPS (where K stands for KNMI) has been running experimentally at KNMI only since winter 2017. The period was split into three seasons; winter, spring, and summer using a three-fold cross-validation framework for each season. More than 40 atmospheric parameters were used as potential predictors for wind speed and they can be divided into variables related to model wind speed, other meteorological parameters and geomorphology. Three main statistical methods were used to improve the probabilistic forecasts, including two parametric (Ensemble Model Output Statistics (EMOS) and Member by Member (MBM)), and one non-parametric method (Quantile Regression Forests (QRF)).

Common verification methods were used to compare the post-processed and raw forecasts: Brier Score (BS), Continuous ranked probability score (CRPS), reliability diagrams and Brier Skill Score (BSS) and Continuous ranked probability skill score (CRPSS) using the seasonal sample climatology. The scores of the calibrated and raw forecasts were compared for all stations and the skill scores were computed per station. Using the EMOS technique a number of distributions were tested, like Normal, Normal truncated (NOtr), BOX COX t etc., to the raw forecast data in order to calibrate the forecast. In the first stage only two predictors were used in order to choose the distributions which provided more skilful forecasts: mean and standard deviation of wind speed ensemble members. At the later stage a stepwise selection was used to choose more and different number of predictors using the NOtr distribution, which was the most skilful of the tested distributions. Also the QRF was used including all the potential predictors, as well as three different subsets per season with the most common predictors being tested. NOtr provided more skilful forecasts, in terms of better reliability, than the forecasts using QRF and QRF-subsets and the raw forecasts for winter and spring. In summer the results show that QRF performed a bit better than NOtr in terms of the CRPSS. Ensemble mean wind speed and land type are the two best predictors for the mean of NOtr in every season, while for the standard deviation of NOtr the ensemble standard deviation of wind speed and land type are the two most selected predictors. For the higher thresholds,

the BSS for QRF is worse than for NOtr and the raw forecasts, despite QRF scoring equally compared to the other methods in the CRPSS.
Finally, two approaches were used, to correct the ensemble members individually (member-by-member (MBM)): the so-called CRPS MIN(imum) and BEST REL(iability) methods, using two predictors per season. The corrected forecasts have been improved compared to the raw forecasts and the CRPS MIN and the BEST REL methods are about equally skilful. The results, concerning the methods, are consistent between the seasons, while the chosen predictors are different per season.

Contents

Acronyms

BMA: Bayesian model averaging
BMRC: Bureau of Meteorology Research Centre
BS: Brier Score
BSS: Brier Skill Score
CRPS: Continuous ranked probability score
EMOS: Ensemble Model output statistics
EPS: Ensemble prediction system
ECMWF: European Centre for Medium-Range Weather Forecasts
H-A: Harmonie-Arome
HMS: Hungarian Meteorological Service
KEPS: KNMI Ensemble Prediction System
LN: Log-Normal
MBM: Member by Member method
MEPS: MetCoOp Ensemble Prediction System
MOS: Model output statistics
NCEP: National Centers for Environmental Prediction
NGR: Nonhomogeneous Gaussian regression
NOtr: Normal truncated distribution
NWP: Numerical Weather Prediction
QRF: Quantile regression forest
RMSE: Root mean square error
SLAF: Scaled Lagged Average Forecasting
TN: normal truncated
UWME: University of Washington Mesoscale Ensemble
WMO: World Meteorological Organization

1. Introduction

Recently more and more public services and people demand higher accuracy weather forecasts. But the chaotic nature of the atmosphere and the physical processes within it lead to unavoidable uncertainties in weather prediction, especially on the local scale and in the long term. Also, the ongoing climate change (IPCC 2014) and extreme weather phenomena have increased these uncertainties, putting more pressure to forecasters.

This has lead the scientific community to focus more on research to minimize forecast errors. A very efficient way to do this (e.g. Hermi et al, 2014) is by correcting the output/ensemble members from numerical weather prediction (NWP) models, using statistical post-processing or calibration methods. The statistical post-processing technique corrects for systematic errors of the model and accounts for local influences which are not completely resolved in the grid box representation of the model output.

Bremnes (2004) used a local quantile regression methodology (an extension of quantile regression) and he applied it to wind speed data. These data were derived from the Hirlam model using a cross-validation methodology in which 10 predictor combinations were tested, including wind speed, wind direction and month. On the other hand, Buhari (2006) used Weibull and Rayleigh distributions to estimate wind power in Taiz, Yemeni, using wind speed observations. Also, Amaya-Martinez (2014) used Weibull, Rayleigh, gamma and log-normal distributions to estimate the wind power density using wind speed observations from six stations in Antioquia, Colombia. She found that the observed wind behavior is represented by a Weibull distribution at two of the locations, by the log-normal distribution at three locations and by the Gamma distribution at one location.

As a more complicated method, Baran (2014) used Bayesian model averaging (BMA), which is a mixture of normal truncated (to the left at zero) distributions comparing his result with an another version of the BMA method, based on the gamma distribution. Applying these methods to the 11-member ALADIN-HUNEPS ensemble of the Hungarian Meteorological Service (HMS) and to 8-members from the University of Washington Mesoscale Ensemble (UWME) he found that the BMA model based on the normal truncated distributions performs better than the BMA gamma model. Baran and Lerch (2015) using the 50-member European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble, the 11-member ALADIN-HUNEPS ensemble and the 8-member University of Washington mesoscale ensemble developed two different models: Ensemble Model output statistics (EMOS) based on the Log-Normal (LN) distribution and a normal truncated and Log-Normal (TN-LN) regime-switching mixture model. A comparison between these two models and other models like the

normal truncated based EMOS method and the truncated normal and general extreme value (GEV) mixture model from Thorarinsdottir and Gneiting (2010) and Lerch and Thorarinsdottir (2013), respectively, indicated that the TN-LN mixture model performs better.

On the other hand, Taillardat et al. (2016) used a non-parametric approach to calibrate the wind speed ensemble forecast in France, between 2001 and 2014. More specifically, they compared the quantile regression forest (QRF) methodology with some parametric distributions, like normal, normal truncated, log-normal, gamma, beta, logistic distribution. In the case of QRF they used a whole list of potential predictors, while in the case of distributions they only used potential predictors from the wind speed ensemble forecasts. Thus, they found that QRF performed better than the parametric distributions and also the necessity to use extra predictors. Finally, Hermi et al. (2014) used a modified version of the EMOS model based on a left-truncated (at zero) normal distribution in order to evaluate the evolution of the difference in skill between the raw ensemble and the post-processed forecasts.

The goal of our study was to improve the wind speed forecasts using 10 ensemble members from Harmonie MEPS for Denmark between December 2016 and August 2017. We have generated the probabilistic wind speed forecasts using five main distributions and for different thresholds: Normal, Log-Normal, Box COX t, Gamma and Weibull and some modifications of them using the “gamlss” package in R and we have compared them using different verification methods. Also, a comparison between some parametric distributions and a non-parametric approach, the Quantile regression forest (QRF) using 46 potential predictors in the stepwise selection process took place. Finally, we have tested the Member by Member method (MBM; Van Schaeybroeck and Vannitsem 2015) to calibrate the raw ensemble data, using the CRPS min and BEST REL approaches in order to correct the ensemble members individually.

The report’s outline is as follows: data and methodology are presented in Section 2, results and discussion in Section 3 and finally the conclusions in Section 4.

2. Data and methodology

2.1 Data

Harmonie-Arome (H-A) is a high-resolution (2.5 Km x 2.5 Km), non-hydrostatic Numerical Weather Prediction (NWP) model that is used for operational short-range weather forecasts in Denmark, Estonia, Finland, Iceland, Ireland, Lithuania, Norway, Spain, Sweden and the Netherlands. H-A has 65 levels in the vertical, with model top at 10hPa and lowest level at 12m and the model time step is 75s (Bengtsson et al. 2017).

In this study an ensemble of Harmonie (Harmonie-MEPS (MetCoOp-Ensemble Prediction system)) is used. Harmonie-MEPS (Andrae 2017; Frogner 2017) is a convection-permitting atmosphere ensemble model, covering Scandinavia and the Nordic Sea. Harmonie MEPS uses 10 members. It is run four times daily (at 00, 06, 12, and 18 UTC) with three-hourly cycling for data assimilation. More specifically, member 0 and 1 are run up to 66 hours, while the rest up to 48 hours. The perturbed boundary and initial conditions are based on the Scaled Lagged Average Forecasting (SLAF) method (Ebisuzaki & Kalnay, 1991, Kalnay, 2003) and the model code based on Harmonie cycle 40h1.1. The general idea of SLAF is that perturbations are taking single deterministic model (HRES) forecasts valid at the same time but with different forecast lengths and initial times:

$$IC\_m = A\_c + K\_m * (IFS\_N - IFS\_N\text{-}6)$$

$$BC\_m = IFS\_0 + K\_m * (IFS\_N - IFS\_N\text{-}6)$$

Where IC_m is the initial condition for member m, BC_m is the lateral boundary condition for member m, A_c is the control analysis, K_m a scaling factor, IFS_0 is the latest available IFS forecast, IFS_N is a forecast with length N and IFS_N-6 is a 6h shorter forecast, both valid at the same time as the analysis (see more details in Harmonie System Documentation;https://hirlam.org/trac/wiki/HarmonieSystemDocumentation/EPS/SLAF).

It is also using lateral boundary conditions from the European Centre for Medium-Range Weather Forecasts (ECMWF) model. In The Netherlands the Harmonie KEPS cycle 40h 1.1 has been running experimentally since winter 2017, while the Harmonie MEPS has been running already for more than 1 year. For this reason we chose the study area to be Denmark as it has similar geographical morphology as The

Netherlands and thus the results from this study are expected to be similar for The Netherlands.

We have used a forecast data set with 10 ensemble members, including wind speed at 10m, for the period 11th November 2016 until 09th September 2017, covering a smaller domain, Denmark and surrounding areas (51.81-58.14°N, 0.03°W-15.66°E). For the purposes of this study (as they are described in Section 2.2), we have split the whole period in three seasons (winter, spring, and summer) thus excluding November 2016 (19 days) and September 2017 (9 days). The forecast data are daily, with output saved every three hours and run for +48 hours from 00 UTC. In addition to Harmonie-MEPS wind speed, we have used 45 other potential predictors (Table 1) that are partly downloaded from the Norwegian Meteorological Institute (MET Norway Thredds Service :http://thredds.met.no/thredds/catalog.html).

We have used hourly observations for wind speed at 10 m (last ten minutes mean of each hour), using data from 97 stations in Denmark (Figure 1, See Table 1 in Annex for station's number, height, latitude and longitude) and in surrounding areas, for the mentioned period, to verify and calibrate wind speed at the closest Harmonie-MEPS grid point to each location. We examined lead-times between +0 (or +3) and +48 h (per 6 h).

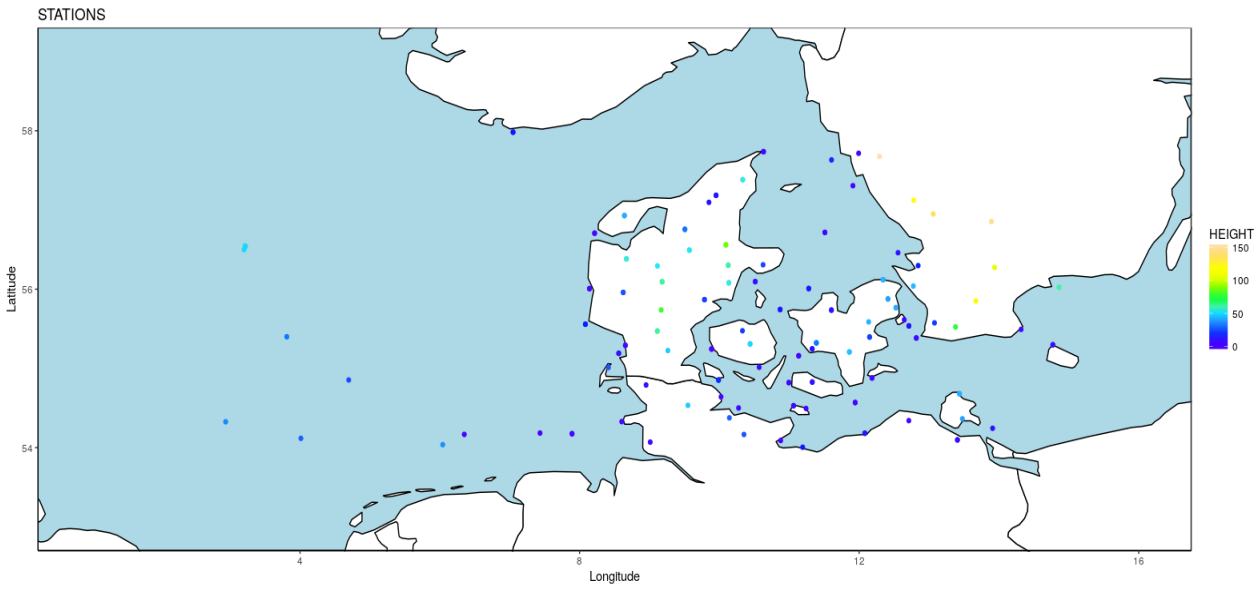


Figure 1: Study domain with stations in Denmark and surrounding (sea) areas (North: 58.14, South: 51.81, East: 15.66, West: 0.03).

Table 1: Potential Predictors

	Potential Predictors[1]
1	0m Air Temperature (°C)
2	Precipitation (mm/h)[2]
3	10m Wind Speed (m/s)
4	Roughness index (from the 250m average resembling data)
5	Land type code from the Corine Land Cover data set (~100m)[2]
6	Cosine of 10m Wind Direction
7	Convective Inhibition
8	mean Sea-level Pressure (Pa)
9	10m Wind Gust (m/s)
10	Latitude
11	Longitude
12	Elevation (m)
13	2m Relative Humidity
14	Surface Air Pressure (Pa)
15	Surface Geopotential
16	Convective available potential energy
17	Roughness Length for Momentum
18	10m Zonal Wind (m/s)
19	10m Meridian Wind (m/s)
20	Atmosphere Boundary layer Thickness
21	Atmosphere level of Neutral Buoyancy
22	Potential Vorticity at 500, 700, 850 and 925 hPa
23	Geopotential at 500, 700, 850 and 925 hPa
24	Turbulent Kinetic Energy at 500, 700, 850 and 925 hPa
25	Upward air velocity at pressure levels at 500, 700, 850 and 925 hPa (m/s)
26	Zonal wind at pressure levels at 500, 700, 850 and 925 hPa (m/s)
27	Meridional wind at pressure levels at 500, 700, 850 and 925 hPa (m/s)

[1] Every meteorological variable consists of 2 predictors, namely the mean and standard deviation of the values of 10 ensemble members except Atmosphere Boundary layer Thickness which consists of the mean and standard deviation of only2 members. The Land Type, Roughness index, latitude, longitude and elevation are not based on the ensemble

[2] See in Annex

2.2 Methodology

2.2.1 Verification

Forecasts for wind speed were verified both deterministically and probabilistically. We evaluated the ensemble mean and the probability of wind speed greater than 2, 5, 8, 11, 14, 17, 20, 23 and 25 m/s compared to observations. Some verification methods were used (Table 2), like the root mean square error (RMSE), the continuous ranked probability score (CRPS), scatter diagrams, reliability diagrams, rank histograms, Brier Score (BS) and Brier Skill Score (BSS) (Wilks, 2011; Hamill, 2000). The BSS was calculated using the seasonal station sample climatology as a reference. It has been noted that some of the largest operational centers (like ECMWF, NCEP, Met Office, and BMRC) and the WMO are using several verification methods, including BSS, reliability diagrams, RMSE and rank histograms among others (information has been taken from BMRC).

2.2.2 Calibration

Nonhomogeneous Gaussian regression (NGR) is one of the two most frequently used EMOS methods, besides Bayesian model averaging (BMA; Raftery et al, 2005).NGR was proposed by Gneiting et al. (2005) and independently by Jewson, Brix and Ziehmann (2004).

We used the version of EMOS as described by Gneiting et al.(2005),where the parameters of the forecast distribution depend on a set of selected predictors. It is a regression-based method, in that the conditional mean and variance of the predictive distribution are defined as optimized linear combinations of the ensemble mean and variance, respectively(Wilks, 2011).

We verified the wind speed forecasts of five different distributions for the mentioned thresholds, based on Brier Score and Brier Skill Score values: Log Normal, Gamma, Normal, Weibull and Box-Cox T (BCT) distributions (Table 3). We used the “gamlss” package (see for more details Stasinopoulos and Rigby, 2007; Stasinopoulos et al. 2008 and references therein). In the case of “Log Normal Distribution”, we used the LNO (Box-Cox) and LOGNO2 distributions from the “gamlss” package in R. According to Stasinopoulos et al. (2008), the LOGNO2 distribution uses μ as the median, so $\mu = (0, +\infty)$, while LNO is more general and can fit a Box-Cox transformation to data. In the case of the Weibull distribution, three different functions were used. Their differences are due to different parameterizations (Stasinopoulos et al. 2008). The Normal truncated (NOtr) and Log Normal left truncated distributions (Nadarajah S. and Kotz S., 2006) were used also in order to exclude the negative values in the case of the Normal distribution (10m wind speed has

zero or positive values) and to study whether better results can be obtained in the case of the Log Normal distribution (LOGNO2). Based on the highest values of BSS, the best two distributions which were fitted to our data were chosen. In order to find the best fit to our data, the forward and backward stepwise selection was used. The goal was to minimize the Akaike Information Criterion (AIC, Akaike, 1974; Sakamoto et al., 1986) using 45 potential predictors (Table 1), to predict the parameters of the distribution (e.g. μ, σ and ν for the BOX COX t distribution). We run several tests using different number of predictors trying to minimize BS and to avoid over fitting. Also, we run different tests, using at the beginning only the mean value of every potential predictor as predictor for μ and the standard deviation of every potential predictor for σ and at a later stage we used mean and standard deviation as potential predictors for μ and σ. Also, we chose predictors which physically and meteorologically can improve the 10m wind speed forecasts per season most.

A machine learning methodology was also used. More specifically, we compared our parametric EMOS results with those of the Quantile regression forests (QRF, Meinshausen, 2006; Taillardat et al., 2016). QRF is a generalization of random forests (Breiman 2001) and it is a non-parametric way to estimate conditional quantiles. The biggest advantage of this method is that it does not assume any distribution and it always builds a distribution according to the data. In this study we used the default values for the number of trees (500 trees) and the terminal node size (5).

Data from all stations were pooled in both the training and testing data sets and we used a three-fold cross-validation framework (Wilks, 2011) to verify forecasts on an independent data set. We split the period into three seasons (winter, spring and summer) and for every season we used two months as a training data set and the remaining month as a test data set. More specifically, we separated our data in three seasons (winter: December, January, and February. Spring: March, April, May. Summer: June, July, August), between December 2016 and August 2017. Because of lack of data autumn has been excluded from this study. We then combined the three months of independent forecasts and verify them together, using the seasonal station climatology as a reference.

Table 2: Verification methods (Brown, 2015). f_f: ensemble forecast, f_o: mean observations. p_i: probabilistic forecast, o_i: binary observations, F: Cumulative density function of the ensemble forecast, F_o: step-function observation

Measure	Attribute evaluated	Comments	Definition
Ensemble forecasts			
RMSE	Skill	Perfect Score: 0	$\sqrt{\frac{1}{N} * \sum_{i=1}^{N} (f_f - f_o)_i^2}$
Probability forecasts			
Brier Score (BS)	Accuracy	Measures the mean squared probability error. Perfect Score: 0	$\frac{1}{N} * \sum_{i=1}^{N} (p_i - o_i)^2$
Reliability diagram	Calibration	Measures how well the predicted probabilities of an event correspond to their observed frequencies (reliability)	Plot observed frequency against binned forecast probability
Brier Skill Score (BSS)	Skill	Measures the relative skill of the forecast based on the climatology using the BS as a metric. Perfect Score: 1	$1 - \frac{BS}{BS_{clim}}$
Ensemble distribution			
Rank Histogram	Calibration	Measures how well the ensemble spread of the forecast represents the true variability (uncertainty) of the observations	Plot rank of observations in ensemble members
CRPS	Accuracy	Measures how well the forecast distribution matches the observation. Perfect Score 0.	$\frac{1}{N} * \sum_{i=1}^{N} \int_{-\infty}^{\infty} (F(x) - F_o(x))^2 dx$

Finally, we used MBM methods to correct each member individually, by a linear mapping. Van Schaeybroeck and Vannitsem (2015) have introduced this methodology and they and Schefzik (2017) have applied it for 2m temperature. These methods include the classical Linear Model Output Statistics (MOS) approach (Glahn and Lowry, 1972), the Error-in-Variables Model Output Statistics (EVMOS) approach, the bias correction and finally two ensemble-spread calibration techniques, BEST REL and CRPS MIN (Van Schaeybroeck and Vannitsem 2015). Van Schaeybroeck and Vannitsem (2015) found that the BEST REL and CRPS MIN methods have more skill

than the other three methods. In this study we compared the corrected ensemble members for the wind speed per season with the uncorrected raw data, using the BEST REL and CRPS MIN method and two predictors: wind gust in the case of winter and spring and 0m air temperature for summer.

2.2.3 QRF sensitivity tests

In the initial test of QRF all potential predictors were included. In order to test its performance using fewer predictors, we created and tested three different QRF-subsets per season. The predictors which appeared more often and higher in the list per month and lead-time have been chosen, creating three different combinations. These three QRF-subsets per season are described below:

1. Winter

1.1. QRF Subset 1

- Mean value 10m wind speed
- Mean values wind gust
- Mean value 10m zonal wind speed
- Land type
- Mean value 0m air Temperature
- Standard deviation 0m air Temperature
- Mean Surface geopotential
- Mean value roughness of momentum
- Mean value wind speed observations

1.2. QRF Subset 2

- Mean value 10m wind speed
- Mean value wind gust
- Standard deviation wind gust
- Standard deviation 10m wind speed
- Mean value convective inhibition
- Mean value 0m air Temperature
- Rough near (index)
- Mean value Mean Sea level pressure
- Mean value wind speed observations

1.3. QRF Subset 3

- Mean value 10m wind speed
- Latitude
- Mean value 10m meridian wind speed
- Mean value surface air pressure
- Mean value 0m air Temperature
- Mean value x wind speed at 500 hPa

- Mean value 2m relative humidity
- Mean values precipitation
- Mean value wind speed observations

2. Spring

2.1 QRF Subset 1

- Mean value wind speed
- Mean value wind gust
- Mean value atmosphere boundary layer thickness
- Mean value convective inhibition
- Standard deviation convective inhibition
- Mean value 0m air temperature
- Longitude
- Land Type
- Mean value wind speed observations

2.2 QRF Subset 2

- Mean value wind speed
- Mean value 10m zonal speed
- Mean value 10m meridian wind speed
- Mean value surface geopotential
- Standard deviation 0m air temperature
- Mean value x wind speed at 925 hPa
- Standard deviation wind speed
- Standard deviation wind gust
- Mean value wind speed observations

2.3 QRF Subset 3

- Mean value wind speed
- Latitude
- Mean value 2m relative humidity
- Mean value turbulent kinetic energy at 925 hPa
- Standard deviation mean sea level pressure
- Mean value momentum of roughness
- Standard deviation 0m air temperature
- Mean value 0m air temperature
- Mean value wind speed observations

3 Summer

3.1 QRF Subset 1

- Mean value wind speed
- Mean value wind gust
- Mean value momentum of roughness
- Mean value wind direction

- Rough near Index
- Longitude
- Mean value 0m air temperature
- Mean value convective available potential energy
- Mean value wind speed observations

3.2 QRF Subset 2

- Mean value wind speed
- Mean value wind direction
- Mean value 2m relative humidity
- Mean value 0m air temperature
- Mean value convective inhibition
- Mean value 10m meridian wind speed
- Mean value 10m zonal wind speed
- Latitude
- Mean value wind speed observations

3.3 QRF Subset 3

- Mean value wind speed
- Mean value wind gust
- Mean value convective inhibition
- Mean value atmosphere level of neutral buoyancy
- Mean value atmosphere boundary layer thickness
- Mean value convective available potential energy
- Mean value precipitation
- Mean value surface air pressure
- Mean value wind speed observations

Table 3: Distributions and characteristics. WS: wind speed; all: mean value and standard deviation of all meteorological variables as potential predictors; sd: standard deviation

Distributions	Gamlss Name	Probability density function	Parameters	Predictors
Log-Normal	LNO	$f(y \vee mu, sigma, nu) = \left(\frac{1}{\left(\sqrt{2*pi}\right)*sigma}\right) * (y^{nu-1}) * e^{-\left(\frac{(z-mu)}{2*sigma}\right)^2}$	mu[3], sigma[4], nu (default value equal to zero)	Mu = mean WS Sigma = sd WS
Log-Normal	LOGNO2	$f(y \vee mu, sigma) = \left(\frac{1}{\left(y*\sqrt{2*pi}*sigma\right)} * e^{\left(-0.5*\left(\frac{(logy-mu)^2}{sigma}\right)\right)}\right)$	mu, sigma. mu=(0,+Inf)	1st) mu = mean WS, sigma = mean WS 2nd) mu = mean WS, Sigma = sd WS
Gamma	GA	$f(y \vee mu, sigma) = \frac{\left(y^{\left(\left(\frac{1}{sigma}^2\right)-1\right)} * e^{\left[\frac{-y}{\left((sigma^2)*mu\right)}\right]}\right)}{\left((sigma^2*mu)^{\frac{1}{sigma^2}} * Gamma\left(\frac{1}{sigma^2}\right)\right)}$	mu, sigma	Mu = mean all Sigma = standard deviation all
Normal	NO	$f(y \vee mu, sigma) = \left(\frac{1}{\left(\sqrt{2*pi}*sigma\right)} * e^{\left(-0.5*\left(\frac{(y-mu)^2}{sigma}\right)\right)}\right)$	mu, sigma	Mu = mean WS, Sigma = standard deviation WS
Weibull	WEI1	$f(y \vee mu, sigma) = \left(\frac{sigma*y^{(sigma-1)}}{mu^{sigma}}\right) e^{-\left(\left(\frac{y}{mu}\right)^{sigma}\right)}$	mu, sigma	Mu = mean WS, Sigma = standard deviation WS
Weibull	WEI2	$f(y \vee mu, sigma) = sigma*mu*y^{(sigma-1)} * e^{-mu*y^{sigma}}$	mu, sigma	Mu = mean WS, Sigma = standard deviation WS

[3] We used the mean value of every meteorological variable as potential predictor for mu.

[4] We used the standard deviation of every meteorological variable as potential predictor for sigma.

Weibull	WEI3	$f(y \vee mu, sigma) = \left(\frac{sigma}{beta}\right) * \left(\frac{y}{beta}\right)^{(sigma-1)e^{-\left(\frac{y}{beta}\right)^{sigma}}}$	mu, sigma	Mu = mean WS, Sigma = standard deviation WS
BOX-Cox-t	BCT	$f(y \vee mu, sigma, nu, tau) = \frac{\left(\frac{1}{y*sigma}\right) * \left(\Gamma\left(\frac{(tau+1)}{2}\right)\right)}{\left(Gamma\left(\frac{1}{2}\right) * Gamma\left(\frac{tau}{2}\right) * tau^{0.5}\right)} * \left(1 + z^{\frac{2}{tau}}\right)^{\frac{-(tau+1)}{2}}$	mu, sigma, nu, tau[5]	Mu = mean WS, Sigma = standard deviation WS, Nu = mu + sigma, Tau = 1
Normal Truncated	NOtr	Normal distribution truncated on the left	mu, sigma	1st) Mu = mean WS, Sigma = standard deviation WS 2nd) Mu = mean/sd all Sigma = sd/mean all
Log-Normal Truncated	LOGNOtr	Log-Normal distribution truncated on the left	mu, sigma	Mu = mean WS, Sigma = standard deviation WS

[5] We defined tau equal to 1 (Based on Domenech et al, 2017)

3. Results

Results are presented in the following three sub-sections. First, insub-section3.1 the RMSE of the raw wind speed of the 10 ensemble members is computed, as well as the correlation coefficient between the observations and the most important atmospheric parameters in order to choose potential predictors. Also, the results for the probabilistic wind speed forecasts fitting 10 different distributions are analyzed for the examined period per month and then using the cross-validation method per season for the most important distributions based on the lowest Brier score. Second, in sub-section3.2 the BSS results from the two best distributions are presented using stepwise selection for a different number of predictors, and compared to quantile regression forests. Also, attribute diagrams and CRPSS are shown for the distributions which are more skillful (smallest BS) compared to the raw data. Finally, in sub-section 3.3 we compared the corrected wind speed ensemble members using the MBM method (based on the Best Rel and CRPS Min approaches) with the uncorrected raw ensemble members and subsequently the best method with the distributions which fitted better to our data.

3.1 The role of one predictor

3.1.1 Verification of wind speed ensemble members

In this sub-section the comparison between raw wind speed ensemble members, observations and some potential predictors are presented and discussed. Over all forecast times, the RMSE between the ensemble members and the observations increases with lead-time, while the correlation coefficient decreases, as expected. The ensemble members are not entirely equally probable as they depend on the initial conditions and SLAF generation. In table 4 the RMSE and the correlation between the 10 wind speed ensemble members and observations are presented for lead-time 0h, 12h, 24h and 48h. For lead-time 0h and control member (WS10.0) the correlation coefficient is equal to 0.90 and the RMSE equal to 1.5 m/s. For the other ensemble members the correlation coefficient is lower and the RMSE is higher. For lead-time 48h, the correlation coefficient has decreased (R ~ 0.80) and the RMSE (~2 m/s) has increased compared to the results for lead-time 0h (See also the scatter plots in Figure 1a, b in Annex for lead time 0h and 24h).

Despite the fact that the differences between R and RMSE for different lead-times are relatively small, these results show that with increasing lead-time the uncertainty in the wind speed forecast is higher and the correlation coefficient and RMSE decreases and increases, respectively, as expected.

Table 4: RMSE (m/s) and correlation coefficient (R) between the 10m wind speed ensemble members and observations, for 4 different lead-times based on scatter plots. WS10.0 is the control member. WS: Wind Speed.

	Lead-time 0h		Lead-time 12h		Lead-time 24h		Lead-time 48h	
	R	RMSE	R	RMSE	R	RMSE	R	RMSE
Observations-WS10.0	0.9	1.5	0.86	1.6	0.88	1.6	0.84	1.9
Observations-WS10.1	0.88	1.6	0.83	1.8	0.85	1.8	0.82	2
Observations-WS10.2	0.88	1.6	0.83	1.8	0.84	1.8	0.8	2.1
Observations-WS10.3	0.88	1.6	0.83	1.8	0.85	1.8	0.81	2
Observations-WS10.4	0.88	1.7	0.83	1.8	0.84	1.8	0.78	2.2
Observations-WS10.5	0.88	1.6	0.84	1.7	0.85	1.8	0.81	2
Observations-WS10.6	0.88	1.6	0.83	1.8	0.85	1.8	0.81	2
Observations-WS10.7	0.89	1.6	0.84	1.7	0.85	1.8	0.81	2.1
Observations-WS10.8	0.89	1.6	0.83	1.8	0.85	1.8	0.81	2
Observations-WS10.9	0.89	1.6	0.84	1.7	0.86	1.8	0.82	2

In order to be able to choose potential predictors for the wind speed's prediction (Section 3.2) we compared some atmospheric parameters with the wind speed observations. Using scatter plots we estimated the correlation coefficient. As we found in Table 4, there are small differences in R for different ensemble members and lead-time. For this reason the results from the comparison between the observed wind speed and some potential predictors are presented (Table 5) and concern the control (WS10.0) and last member (WS10.9), and the ensemble mean for lead-time 0h, 24h and 48h.

Based on R (Table 5), the wind gust has the strongest relationship with wind speed. There are small differences for different lead-times and different ensemble members. On the other hand, R for the other atmospheric parameters - potential predictors, shows low correlation with wind speed.

Table 5: Correlation between the 10m wind speed observations and some potential predictors for the control member, the last member (WS10.9) and the mean value over the ten members and for lead-time 0h, 24h and 48h. CIN: Convective inhibition. FF: Observations. Cos(WD): cosine of Wind Direction. WG: Wind gust.

Comparison/ Causes	Lead-time 0h	Lead-time 24h	Lead-time 48h
	R	R	R
Latitude-FF	-0.083	-0.082	-0.081
Rough Near-FF	-0.081	-0.082	-0.081
Elevation - FF	-0.259	-0.26	-0.26
WD10.0 - FF	-0.01	0.009	0.002
WD10.9 - FF	0.007	0.002	-0.01
Mean WD - FF	-0.007	0.006	0.002
WG10.0 - FF	0.85	0.82	0.79
WG10.9 - FF	0.848	0.81	0.76
Mean WG – FF	0.86	0.84	0.81
0m Temperature 0.0 – FF	0.074	0.07	0.065
0m Temperature 0.9 – FF	0.073	0.069	0.065
Mean.0m.Temperature - FF	0.07	0.067	0.067
2mHumidity2.0 - FF	-0.24	-0.24	-0.237
2m Humidity2.9 – FF	-0.2	-0.22	-0.22
Mean 2m.Humidity – FF	-0.23	-0.26	-0.25
CIN.0 – FF	0.21	0.206	0.216
CIN.9 – FF	0.13	0.21	0.22
Mean.CIN - FF	0.18	0.26	0.27

Despite the fact that the correlation between wind speed and some atmospheric parameters (potential predictors for wind speed) was not high, we included them in the stepwise selection procedure for NOtr and QRF distributions, as is described in the following sub-section (3.2).

3.1.2 Distributions' comparison

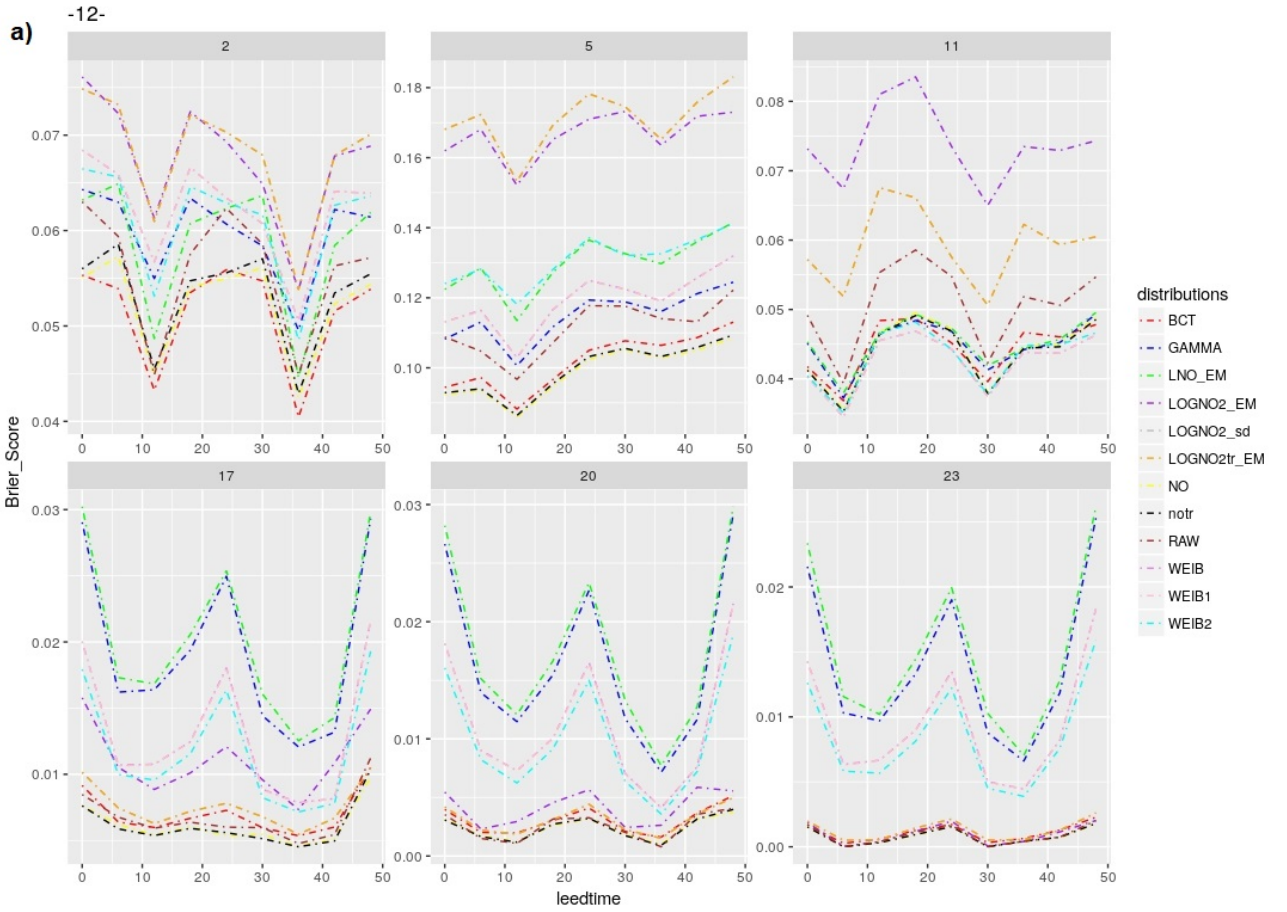
In this sub-section BS results for different distributions are presented. We fitted and compared different distributions to our data. According to literature (Bremnes, 2004; Buhari, 2006; Baran and Lerch, 2015) different distributions have been used for wind speed data, verifying the wind speed for oceanic (Baran and Lerch, 2015) and continental regions (Buhari, 2006), like Weibull, Log-Normal, Normal or a regression technique like local quantile regression. In this study we fit several distributions to our data, using the gamlss package in R. More specifically, we used: BCT, GAMMA, LNO, LOGNO2, LOGNO, NO, WEI, WEI2, WEI3 (see section 2.2 for more details about the name and properties). Also, we tested the NO truncated (at zero) distribution, as well as LOGNO truncated.

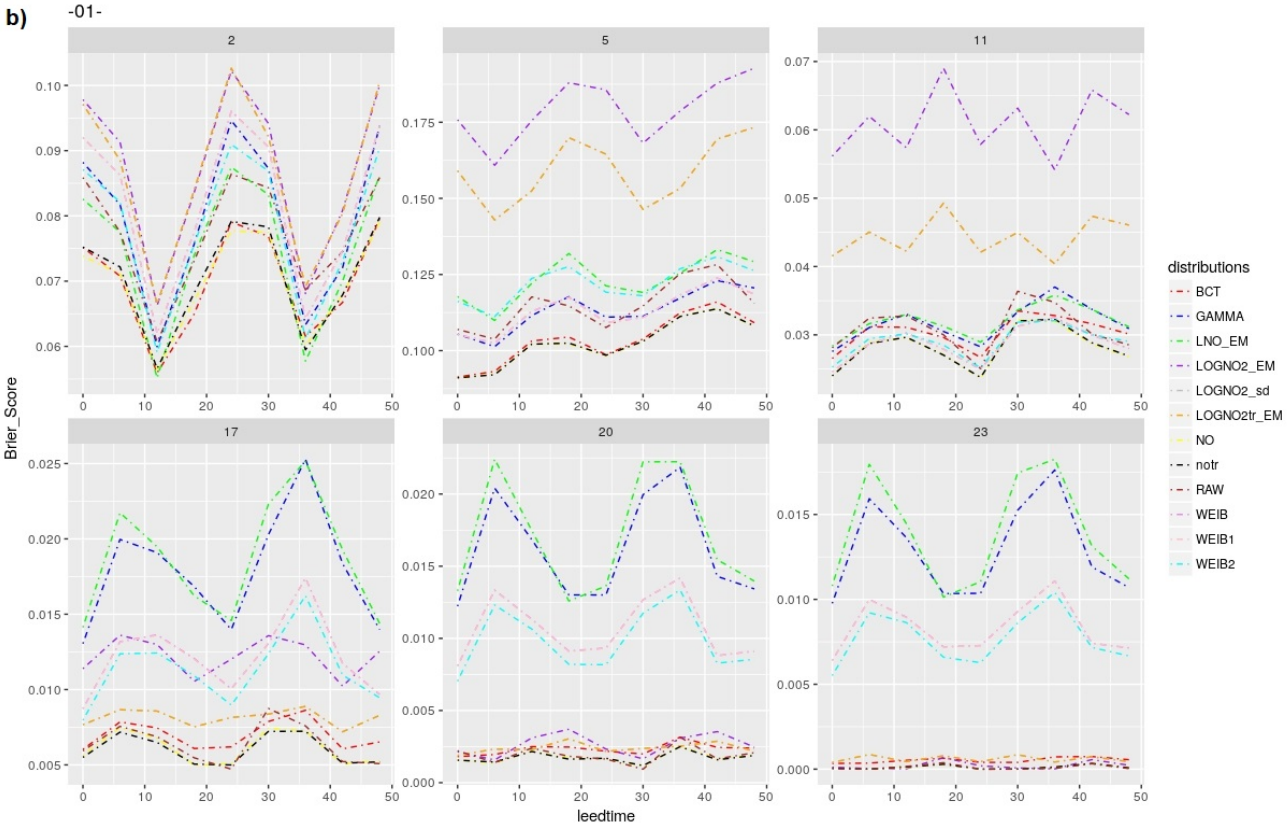
Initially, we fit the distributions per month in order to give a first insight into which distributions fit better to our data (based on the lowest BS). At the next stage, we fit them using the cross-validation method per season. We applied the cross-validation methodology to our data per season, using two of the months for training and using the other month for testing. At the end, we could verify per whole season by taking the independent months together.

For the initial results, the BS comparison (Figure 2a, b, c) is presented only for the winter, because the results for the rest of the months are similar, in terms of which distributions fit better to our data.

In more detail, we used the following distributions: BCT (predicting only mu and sigma at the beginning), GAMMA, LNO, NO, WEI, WEI1 and WEI2, as well as the truncated NO and truncated LOGNO. We used the ensemble mean wind speed as predictor for mu and the standard deviation of the ten ensemble members for wind speed as predictor for sigma. In the case of LOGNO2 distribution we compared the BS using the ensemble mean wind speed as predictor for mu and sigma on the one hand and the ensemble mean wind speed as predictor for mu and the standard deviation as predictor for sigma on the other hand (see Table 3).

In Fig. 2 the BS comparison between the mentioned distributions is presented for lead-times from 0h to 48h with 6h step and for the mentioned thresholds for winter. As has been mentioned in the methodology section (2.2), BS equal to zero indicates the perfect score, while equal to 1 indicates the worst score. Firstly, there is clearly improvement in BS when some distributions are fitted compared to BS based on the raw forecasts. This improvement is larger for low thresholds for the majority of the lead-times, while for middle thresholds (11 and 17 ms^{-1}) the improvement is smaller. For the high thresholds (extreme wind speed) there are not many cases available, but still it seems that some distributions improve upon the raw forecasts even though the differences are really small.


a)
2
5
11
17
20
23
Brier_Score
leedtime
distributions
BCT
GAMMA
LNO_EM
LOGNO2_EM
LOGNO2_sd
LOGNO2tr_EM
NO
notr
RAW
WEIB
WEIB1
WEIB2


b)
2
5
11
17
20
23
Brier_Score
leedtime
distributions
BCT
GAMMA
LNO_EM
LOGNO2_EM
LOGNO2_sd
LOGNO2tr_EM
NO
notr
RAW
WEIB
WEIB1
WEIB2

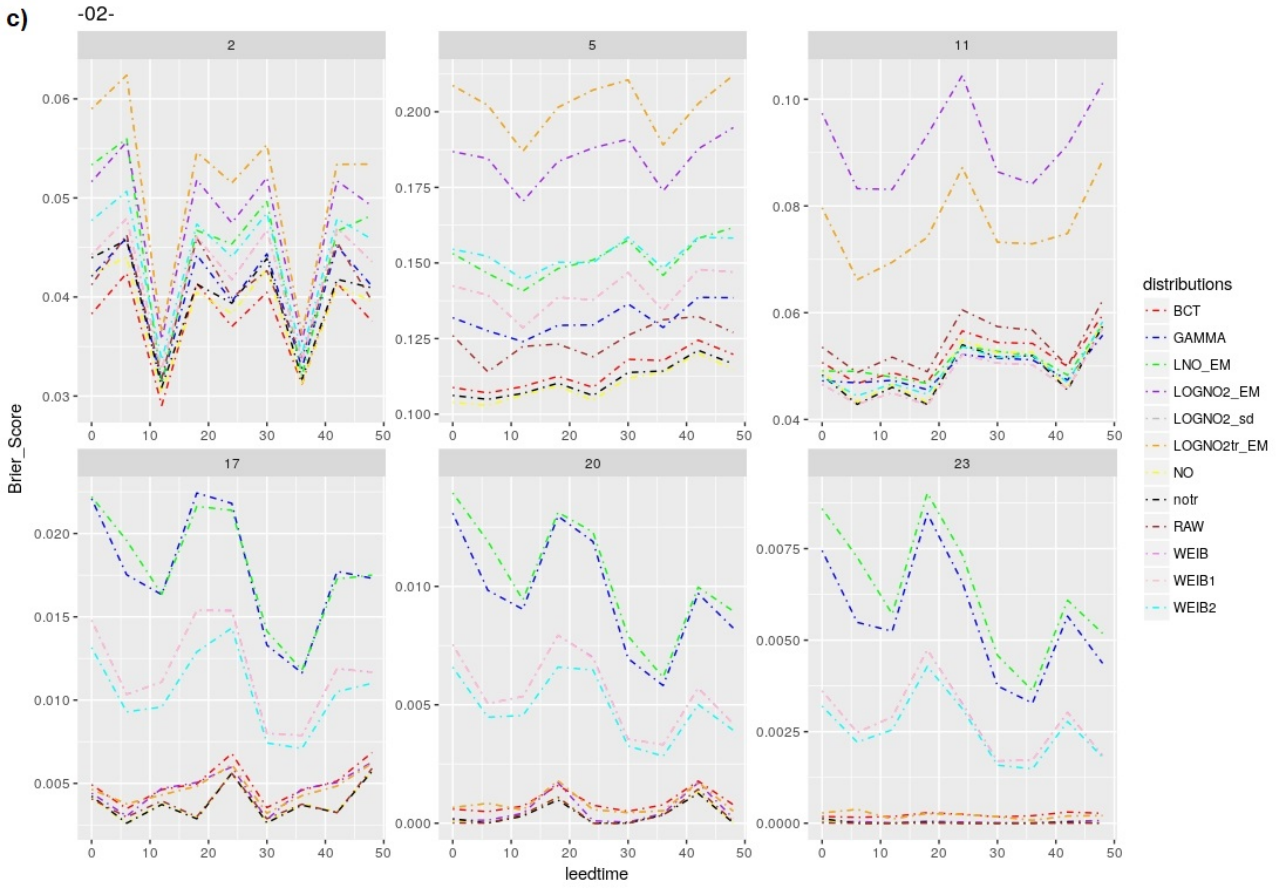


Figure 2: BS comparison for different distributions and for the mentioned thresholds and lead-time for (a) December, (b) January and (c) February.

For low and middle thresholds the BCT, NO and NOtr gave better results than the raw forecasts, while the three different WEI distributions fit also well for low thresholds. In general the rest of the distributions did not reduce the BS compared to the raw ensemble so we did not use them in the cross-validation study in the next sub-section. Also, we excluded the NO distribution from further comparison and implementation, as the wind speed cannot have negative values. Finally, despite the fact that the LNO distribution does not appear to perform well in our study, we included it for further investigation, as it fits generally well to wind speed data according to Amaya-Martinez et. al(2014)and Baran and Lerch(2015).

Next, we compared some of the distributions per season using the cross-validation method.

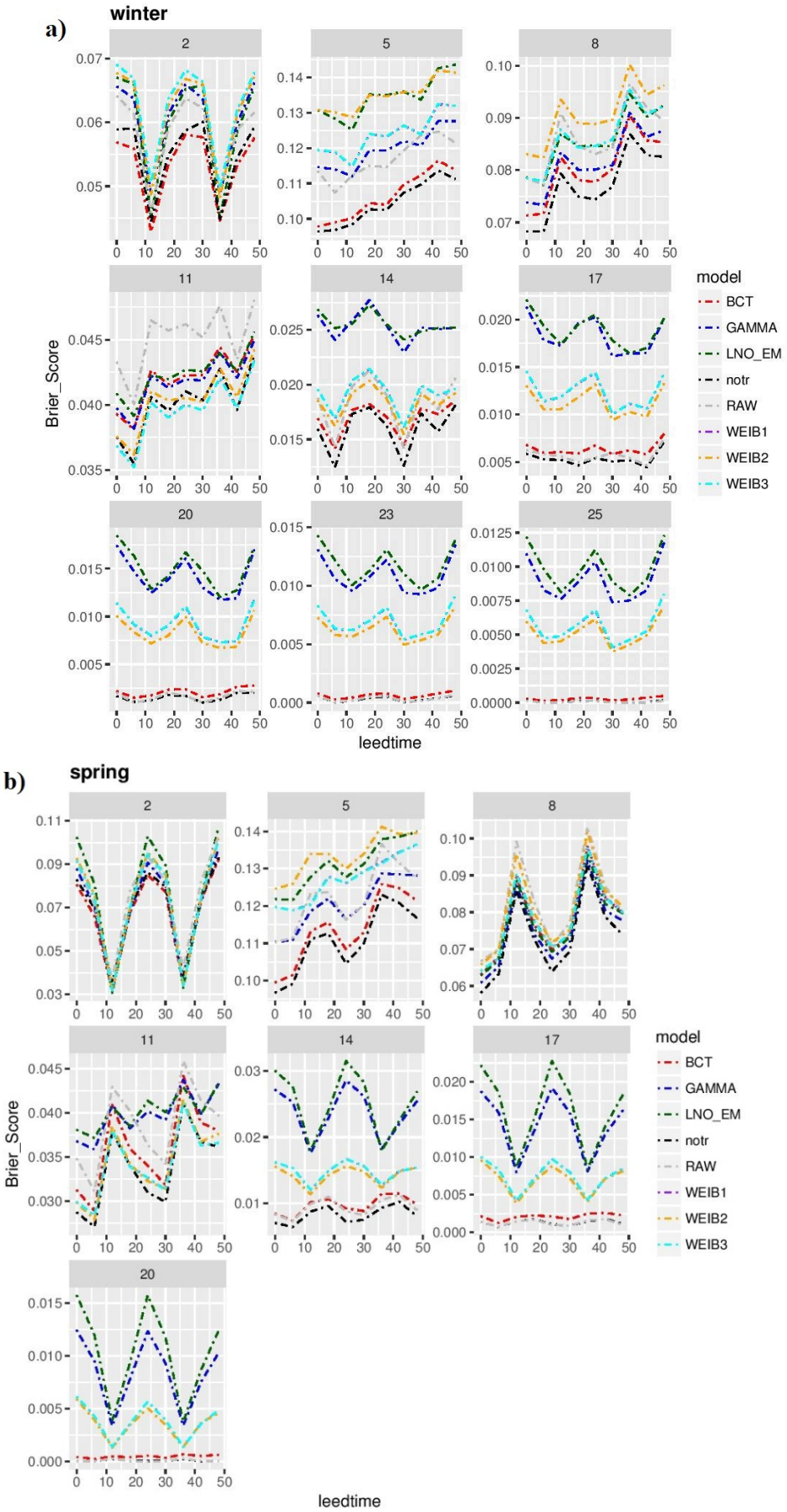
For winter (Fig. 3a) there are mixed patterns for the distributions based on the threshold and lead-time. More specifically, the BS has improved for thresholds of 8 and 11 ms^{-1} using almost every distribution compared to the raw forecasts. For lower thresholds, for example for the threshold equals to 2 ms^{-1} only the NOtr and BCT distributions perform better than the raw forecasts. For the threshold of 11 ms^{-1}WEIB2 and WEIB3 perform equal to or better than NOtr. But it is clear that BCT and especially NOtr perform in

general much better than the other distributions and the raw forecasts, such as for thresholds 2, 5, 8, and 14 ms^{-1}. For the extreme thresholds (23 and 25 ms^{-1}), while every distribution performs worse than the raw forecasts, including the BCT distribution, there are no differences between the NOtr distribution and the raw forecasts. For these thresholds the BS is almost equal to zero.

In spring (Fig. 3b) it is clear that the NOtr distribution fits better to our data, because the forecast is more skillful. Also, in this case there are some exceptions for some thresholds (such as 11 ms^{-1}) and high lead-times where WEIB3 performs slightly better, but this cannot lead to the conclusion that the WEIB3 distribution performs better than NOtr. On the contrary, NOtr performs better for every threshold and almost every lead time.

Similar patterns can also be seen in summer (Fig. 3c). The NOtr distribution performs better than the rest of the distributions for low, middle and high thresholds and lead-times. Also, there are some exceptions in which other distributions perform better for specific thresholds and lead-times, such as for threshold 2 ms^{-1}, the BCT performs better for lead-time 0 and 24h but the differences are really small.

For every season and threshold there is a general pattern: apart from a diurnal cycle the BS is increasing with lead-time. This means that the forecast is more skillful for low and middle than longer lead-times. That is understandable: the chaotic behavior of the atmosphere is the main reason that there are uncertainties in the weather forecast, especially for long lead-times.


a)
winter
2
5
8
11
14
17
20
23
25
Brier_Score
leedtime
model
BCT
GAMMA
LNO_EM
notr
RAW
WEIB1
WEIB2
WEIB3
b)
spring
2
5
8
11
14
17
20
Brier_Score
leedtime
model
BCT
GAMMA
LNO_EM
notr
RAW
WEIB1
WEIB2
WEIB3

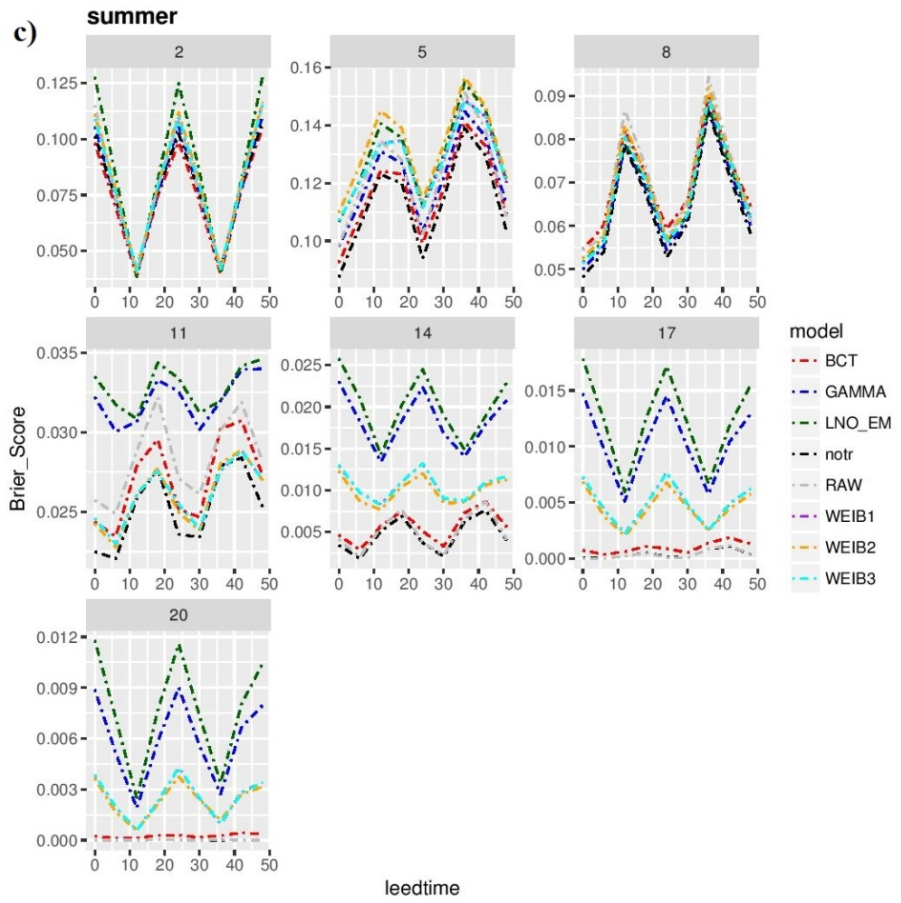


Figure 3: Brier score comparison for different distributions for a) winter, b) spring and c) summer based on cross-validation and for different thresholds and lead-times (0 to 48h, per 6h).

3.2 Stepwise selection

From the analysis so far, it is clear which distributions can be used best for probabilistic forecasts of wind speed, using the ensemble mean wind speed as predictor for mu and the ensemble standard deviation of wind speed as predictor for sigma. In the next sub-sections we describe some sensitivity tests using the stepwise selection method and different maximum number of potential predictors (see Table 1) in the case of NOtr. We compare with the BCT distribution, as it is used semi-operationally and by Domenech et al. (2017) and we compared these two parametric methods with a machine learning methodology, QRF, using different subsets of potential predictors (sub-section 2.2.3). Finally, we used three more verification methods (BSS, Reliability diagrams and CRPSS) to evaluate which statistical post-processing methods were most skillful.

3.2.1 Sensitivity tests and comparison between NOtr, BOX COX t and QRF methods

In this sub-section the comparison between the NOtr, BCT distributions (with the stepwise selection (minimization of AIC value)) and QRF (using different numbers of potential predictors) is presented. First, we compared the skill of the different forecasts using the BS.

We calculated the mean value and standard deviation over the ensemble members for the 45 potential predictors (see Table 1). In the case of the NOtr distribution, we used the mean value and the standard deviation of every parameter as potential predictors for mu and sigma. Also, we tested different maximum number of predictors looking for the lowest BS. Also, as has been mentioned in literature (Taillardat et al., 2016) the use of many predictors can lead to overfitting. Thus, we tested the NOtr using 2, 5, 10 and 15 predictors out of the 45 potential predictors for mu and sigma. In the case of the BOX COX t distribution, we used the form which according to Domenech et al. (2017) provided better results for wind speed across Europe. Thus, we used the mean value of wind speed as predictor for mu, the standard deviation of wind speed as predictor for sigma, the mean and standard deviation of wind speed as predictors for nu, while we defined the tau equal to 1. Finally, in the case of QRF we included the mean value and the standard deviation of every predictor in order to estimate the probabilities. In this case we did not test the QRF using different number of predictors, thus the total number of predictors depended on the lead-time and the season.

Firstly, the BS is presented for every distribution and sensitivity test per season. In Fig. 4 the BS is presented for the three seasons. The BS was estimated for the NOtr and BCT distributions, and QRF, for different numbers of predictors as explained above and in the methodology section. In the case of winter (Fig. 4a) there are not big differences between the distributions for the low thresholds. On the other hand, for middle thresholds, there are differences and it appears that the NOtr using 2, 5, 10 and 15 predictors perform better than the QRF and the three QRF subsets. For a few lead-times the QRF and the QRF subset 1 perform better than the NOtr, while the BCT distribution performs better than the QRF subset 2 and subset 3. For the extreme values of wind speed it seems that the NOtr distribution versions perform better, minimizing the BS, while the BS differences between the raw and NOtr probabilities remain quite high. Based on the observations there are 242 cases of wind speed greater than 20 ms^{-1}, 37 cases greater than 23 ms^{-1} and only 5 cases greater than 25 ms^{-1}. Based on this we can assume that the results for 20 and 23 ms^{-1} are accurate and statistically important. Thus, based on the NOtr distribution the two most selected predictors for both mu and sigma are: ensemble mean wind speed and land type. As land type is considered as a factor in R, for every lead-time different land types are chosen depending on where the station is located (see Annex).

Finally, although the differences between NOtr choosing a different number of maximum predictors (2, 5, 10 or 15) are small, we will exclude the results for the last two cases from the discussion, as there is overfitting between the chosen predictors/atmospheric parameters. Note that in the case of NOtr using 5 predictors, there is not a standard chosen third, fourth and fifth predictor, as it is depending on the lead-time. The most often selected predictors are presented in Table 6 per season.

For short lead-times, the latitude was chosen as the third predictor probably because of the station configuration with more sea stations (with generally higher wind speed) in the south than in the north, which is confirmed by the negative correlation coefficient in Table 5.

In spring (Fig. 4b) we see similar behavior as in winter for low thresholds, in which there is not a clear BS difference. For the middle thresholds (8, 11 and 14 ms^{-1}) a difference appears. More specifically, it seems that the 3 QRF subsets and the BCT perform worse than the NOtr distributions. There is not, also, a clear pattern between the 3 QRF subsets and QRF. NOtr_5 has the lowest BS. For the extremes there is not a clear picture as it seems that the differences between the distributions are really small, excluding the BCT. In spring there are 100 cases of wind speed larger than 17 ms^{-1} based on the observations. The ensemble mean 10m wind speed and land type are also the two first chosen predictors for mu in spring, using the NOtr distribution for every lead-time, while for sigma the ensemble

standard deviation of 10m wind speed and land type are the two most often selected predictors..

Finally, it is remarkable that in summer (Fig. 4c) there are big fluctuations in BS for different lead-times and for every threshold compared to the other seasons. Here the thresholds of 14 and 17 ms^{-1} are quite extreme, as there are only a few cases for a threshold equal to 20 ms^{-1}. In more detail, while there is not a clear difference for threshold equal to 2 ms^{-1}, it seems that for low thresholds (8 and 11 ms^{-1}) the NOtr distributions and the QRF perform better than the QRF subsets and the BCT distribution. Although the differences in BS between the distributions are small, the NOtr minimizes it more than the others. For the extremes there is not a clear picture of the BS differences between the distributions except that the BCT performs worse than the raw forecasts. Also, for a threshold equal to 20 ms^{-1} all the distributions predicted probability of wind speed exceeding 20 ms^{-1}equal to zero, except the BCT, which indicates that the use of more and different predictors leads to better results.

Also, for summer the two most often selected predictors for mu are ensemble mean wind speed and land type and for sigma the ensemble standard deviation of wind speed and land type for every lead-time. For the other predictors there is not a standard atmospheric parameter which is chosen. The standard deviation and mean of 0m air temperature were the most common predictors for almost every lead-time, which indicates the strong relation between thermal activity and wind speed in summer (as the earth's surface is heated by the surface solar radiation, the temperature increases, leading to thermal activity and hence increased wind speed).

Table 6: Most often selected parameters fitting the NOtr distribution and using maximum 5 predictors per season. The symbol asterisk (*) indicates the appearance of every parameter per specific season.

Parameter	*Winter*	*Spring*	*Summer*
Ensemble mean wind speed (mu and sigma)	*	*	*
Land type (mu and sigma)	*	*	*
Ensemble standard deviation of wind speed (sigma)	*	*	*

From the seasonal analysis so far, it is clear which atmospheric parameters can affect the wind speed and can be used for its prediction. Despite that the wind speed patterns are different between the three seasons, some parameters appeared more often than the others, like mean wind speed, land type, precipitation, latitude, 0m air temperature, zonal wind speed at different pressure levels, as well as surface air pressure and relative humidity. Concerning the land type, the following land types have been

chosen and they are: Land type 4, 6, 10, 11, 12, 18, 21, 24, 26, 27, 29, 30, 31, 32, 37, 42 and 44 (see in Annex, Table 2 for an explanation). Also, some parameters, like wind direction and momentum of roughness, which we expected to influence the wind speed at 10m more, were not chosen as predictors, neither using the different kind of distributions nor with QRF. Finally, parameters like turbulent kinetic energy and potential vorticity were chosen a few times, especially for long lead-times and it seems that they are good predictors for the extremes, as they are strongly connected with convection and the air parcel's circulation, respectively. Because of the peculiarity of the latitude as a predictor, which is probably a result of the station configuration in this study, the latitude was excluded as a predictor for the parameters of the truncated normal distribution in the remainder of this section.

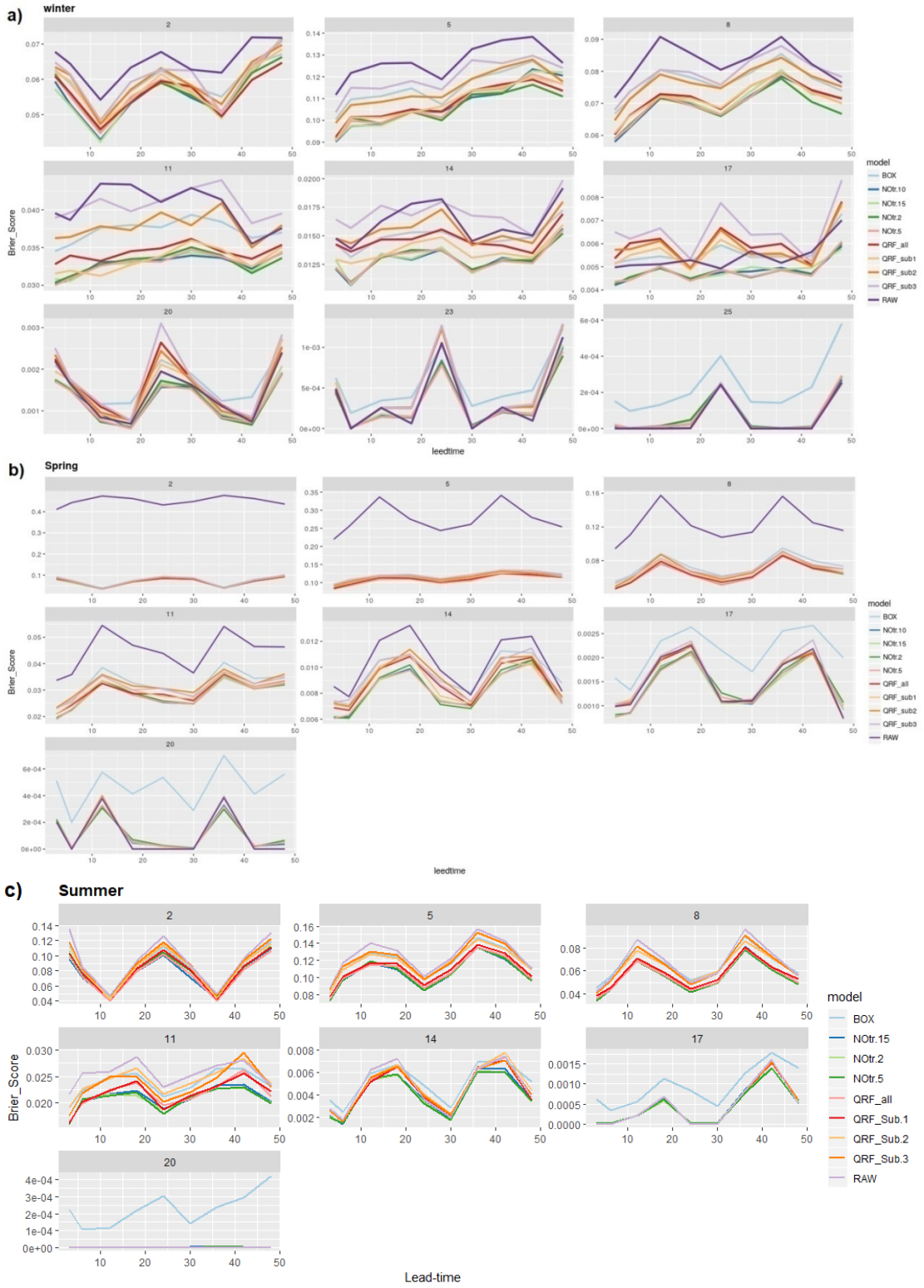
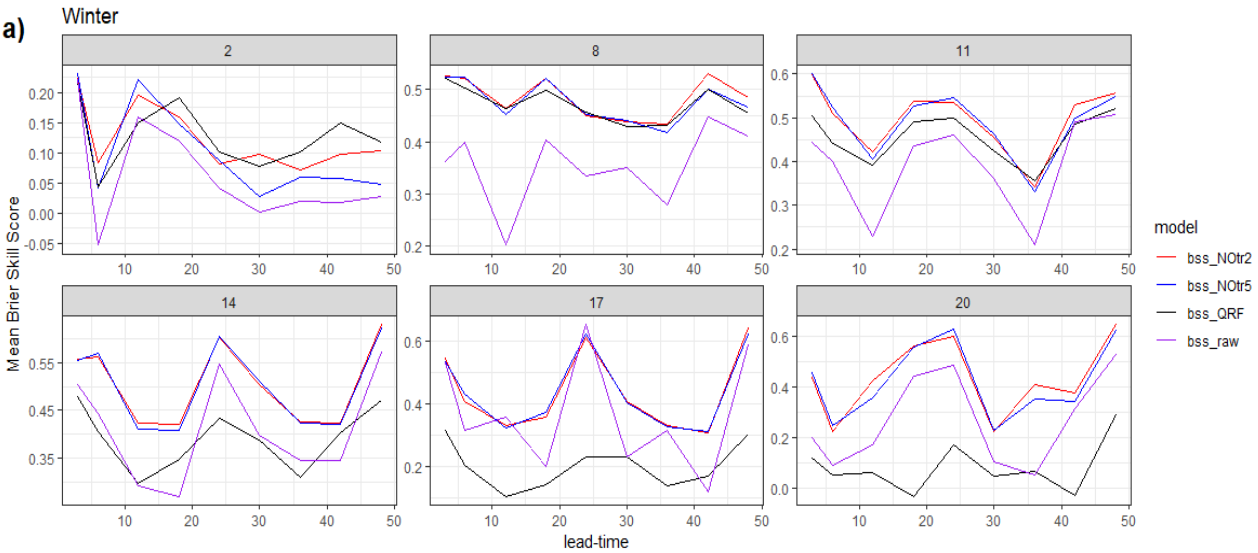


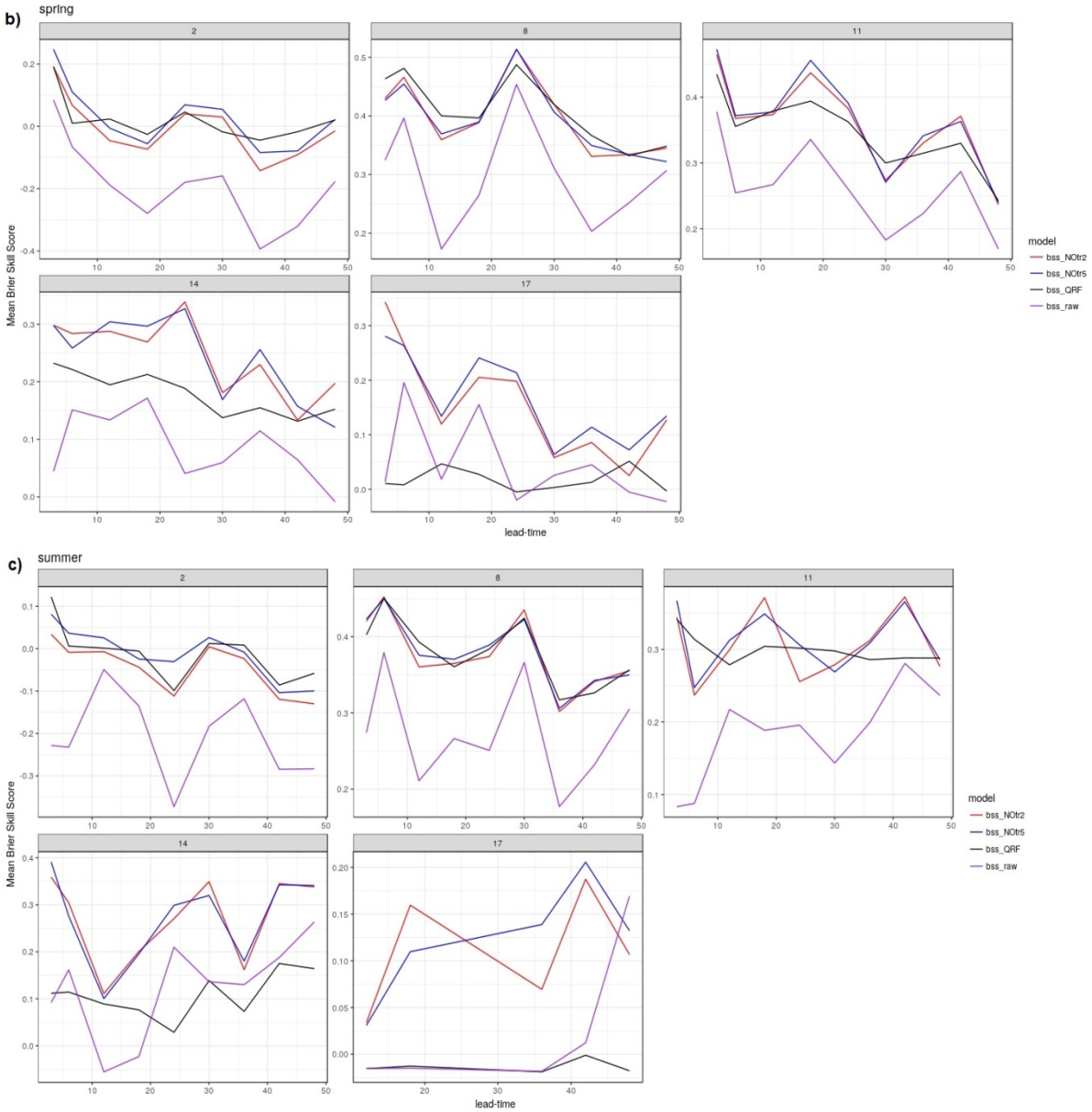
Figure 4: BS comparison between the NOtr, BCT distributions and QRF and QRF subsets, using the stepwise selection and different numbers of predictors for a) winter, b) spring and c) summer.

3.2.2 Verification of NOtr and QRF: BSS

From the analysis so far we can conclude which methods fitted better to our data. These are the NOtr_2 and NOtr_5 distributions and the QRF method for every season. Thus, in this subsection a comparative verification of these methods is presented using the BSS. In Fig. 5 the spatial mean BSS of these methods is shown for the three seasons and for specific thresholds. First, the BSS was calculated per station with the seasonal station climatology as the reference, and thereafter averaged. Remember that the best BSS is 1, thus positive BSS means that it is better than climatology and negative BSS that it is worse. In winter (Fig. 5a) for a moderate threshold (8 ms^{-1}) NOtr_2, NOtr_5 and QRF provided a higher mean BSS than the raw forecasts for every lead-time, while NOtr performed better for shorter than longer lead-times. Note that NOtr performed better than the QRF for the middle and higher thresholds. Indeed, the raw forecasts are more skilful than QRF for the higher thresholds (Fig. 5a, thresholds 14, 17, and 20 ms^{-1}). The BSS for the other two seasons (Fig. 5b, c) indicated similar results for the low thresholds as for winter, with the exception of forecasts for wind speeds exceeding the 2 ms^{-1}threshold, which cannot be skillfully forecast after a few hours lead time. More specifically, the BSS differences between NOtr_2, NOtr_5 and QRF are smaller than in Fig. 5a, but NOtr provided lower BSS for the extreme wind speed thresholds. From the comparison so far itis not easy to conclude which version of NOtr performed better, as the differences are quite small and further investigation will answer this question.

We can conclude that the BSS differences between NOtr and QRF are quite large for the higher wind speed thresholds. For the low threshold of 2 ms^{-1}, there are only a few observations lower than that threshold, especially at coastal or sea stations, which probably explains the low BSS values.



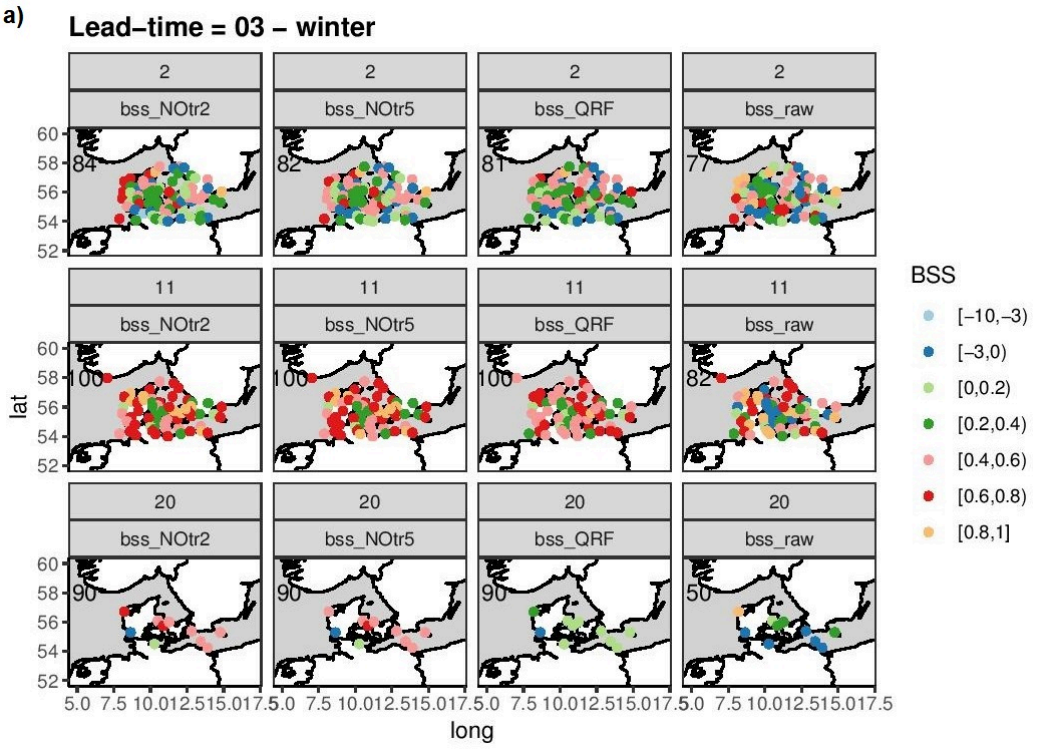


<u>Figure 5:</u> Spatial mean BSS comparison between NOtr_2,NOtr_5 (using the stepwise selection and different number of predictors), QRF-all and the raw forecasts, with the seasonal station climatology as the reference for a) winter, b) spring and c) summer.

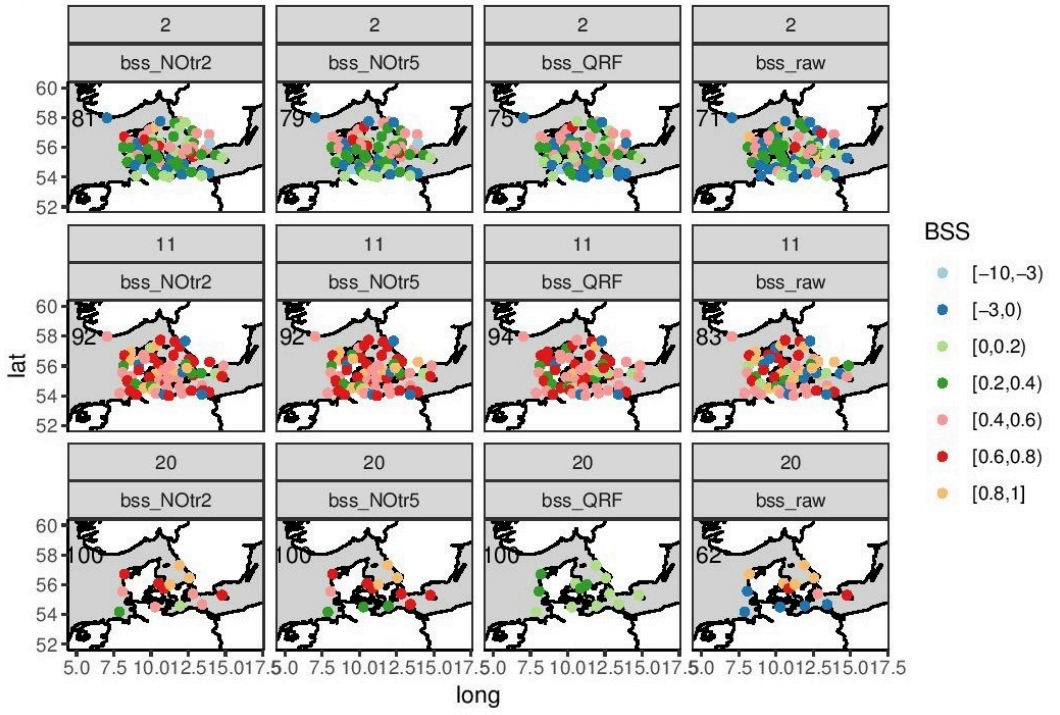
In Figure 6 the spatial BSS is presented in winter, as it is an interesting season for extreme events. The BSS was computed for three different lead-times and three specific thresholds (2, 11 and20 ms^{-1}) and for the NOtr_2, NOtr_5 and QRF methods and the raw forecasts. In general, the number of stations with positive BSS has increased when fitting the two distributions and QRF. Although the differences in low and intermediate thresholds between NOtr and QRF are small, NOtr_2 and NOtr_5 performed better in general. There is not a clear pattern between the coastal, sea and land stations. The forecast is more skillful for any type of station and it seems that the role of land type (ocean, land etc.) is important as a second predictor. In the case of extreme wind speed (>20ms^{-1}) the majority of the stations have BSS values greater than zero, but the number of stations is limited, as it is difficult to forecast extreme wind speed and there are also fewer cases of wind speed

larger than 20 m/s than for the other thresholds. For this threshold NOtr performs better in general than QRF with some coastal stations having BSS values greater than 0.8 (Fig. 6a, b).Note that the number of stations with a positive BSS is also larger for QRF than for the raw forecast, despite the fact that the raw forecast often has a larger mean BSS value for the higher thresholds (e.g. Figs. 5a and 6 for 20 ms^{-1}). This indicates that the raw forecast has larger variability of BSS values between stations, with many negative and higher positive values, while QRF has more values that are only moderately above zero. For a longer lead-time (48h forecast) the skill of the forecast has decreased for both calibrated and raw forecasts but still the NOtr provides a more skillful forecast as reflected by somewhat higher BSS values. Regarding the raw forecasts, there are stations, usually coastal and oceanic, with highly negative skill score.

a)


Lead–time = 03 – winter
2
bss_NOtr2
bss_NOtr5
bss_QRF
bss_raw
84
82
81
77
11
100
100
100
82
20
90
90
90
50
lat
long
BSS
[–10,–3)
[–3,0)
[0,0.2)
[0.2,0.4)
[0.4,0.6)
[0.6,0.8)
[0.8,1]

b)


Lead–time = 24 – winter
2
bss_NOtr2
bss_NOtr5
bss_QRF
bss_raw
81
79
75
71
11
92
92
94
83
20
100
100
100
62
lat
long
BSS
[–10,–3)
[–3,0)
[0,0.2)
[0.2,0.4)
[0.4,0.6)
[0.6,0.8)
[0.8,1]

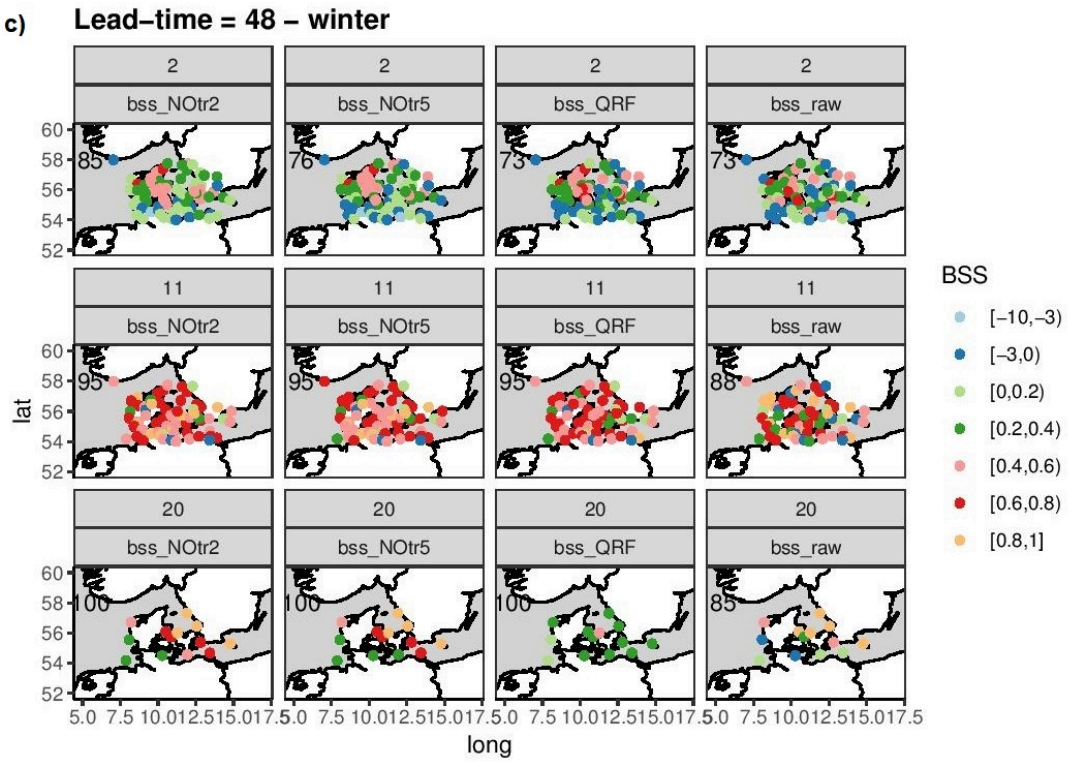


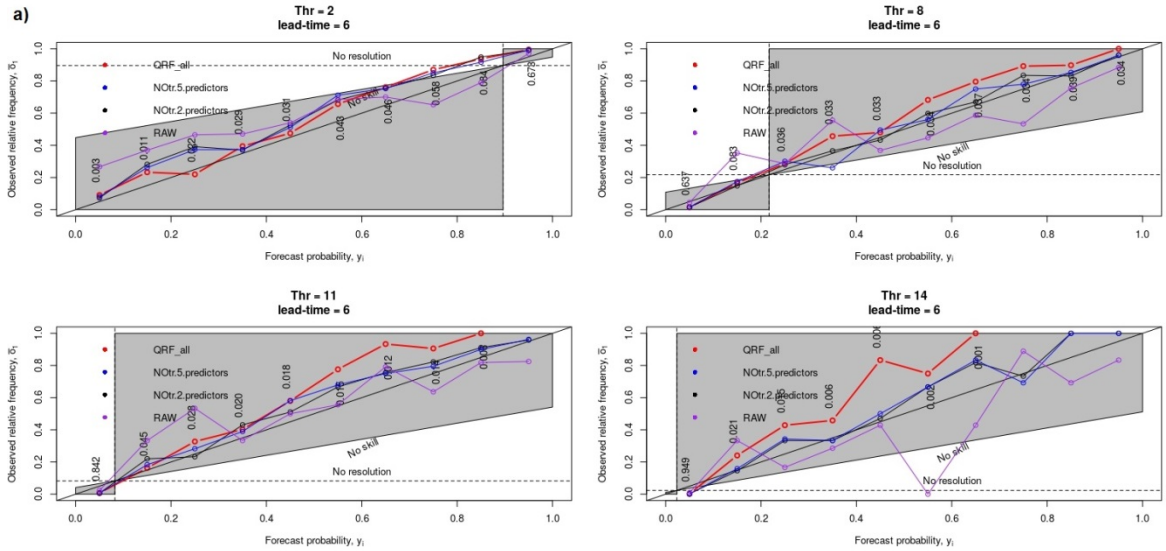
Figure 6: Spatial comparison of BSS values between the NOtr_2and NOtr_5 distributions, QRF and the raw forecasts, with the seasonal station climatology as the reference for winter and specific lead-times a) 03, b) 24 and c) 42 h. The number in the upper left corner of each map indicates the percentage of stations having positive BSS values.

3.2.3 Verification of NOtr and QRF: reliability diagrams and CRPSS

In this sub-section a comparative verification of NOtr and QRF in winter is presented, using reliability diagrams and CRPSS with the seasonal station climatology as the reference. The results for the spring and summer are presented in the Annex.

Firstly, the reliability diagrams are presented (Fig. 7a, b, c) for specific lead-times (6h, 24h and 42h) and thresholds (2, 8, 11 and 14 ms^{-1}), comparing the raw and calibrated forecasts, based on NOtr_2, NOtr_5, and QRF. If the reliability curve lies on the diagonal, it means that the forecast is perfectly reliable, while if itis above or below the diagonal then there is under- or overforecasting, respectively. For the thresholds of 2, 8 and 11 ms^{-1} and for every lead-time (Fig. 7a, b, c) the calibrated NOtr forecasts are more reliable than the QRF forecasts.

Comparing the three methods, there are small differences for low-thresholds, while for 11 $m*s^{-1}$ the calibrated forecasts using NOtr_2 and NOtr_5are more reliable than the raw forecasts and QRF, which show over- and underforecasting for higher forecast probabilities, respectively. In the case of 14 $ms^{-1}$the forecasts are reliable only for lower (≤ 50%) forecast probabilities and only in the case of NOtr_2 and NOtr_5.



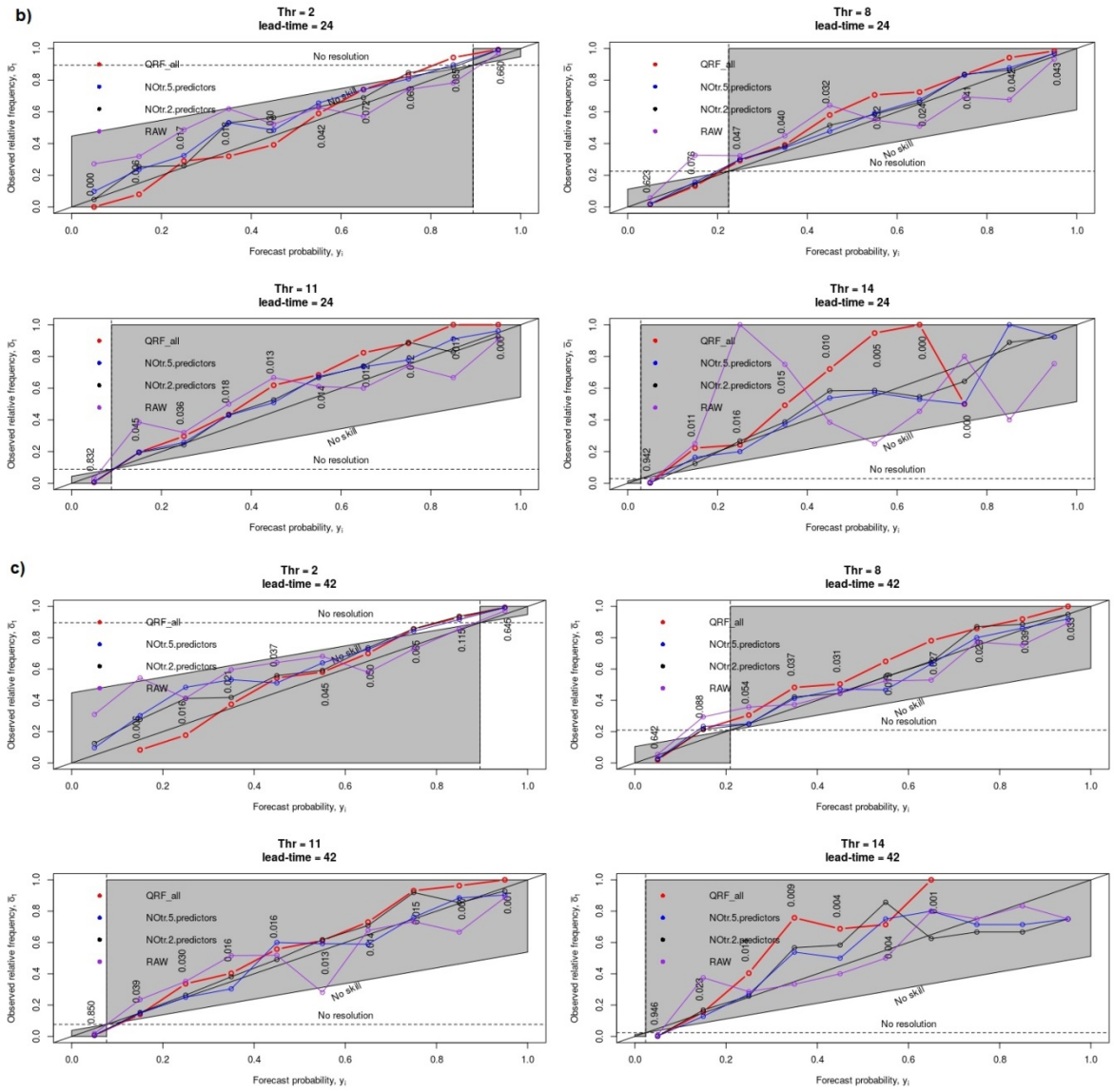


Figure 7: Reliability diagrams for wind speed greater than 2, 8, 11 and 14 ms^{-1} during the winter for raw and calibrated forecasts, using NOtr_2, NOtr_5and QRF, for lead-times a) 6h, b) 24h and c) 42h.

In the next two figures (Fig. 8 and 9) the CRPSS is presented using the station climatology as a reference. In Fig. 8 the mean CRPSS over all the stations is presented per lead-time, while in Fig. 9 the spatial coverage of CRPSS per station and specific lead-times is presented. The comparison concerns the NOtr_2, NOtr_5, QRF and raw forecasts as in the last figures.

Most importantly, the forecasts using the two main statistical post-processing methods are much more skilful than the raw forecasts for every lead-time (Fig. 8). Besides, the diurnal cycle in the CRPSS of 10m wind speed is noticeable, with lower values during day and higher values during night. Apart from the diurnal cycle the CRPSS of both the raw and calibrated forecasts is decreasing with lead-time, as expected.NOtr_2, NOtr_5and QRF perform equally well for lead-times till 36 h, but for longer lead times NOtr_2 and QRF seem to perform a bit better than NOtr_5. Therefore, the comparison between NOtr_2 and NOtr_5 suggests that the use of extra predictors does

not play an important role in improving the probabilistic forecasts. It has to be noted that the results of the mean CRPSS do not appear to be consistent with the results of mean BSS and BS, in terms that the NOtr_2 and NOtr_5 performed better than the QRF. This stems from the fact that the CRPSS compares the whole distribution and is heavily influenced by the bulk of the distribution and is less influenced by rare and extreme events. Indeed, for QRF the results for the BSS for higher thresholds and CRPSS seem not to be consistent. The comparison is more consistent for the low-middle thresholds (not shown), as there are more data available for this range. Also, the small training period probably leads to worse BSS and reliability results for the QRF for the higher thresholds.

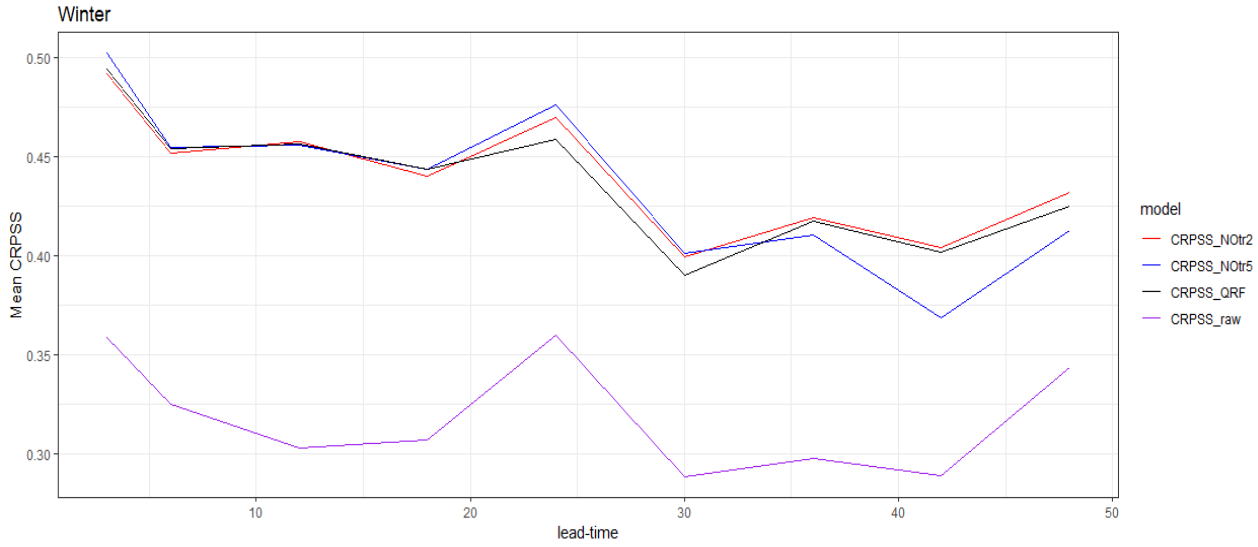


Figure 8: Mean CRPSS comparison between RAW and calibrated data, using NOtr_2, NOtr_5 and QRF for winter.

Finally, in Figure 9 the spatial coverage of the CRPSS per station is presented, for the NOtr_2, NOtr_5, QRF methods and the raw forecasts and for short and long lead-times. Positive CRPSS values appear for almost every station, and for short and long lead-times. Both the raw and calibrated forecasts are more skilful for shorter than longer lead-times, which is consistent with the previous graph (Fig. 8).

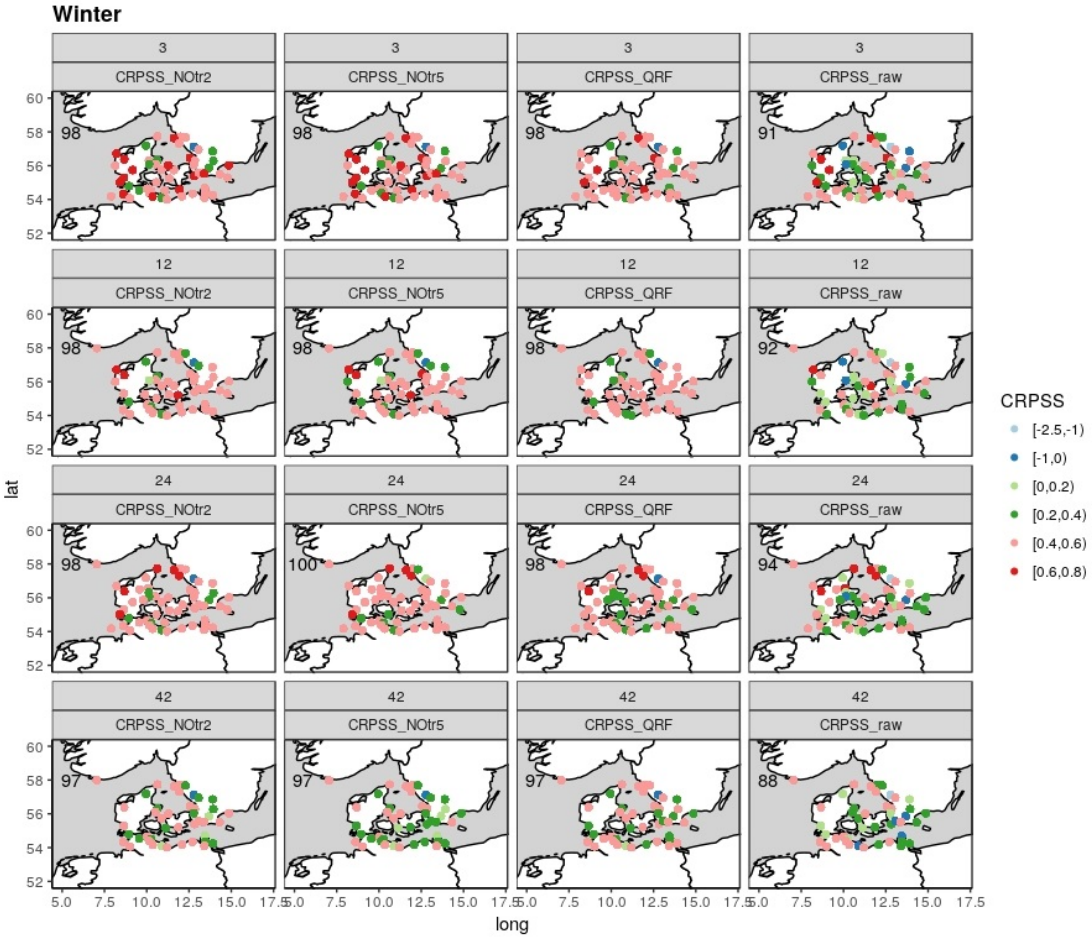


Figure 9: Spatial comparison of CRPSS between the NOtr_2, NOtr_5, QRF and raw forecasts, based on the seasonal station climatology for winter and specific lead-times. The number in the upper left corner of each map indicates the percentage of stations having positive CRPSS values.

3.3 Member by member approach

In this subsection the results using the MBM approach are presented. We split our data in three seasons and using the cross-validation method, as has been described in the methodology section, we applied the CRPS MIN and BEST REL methods. Schefzik (2017) and Schaeybroeck and Vannitsem (2015) have applied these methods, among others, to correct 2m temperature ensemble members and they found that the BEST REL and CRPS MIN performed better to their data than the other tested methods.

Because results of the MBM method have not been tested and/or described for wind speed in the literature, we compared the results also to a simple method, namely bias correction (results are not presented in this report). Also, we tested the MBM method using either one or two predictors. We chose the second predictor based on the results from the distributions and QRF, as has been described in subsection 3.2.

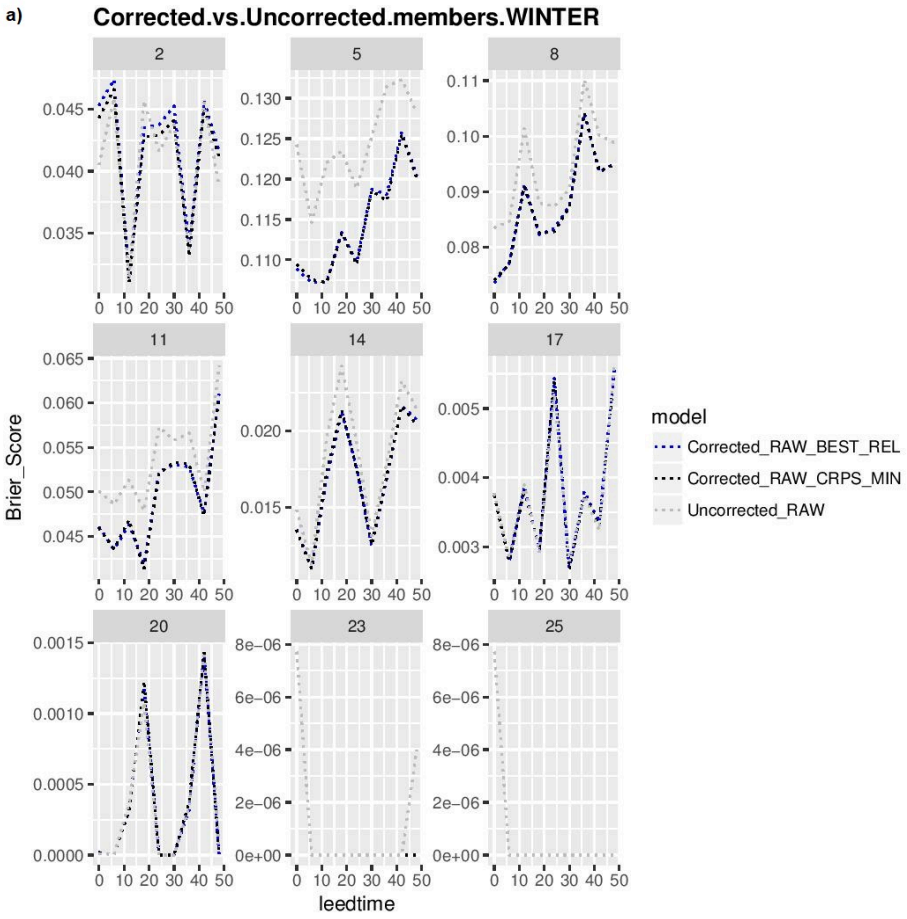
The experiment was performed for the winter and it was split in the middle, namely the first half was the test data and the other half was the training data. Also, we used five different combinations of predictors:

- 1 predictor: mean wind speed
- 2 predictors: mean wind speed and mean wind gust
- 2 predictors: mean wind speed and elevation
- 2 predictors: mean wind speed and latitude
- 2 predictors: mean wind speed and 0m air temperature

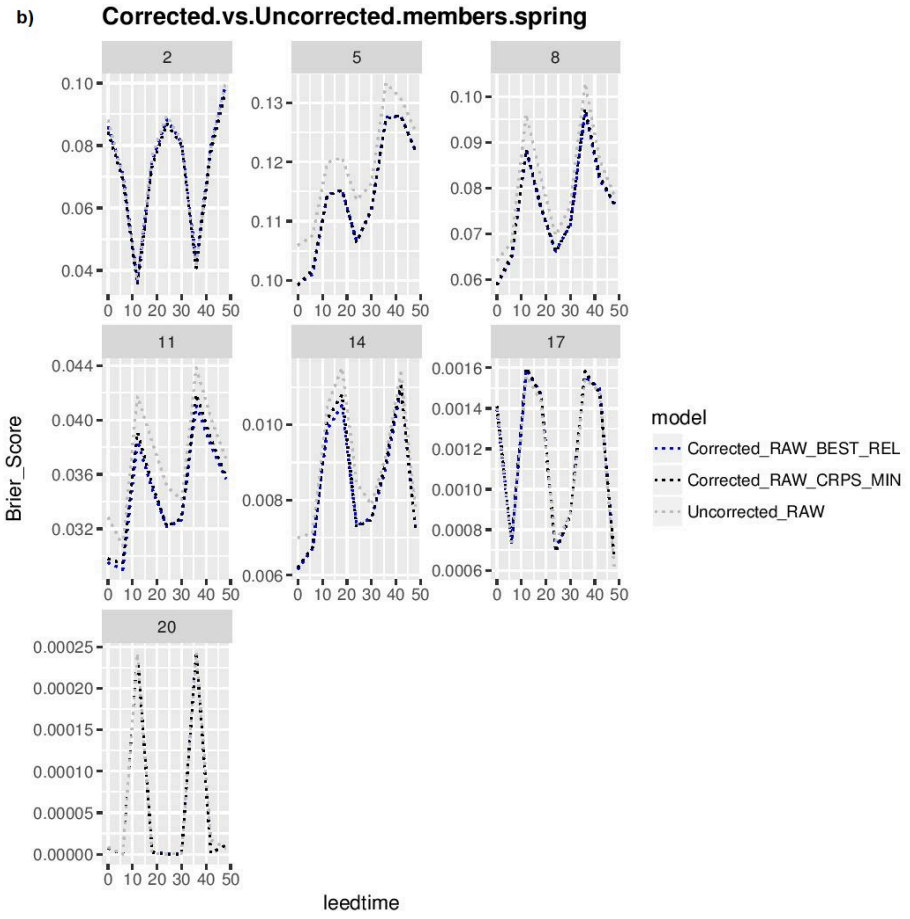
Initially, we found that the BS using the BEST REL and CRPS MIN has the lowest value for every lead-time and threshold. On the other hand, the BS using the bias correction was between the BS of those two methods and that of the uncorrected raw data. Concerning the five different combinations of predictors, we found that the BS has the smallest value when we used a second predictor. Despite the fact that the differences were small, the BS has been improved when the mean wind gust is used as the second predictor in the BEST REL and CRPS MIN methods. It has to be mentioned that the majority of the chosen second predictors were not always included in the top five lists from the NOtr and QRF results. Thus we tested predictors which can explain the meteorological and physical wind speed patterns in winter.

After these initial results for winter, we used a second predictor in the BEST REL and CRPS MIN methods and followed the same methodology for cross-validation as for the fitted distributions and the QRF. We chose the wind gust as the second predictor for winter and spring and the 0m air temperature for summer.

In Fig. 10a, b, and c, the BS comparison between the uncorrected raw data and the two methodologies using the MBM approach is presented for every season, respectively, and for the same thresholds as before.



b)


Corrected.vs.Uncorrected.members.spring
2
5
8
11
14
17
20
Brier_Score
leedtime
model
Corrected_RAW_BEST_REL
Corrected_RAW_CRPS_MIN
Uncorrected_RAW

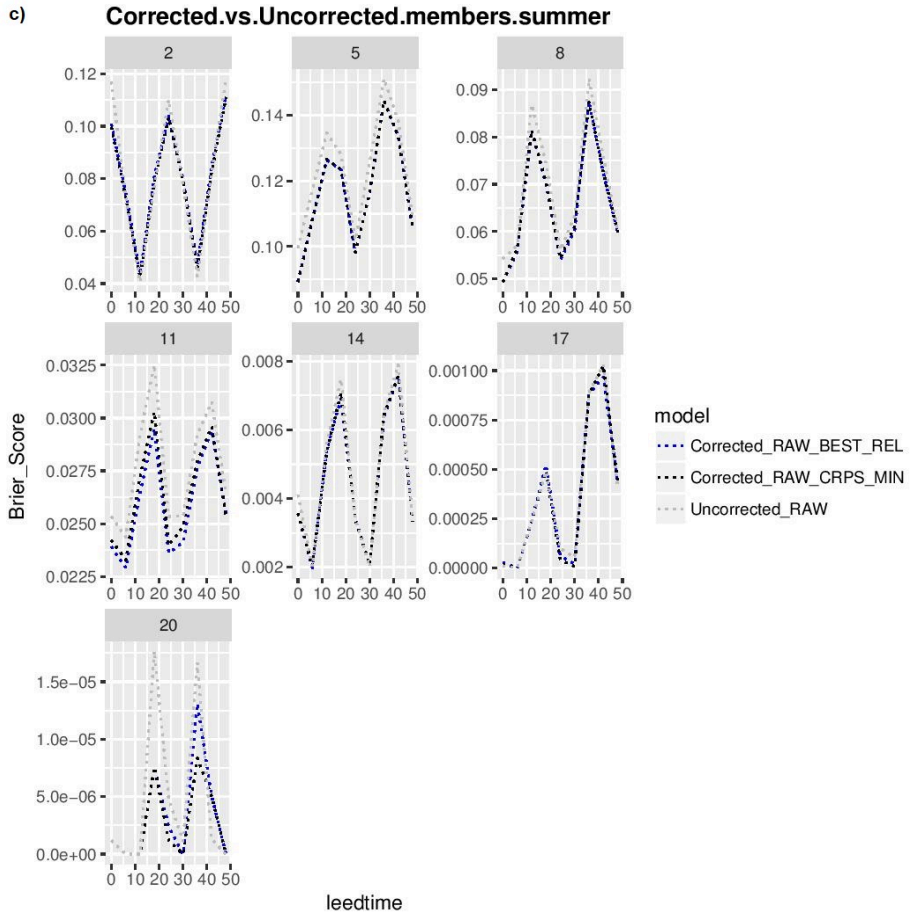


Figure 10: BS comparison between the uncorrected raw data and the corrected data based on the two MBM approaches using wind gust as a second predictor for a) winter, and b) spring, and 0m air temperature for c) summer.

Firstly, it has to be mentioned that the BS of the raw data in Figure 10 does not show the same diurnal cycle as the raw data in Figure 6. This happened because some cases have been excluded (per lead-time and season) in the MBM approach in order to be able to use the cross-validation method. Based on Figure 10, we can notice similar patterns for every season. The corrected data based on the BEST REL and CRPS MIN methods have decreased the BS for almost every lead-time and low to intermediate thresholds. It appears that the two methodologies did not improve the BS for the extremes values (thresholds > 17 m/s) in any season compared to the raw forecasts, while the BS has small improvements for middle thresholds (11 and 14 m/s). There is only one exception: in summer (Fig. 10c) for threshold equal to 20 m/s the BS is very close to zero and for some lead-times there are big differences (in magnitude) with the uncorrected raw data. But this result is probably not statistically significant, because there are not many cases of wind speed larger than 20 m/s. There are also some other exceptions, like in winter for example (Fig. 10a) and for threshold 2 m/s, in which case the BS of the raw forecasts is often slightly better than the BS of the corrected forecasts.

Finally, based on Figure 10, there are only small BS differences between the two methods. Again the BSS has been calculated using the seasonal climatology for every station and the results are in agreement with the BS.

The rest of the analysis is focusing on winter, as an example. The reliability diagrams for specific lead-times and thresholds are presented here. Besides, the CRPS was calculated including all the stations. The results for the other seasons are presented in the Annex.

Firstly, the CRPS of the raw forecasts and the two approaches of the MBM method using wind gust as the second predictor (Fig. 11) is presented. The CRPS is plotted as a function of lead-time and is generally increasing with lead-time. This indicates that the forecast is more skilful for shorter than longer lead-times, as expected. The corrected forecasts are more skillful than the raw forecasts for every lead-time and they show a similar diurnal cycle as the raw. Comparing the BEST REL and CRPS MIN there are very small differences in CRPS.

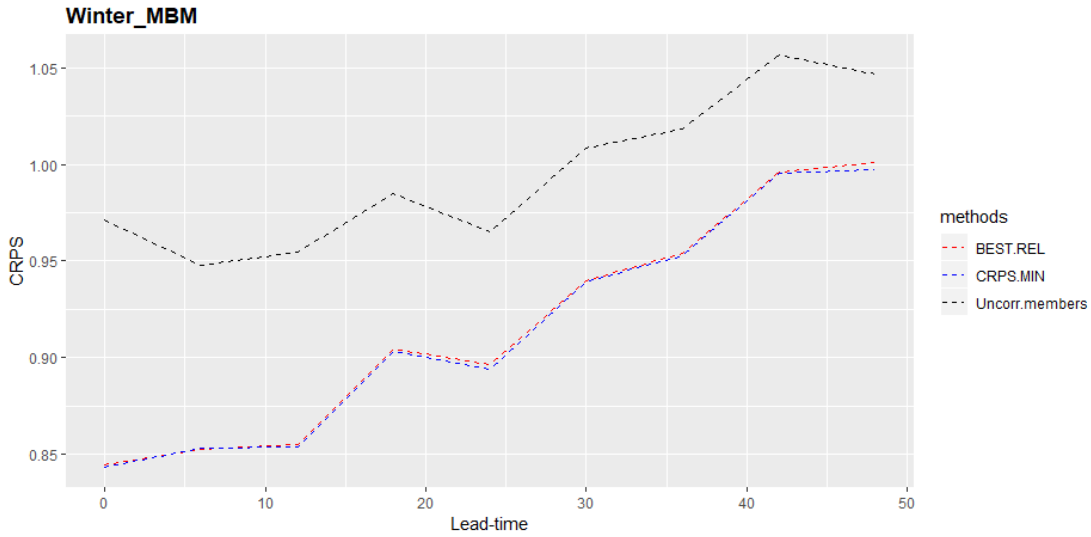
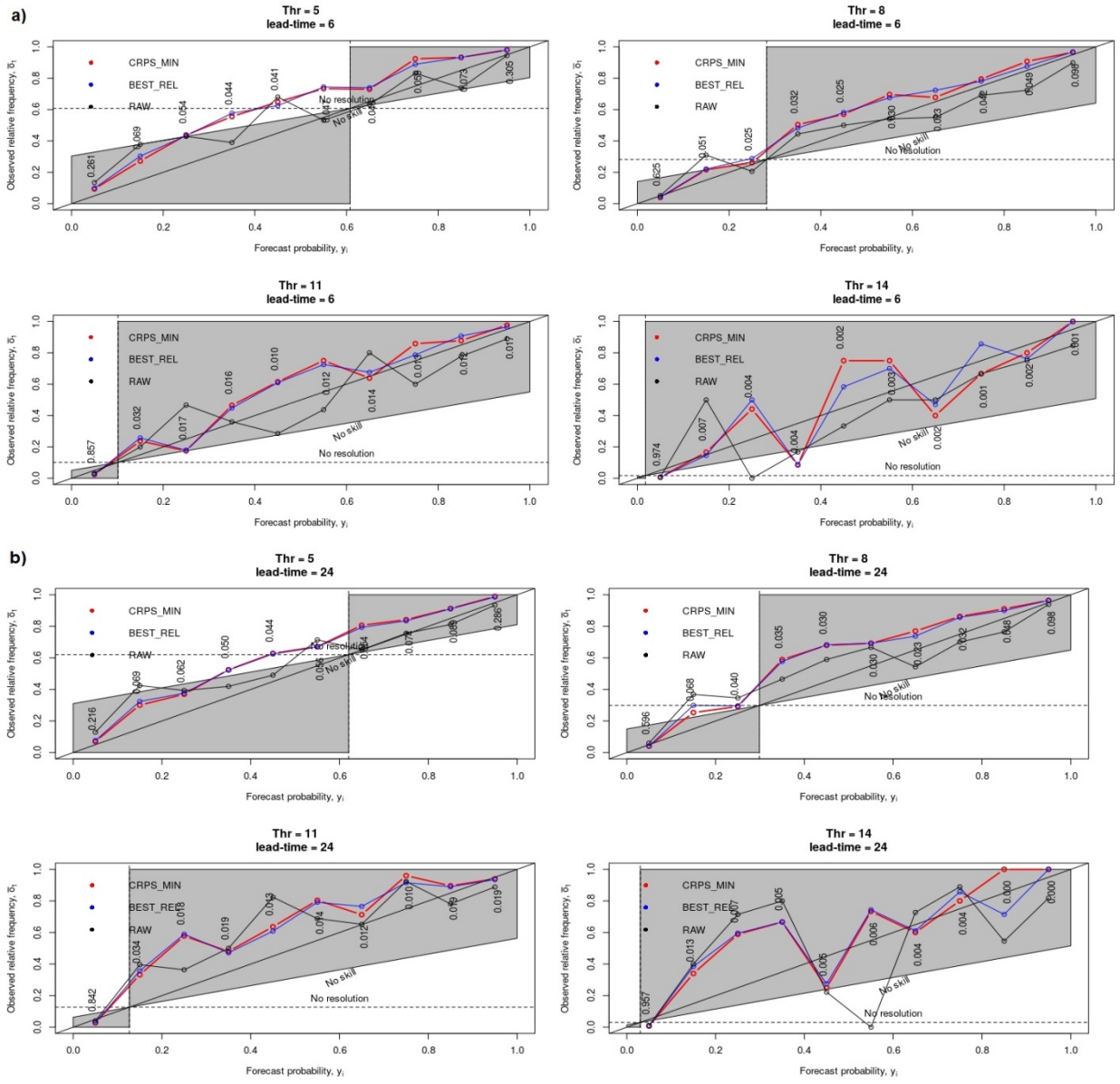


Figure 11: CRPS comparison between the uncorrected raw forecasts and the corrected forecasts based on the two MBM approaches and wind gust as a second predictor for winter.

Verification of wind speed forecasts for specific thresholds and lead-times is presented in Fig. 12a-c using reliability diagrams. Exceedance probabilities for all wind speed thresholds are generally under forecast. In general the post-processed forecasts have improved in terms of resolution, but decreased in terms of reliability compared to the raw forecasts.


a)
Thr = 5
lead-time = 6
Thr = 8
lead-time = 6
Thr = 11
lead-time = 6
Thr = 14
lead-time = 6
b)
Thr = 5
lead-time = 24
Thr = 8
lead-time = 24
Thr = 11
lead-time = 24
Thr = 14
lead-time = 24
CRPS_MIN
BEST_REL
RAW
No skill
No resolution
Observed relative frequency, $\bar{o}_i$
Forecast probability, y_i

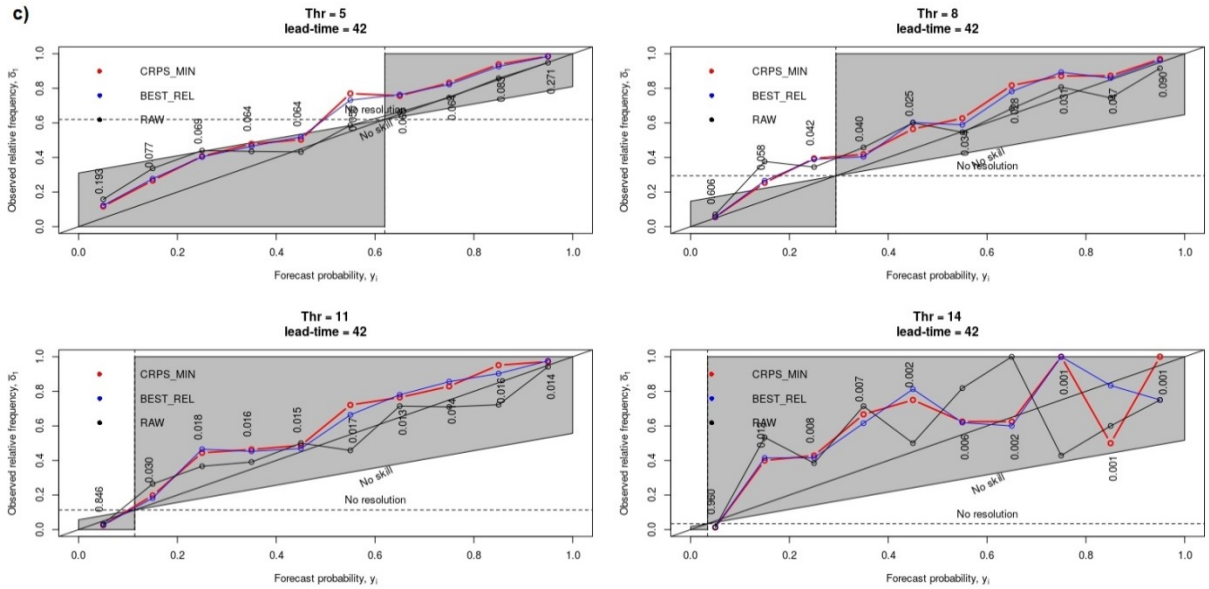


Figure 12: Reliability diagrams for wind speed greater than 5, 8, 11 and 14 m/s during the winter for the raw forecasts and the BEST REL and CRPS MIN MBM approaches for different lead-times: a) 6h, b) 24h and c) 42h.

4. Conclusions and discussion

Probabilistic wind speed forecasting using parametric and non-parametric post-processing methods has been studied. 10 ensemble members from the Harmonie MEPS and in-situ observations have been used, covering Denmark and the surrounding areas. The study period was between December 2016 and August 2017, and was split into three seasons, to study the wind speed forecast seasonally. Three main statistical post-processing methods have been used to improve the probabilistic wind speed forecast: 1) EMOS, 2) quantile regression forest (QRF) and 3) a member by member (MBM) method. More than 40 atmospheric parameters were used as potential predictors for 10m wind speed. The following main conclusions can be drawn from this analysis:

- Using the EMOS approach, a number of different distributions was verified, when the ensemble mean and standard deviation of wind speed were used as predictors. The BOX COX t and NOtr distributions fitted best to our data and provided more skilful forecasts (using the BS) than the other distributions. The calibrated forecasts were more skilful than the raw ensemble forecasts for every lead-time, threshold and season.
- Several sensitivity tests for NOtr and QRF were run, including testing different numbers and combinations of atmospheric variables as predictors, respectively. Based on BSS and reliability diagrams, the forecasts calibrated with NOtr_2 and NOtr_5 are more skilful than QRF, particularly at higher thresholds; they are more reliable for every season, lead-time and low to middle thresholds. QRF using all potential predictors performed better than the three QRF-subsets. When the CRPSS is computed per season, and averaged over all the stations, the differences between the three methods were small, but the NOtr with two predictors provided more skilful forecast for longer lead-times (42h and 48h) – in the winter. Comparing the BSS and the CRPSS in the case of QRF the results may appear not to be consistent, but this is due to the high weighting given to the bulk of the distribution in the CRPS. Some other reasons for the worse BSS results of QRF can be the lack of data for the higher thresholds, as well as the small training period which seems to affect QRF most.
- In every season and for every subset and lead-time two predictors were selected always for the mean in NOtr_5: ensemble mean wind speed and land type, while for the

standard deviation the two most common predictors were the ensemble standard deviation of wind speed and land type. The third, fourth and fifth predictors were not consistent between the lead-times, subsets and seasons and consisted of geomorphological and other meteorological parameters, like turbulent kinetic energy at different pressure levels.

- Finally, the MBM method was applied to wind speed data for the first time to our knowledge and yields more skilful forecasts than the raw forecasts, using the BEST REL and CRPS MIN approaches. More specifically, the forecast has improved for low and middle thresholds and for most lead-times, while for the extremes there was not a noticeable improvement. In this case only two predictors have been used: ensemble mean wind speed as the first predictor and the second predictor was dependent on the season (wind gust for winter and spring and 0m air temperature for summer).

In contrast to these results, Taillardat et al. (2016), comparing QRF and EMOS using four years of 35-member ensemble forecasts of wind speed and 87 French stations, found that QRF performed better than EMOS. However, in their study a long list of potential predictors is used only for QRF instead of our study in which more than 40 atmospheric parameters have been used as potential predictors for every method. They tested also more than one distribution (NOtr, Gamma, log-normal) and found that a version of the normal truncated distribution fitted better to their data set, as in our study.

Bremnes (2004) used a local quantile regression to forecast wind speed. He tested ten predictor combinations including extra predictors like wind direction and month, but he did not compare these results with other methods.

On the other hand, Han et al. (2018) compared six post-processing methods, among others EMOS and BMA. In the case of EMOS, they used the normal truncated distribution to fit to wind speed data over 51 ensemble member forecasts and observations from 26 stations in Pyeong Chang for a period of three years. They found that the skill of probabilistic forecasts using EMOS and BMA were better than for the other methods but with small differences.

Other studies, like Al Buhairi (2006) and Amaya-Martinez et al. (2014) compared different distributions for wind speed forecasts, including the Weibull, Gamma, Log-normal and Rayleigh distributions without using any extra predictor. According to our results the first three mentioned distributions performed worse than NOtr and non-parametric QRF in 3-month data sets of winter, spring and summer.

However, a further investigation of the NOtr and QRF methods would be advisable, as the study period was only three seasons and according to the literature QRF performs better using longer time periods. Concerning the MBM method, some more sensitivity tests using different second predictors,

as for example turbulent kinetic energy at different pressure levels (925 or 1000 hPa) might provide more skilful forecasts, as the wind gust was chosen only once in NOtr_5 in winter. Finally, as this study was a case study for Denmark, some more testing for every method is advisable for the Netherlands using the Harmonie KEPS output.

Acknowledgments

The authors are grateful to Bert van Schaeybroeck (RMI Belgium) for delivering the Fortran code for the MBM method and to John Bjørnar Bremnes (Met Norway) for compiling and providing the dataset on which this study is mostly based.

References

Akaike H.: A new look at the statistical model indentification. IEEE Transactions on Automatic control, 1974, 716-723, doi: 10.1109/TAC.1974.1100705

Amaya-Martinez P.-A., Saavedra-Montes A.J. and Arango-Zuluaga E.-I.: A Statistical Analysis of Wind Speed Distribution Models in the Aburra Valley, Colombia, CT&F – Ciencia, Tecnologia y Futuro – Vol. 5, Num. 5, Dec. 2014, Pag. 121-136

Andrae U.: Experiences and challenges with MetCOop EPS, SMHI, Head of MetCoOp development, Finnish Meteorological Institute

Baran S.: Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. Computational Statistics and Data Analysis 75 (2014) 227-238

Baran S. and Lerch S.: Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind-speed forecasting, Q.J.R. Meteorol. Soc. 141:2289, July 2015 B DOI: 10.1002/qj.251

Bengtsson L., Andrae U., Aspelien T., Batrak Y., Calvo J., Rooy W.D., Gleeson E., Hansen-Sass B., Homleid M., Hortal M., Ivarsson K.-I., Lenderink G., Niemela S., Nielsen K. P., Onvlee J., Rontu L., Samuelsson P., Munoz D.S., Subias A., Tijm S., Toll V., Yang X. and Koltzow M.O.: The HARMONIE-AROME Model Configuration in the ALADIN-HIRLAM NWP System, 2017 American Meteorological Society. DOI: 10.1175/MWR-D-16-0417.1

Breiman, L.: Random Forests, 2001, Machine Learning, 5-35.

Bremnes J.B.: Probabilistic Wind Power Forecasts Using Local Quantile Regression, Wind Energy 2004; 7:47-54 (DOI: 10.1002/we.107)

Brown B.G.: Verification of ensembles, Copyright UCAR 2015

Buhairi M.H.: A statistical analysis of wind speed data and an assessment of wind energy potential in Taiz-Yemen. Ass. Univ. Bull. Environ. Res. Vol. 9, No.2, October 2006

Domenech M.R., Schmeits M. and Whan K.: Three results around the GLAMEPS wind forecasts, 2017, draft Master thesis

Ebisuzaki, W., and Kalnay. E., 1991: Ensemble experiments with a new lagged average forecasting scheme. WMO, Research activities in atmospheric and oceanic modeling. Report #15, pp6.31-6.32. [Available from WMO, C.P. No 2300, CH1211, Geneva, Switzerland]. See also Kalnay (2003), p. 234

Frogner I.L. and the HIRLAM EPS and predictability team: Representation of model uncertainty in a convection permitting EPS – HarmonEPS, Reading, November 2017

Gneiting T., Larson K., Westrick K., Genton M. G. And Aldrich E.: Calibrated Probabilistic Forecasting at the Stateline Wind Energy Center: The Regime-Switching Space-Time (RST) method, Technical report no.464, Department of Statistics, University of Washington, 2004

Hamill T.M.: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, American Meteorological Society 2001, Volume 129

Han K., Choi J.T. and Kim C.: Comparison of Statistical Post-Processing Methods for Probabilistic Wind Speed Forecasting, Asia-Pac. J. Atmos. Sci., 54(1), 91-101,2018. DOI: 10.1007/s13143-017-0062-z

Hermi S., Scheuerer M., Pappenberger F., Bogner K, and Haiden T.: Trends in the predictive performance of raw ensemble weather forecasts, 2014, Geophysical Research Letters, 41, 9197-9205.

Jewson S., Brix A. and Ziehmann C.: A new parametric model for the assessment and calibration of medium-range ensemble temperature forecasts, 2004, Atmospheric Science Letters, 5, 96-102.

Kalnay, E, 2003: Atmospheric modeling, data assimilation and predictability. Cambridge University Press, 341 pp.

Lerch S. and Thorarinsdottir T.L.: Comparison of non-homogenous regression models for probabilistic wind-speed forecasting, 2013, Tellus A 65: 21206, doi: 10.3402/tellusa.v65i0.21206

Meinshausen N.: Quantile Regression Forests, Journal of Machine Learning Research 7, 2006, 983-999

Meinshausen N.: Package "quantregForest", December 2017, http://github.com//lorismichel/quantregForest.

Nadarajah S. and Kotz S.: R Programs for Computing Truncated Distributions, Journal of Statistical Software, August 2006, Volume 16, Code Snippet 2.

Sakamoto Y., Ishiguro M. and Kitagawa G.: Akaike information criterion statistics, 1986, Dordrecht, The Netherlands: D.Reidel

Stasinopoulos D.M. and Rigby R.A.: Generalized Additive Models for Location Scale and Shape (GAMLSS) in R, Journal of Statistical Software, December 2007, Volume 23, Issue 7.

Stasinopoulos M., Rigby B. and Akantziliotou C.: Instructions on how to use the gamlss package in R, Second Edition, January 2008.

Taillardat M., Mestre O., Zamo M. and Naveau P.: Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics, American Meteorological Society 2016, DOI: 10.1175/MWR-D-15-0260.1

Thorarinsdottir T.L. and Gneiting T.: Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression, 2010. Journal of the Royal Statistical Society L Series A (Statistics in Society), 371-388, doi:10.1111/j.1467-985X.2009.00616.x

Schaeybroeck Van Bert and Vannitsem Stephane: Ensemble post-processing using member-by-member approaches: theoretical aspects, 2015, Q.J.R. Meteorol. Soc. 141:807-818, DOI: 10.1002/qj.2397

Schefzik Roman: Ensemble calibration with preserved correlations: unifying and comparing ensemble copula coupling and member-by-member postprocessing, Q. J. R. Meteorological Society 143:999-1008, January 2017 B DOI: 10.1002/qj.2984

Weigel A.P., Liniger M.A. and Appenzeller C.: The Discrete Brier and Ranked Probability Skill Scores, Monthly Weather Review, Volume 135, 2007 American Meteorological Society, DOI: 10.1175/MWR3280.1

Wilks, D.S.: Statistical Methods in the Atmospheric Sciences. Academic Press., 2011

Annex

Table 1: Station's information in Denmark and surrounding areas.

a/a	Station's Number	Height (m)	Latitude	Longitude
1	2607	18	56.29778	12.84417
2	10091	42	54.68167	13.43667
3	10170	4	54.18167	12.08222
4	10022	7	54.79028	8.95139
5	10028	5	54.32778	8.60278
6	10042	2	54.64084	10.02361
7	10093	40	54.36417	13.47694
8	10097	12	54.24361	13.91000
9	10130	3	54.06917	9.01056
10	10150	26	54.16528	10.35167
11	10152	1	54.08917	10.87722
12	10161	15	54.00445	11.19222
13	2518	2	57.30667	11.91083
14	2618	135	56.94950	13.06017
15	2629	103	56.27445	13.93778
16	2611	44	56.03972	12.77556
17	2635	21	55.57472	13.07806
18	2513	3	57.71567	11.99250
19	2623	114	55.85222	13.67056
20	2539	122	57.12278	12.78194
21	2628	60	56.02333	14.86000
22	2622	148	56.85444	13.89278
23	2615	4	55.38367	12.81667
24	2625	5	55.49278	14.31778
25	2605	10	56.46056	12.55639
26	2516	19	57.63217	11.60483
27	1400	52	56.54667	3.21194
28	1436	13	57.98278	7.04778
29	6206	28	54.11667	4.01222
30	6239	24	54.85389	4.69611
31	6201	36	54.32567	2.93575
32	6205	33	55.39917	3.81028
33	6214	37	54.03697	6.04161
34	6141	8	54.82750	11.32889
35	6174	3	54.87889	12.18417
36	6151	2	55.15917	11.13472
37	6168	37	56.11917	12.34222
38	10035	47	54.53333	9.55000
39	6060	53	56.29345	9.11384
40	6070	23	56.30825	10.62542
41	6120	21	55.47488	10.33049
42	6193	9	55.29795	14.77177

43	6019	41	56.92917	8.64111
44	6056	53	56.38306	8.67028
45	6068	60	56.09389	9.18111
46	6073	4	56.09556	10.51361
47	6079	2	56.71695	11.50972
48	6081	19	55.55750	8.08278
49	6093	3	55.29111	8.65500
50	6096	6	55.19056	8.56000
51	6119	17	54.85278	9.98806
52	6135	33	55.32167	11.38806
53	6138	2	54.82056	10.99389
54	6174	21	55.39556	12.14917
55	6181	40	55.76639	12.52639
56	6183	7	55.53639	12.71139
57	6031	13	57.18516	9.95268
58	6109	62	55.47147	9.11237
59	6124	7	55.01436	10.56934
60	6149	5	54.56869	11.94347
61	6159	16	55.743484	10.86936
62	6169	16	56.00830	11.27872
63	6188	41	55.87645	12.41208
64	6032	56	57.38277	10.33492
65	6041	5	57.73639	10.63164
66	6049	89	56.56043	10.09289
67	6052	4	56.70679	8.21495
68	6058	4	56.00725	8.14128
69	6065	33	56.75582	9.50674
70	6072	62	56.30270	10.12721
71	6074	56	56.08035	10.13529
72	6082	25	55.95905	8.62420
73	6102	24	55.86795	9.78719
74	6110	47	55.22516	9.26336
75	6123	3	55.24436	9.88817
76	6126	51	55.30878	10.43983
77	6136	12	55.24646	11.32851
78	6154	46	55.20747	11.86045
79	6156	13	55.73578	11.60352
80	6170	43	55.58685	12.13625
81	2526	154	57.67611	12.29194
82	2636	73	55.52306	13.37889
83	10015	4	54.17500	7.89167
84	10020	26	55.01111	8.41250
85	10055	3	54.52973	11.06194
86	10184	2	54.09778	13.40750
87	10004	0	54.16667	6.35000
88	10007	0	54.18334	7.43333
89	10044	5	54.50028	10.27472
90	6104	80	55.73794	9.16741

91	10046	28	54.37611	10.14332
83	10180	4	54.34058	12.71083
93	6180	5	55.61404	12.64535
94	2616	5	55.38389	12.81944
95	6030	13	57.09627	9.85051
96	6069	58	56.49315	9.57095
97	10067	5	54.49361	11.24056

Table 2: Corine Land Cover data (https://land.copernicus.eu/pan-european/corine-land-cover).

Grid code	CLC code	Label
1	111	Continuous urban fabric
2	112	Discontinuous urban fabric
3	121	Industrial or commercial units
4	122	Road and rail networks and associated land
5	123	Port areas
6	124	Airports
7	131	Mineral extraction sites
8	132	Dump sites
9	133	Construction sites
10	141	Green urban sites
11	142	Sport and leisure facilities
12	211	Non-irrigated arable land
13	212	Permanently irrigated land
14	213	Rice fields
15	221	Vineyards
16	222	Fruit trees and berry plantations
17	223	Olive groves
18	231	Pastures
19	241	Annual crops associated with permanent crops
20	242	Complex cultivation patterns
21	243	Land principally occupied by agriculture, with significant areas of natural vegetation
22	244	Agro-forestry areas
23	311	Broad-leaved forest
24	312	Coniferous forest
25	313	Mixed forest
26	321	Natural grasslands
27	322	Moors and heath land
28	323	Sclerophyllous vegetation
29	324	Transitional woodland
30	331	Beaches, dunes, sands
31	332	Bare rocks
32	333	Sparsely vegetated areas
33	334	Burnt areas

34	335	Glaciers and perpetual snow
35	411	Inland marshes
36	412	Peat bogs
37	421	Salt marshes
38	422	Salines
39	423	Intertidal flats
40	511	Water sources
41	512	Water bodies
42	521	Coastal lagoons
43	522	Estuaries
44	523	Sea and ocean
48	999	No Data
49	990	Unclassified land surface
50	995	Unclassified water bodies

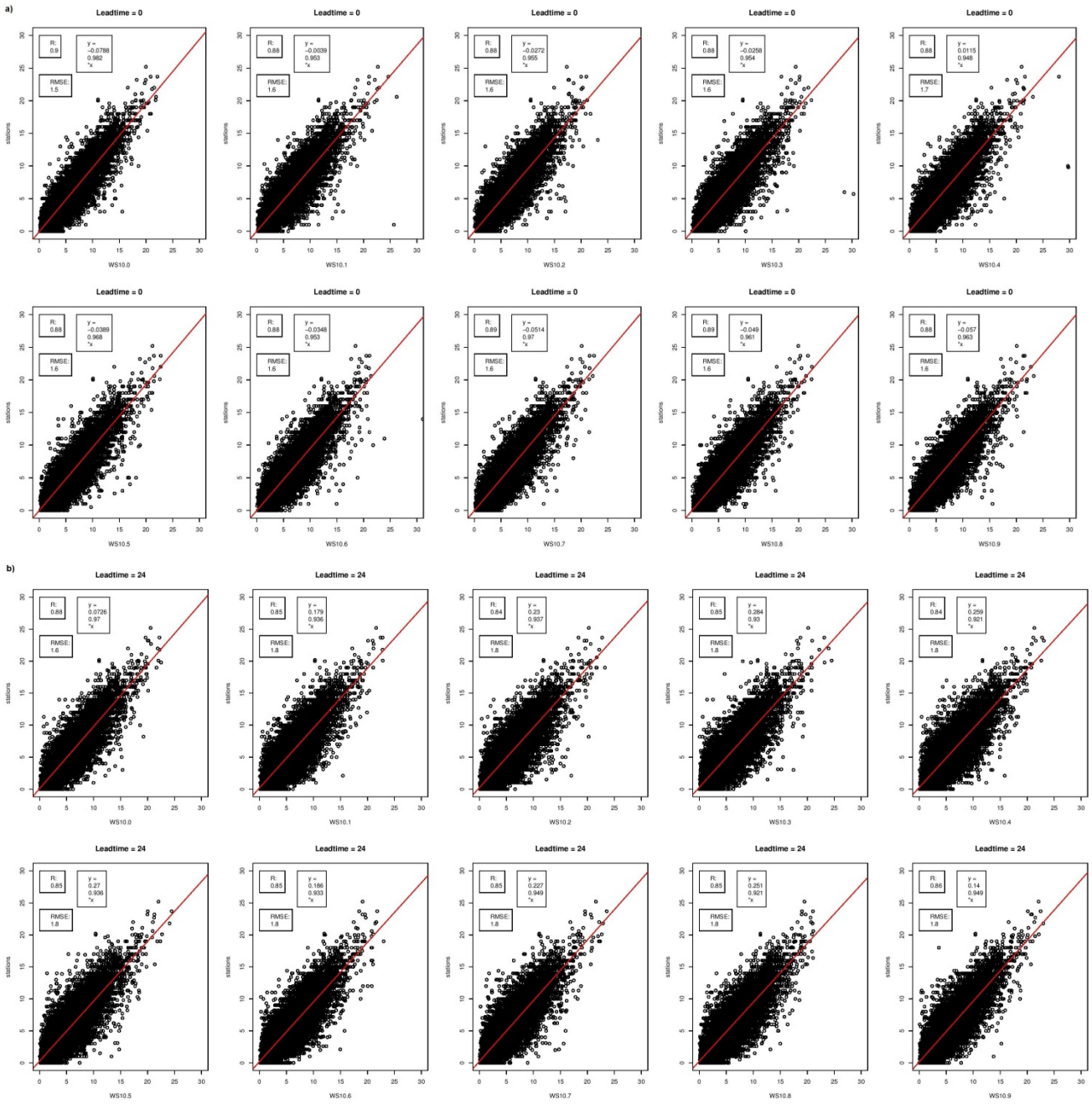
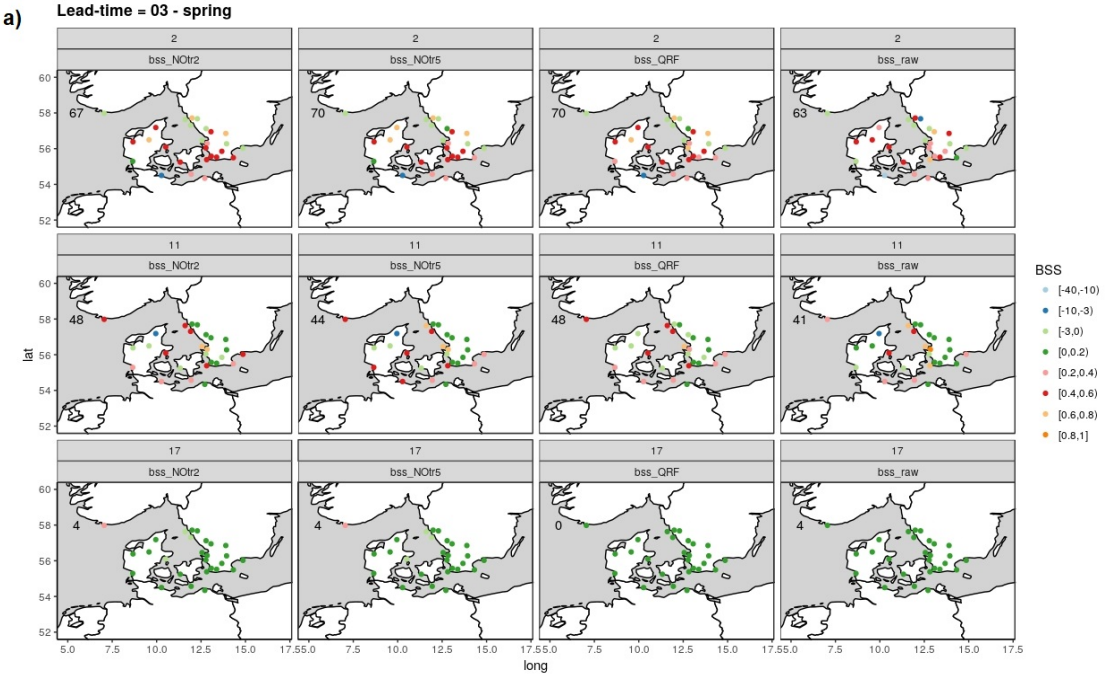
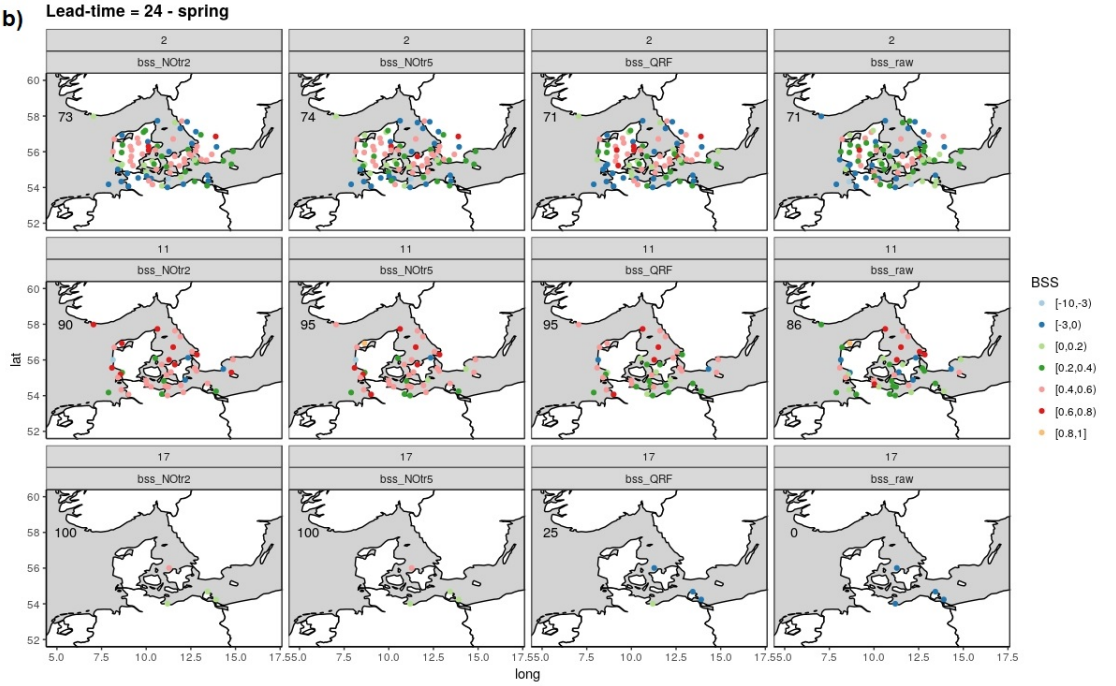


Figure 1: Scatter plots of 10m wind speed observations versus the ten ensemble members for lead-time a) 0 and b) 24h. R and RMSE are presented in every panel.

a) **Lead-time = 03 - spring**



b) **Lead-time = 24 - spring**



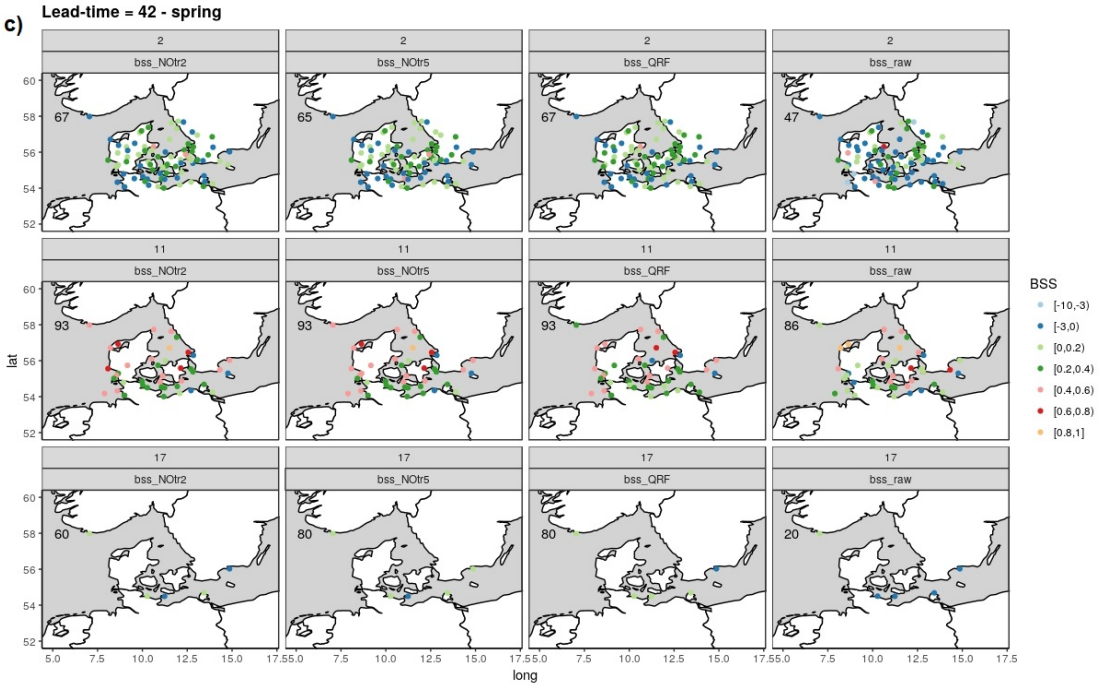
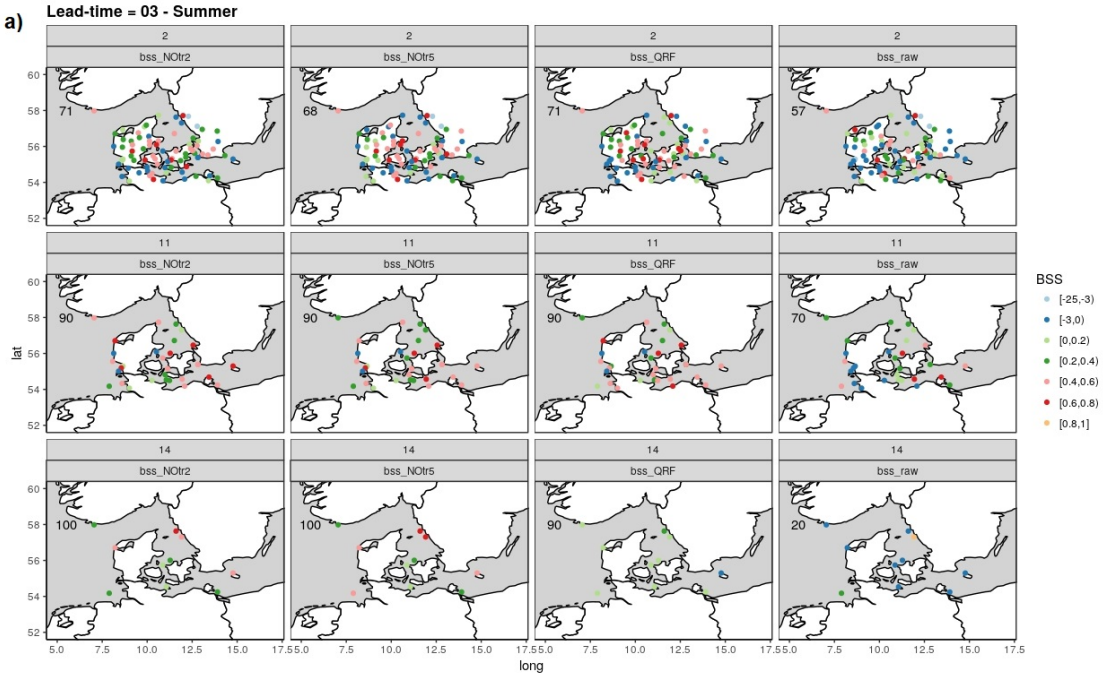


Figure 2: BSS comparison between NOtr_2, NOtr_5 distributions and QRF, based on the seasonal station climatology for spring and specific lead-times a) 03, b) 24 and c) 42 h. The number in the upper left corner of each map indicates the percentage of stations having positive BSS values.



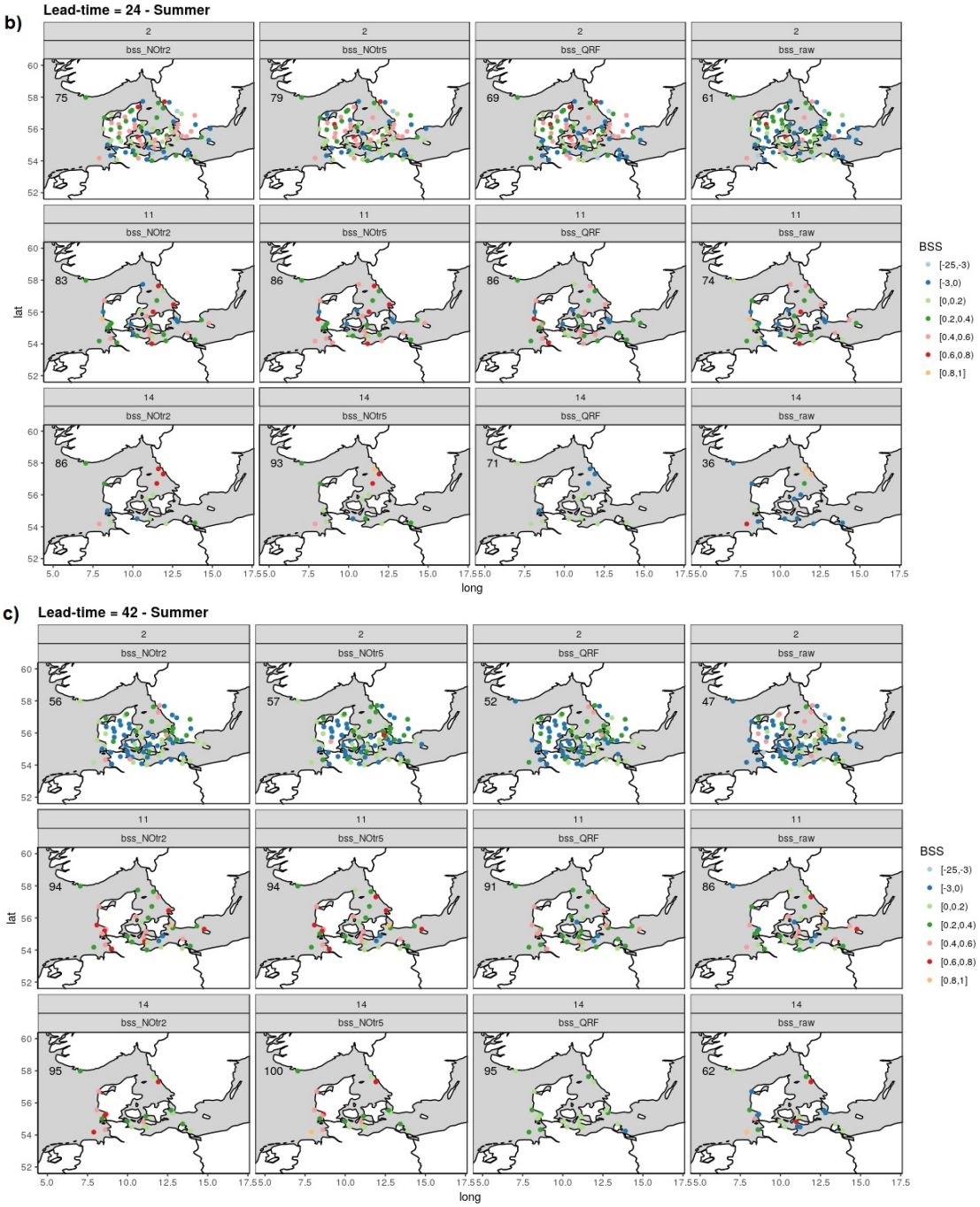


Figure 3: As Figure 2 but for summer.

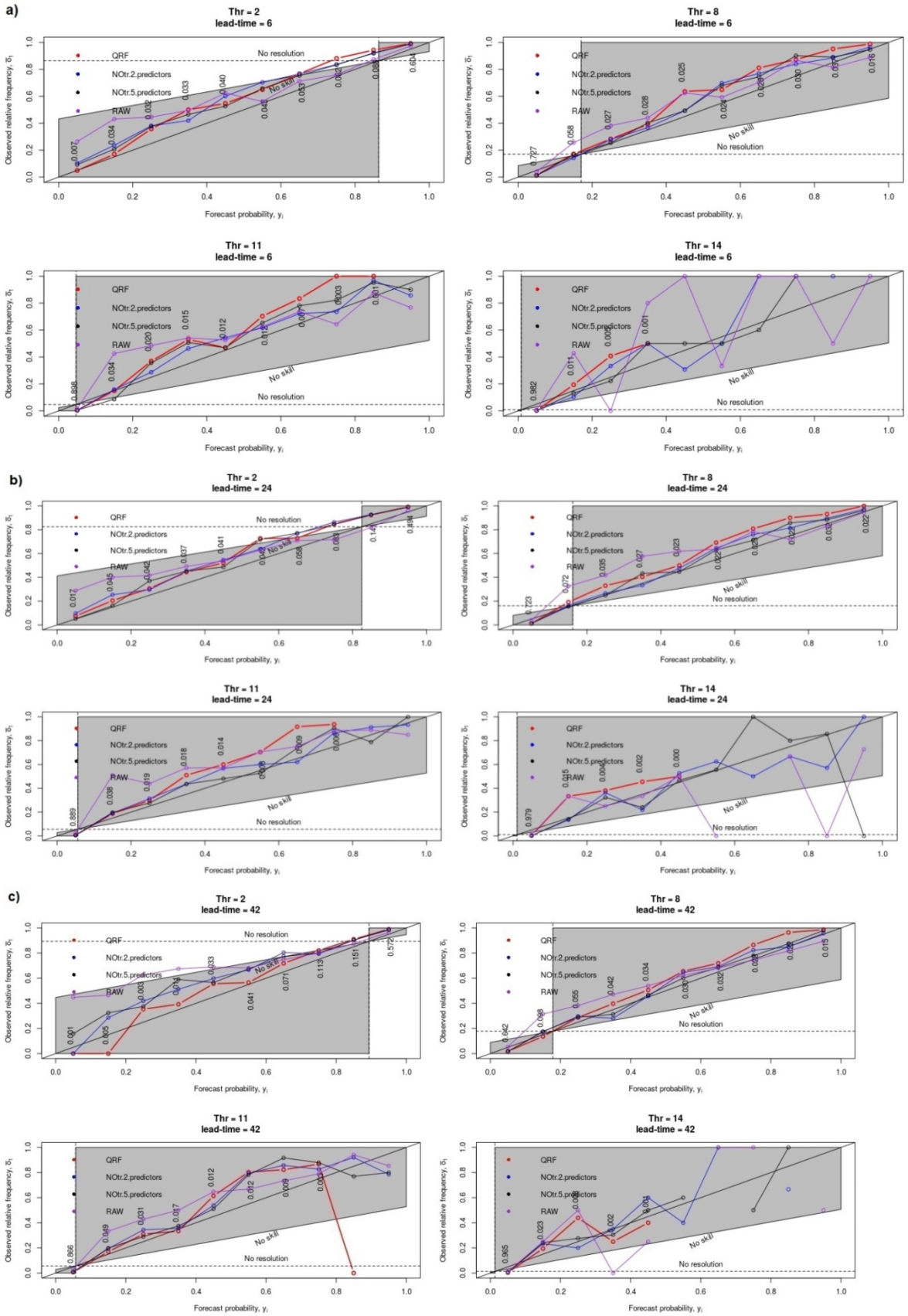
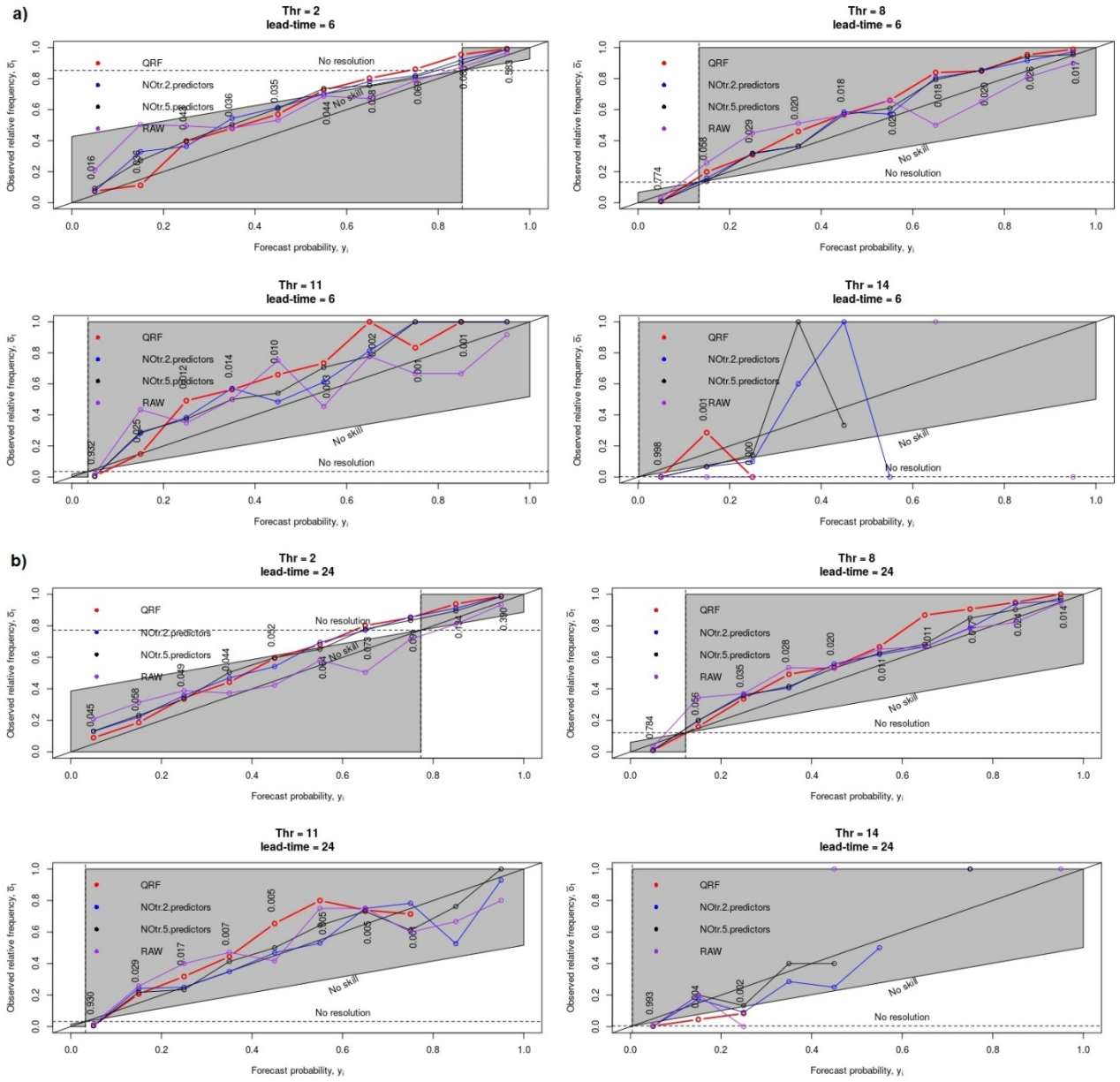


Figure 4: Reliability diagrams for wind speed greater than 2, 8, 11 and 14 ms^{-1} during the spring for raw and calibrated forecasts, using NOtr_2, NOtr_5 and QRF, for lead-times a) 6h, b) 24h and c) 42h.


a)
Thr = 2
lead-time = 6
Thr = 8
lead-time = 6
Thr = 11
lead-time = 6
Thr = 14
lead-time = 6
b)
Thr = 2
lead-time = 24
Thr = 8
lead-time = 24
Thr = 11
lead-time = 24
Thr = 14
lead-time = 24
QRF
NOtr.2.predictors
NOtr.5.predictors
RAW
No resolution
No skill
Observed relative frequency, $\bar{o}_i$
Forecast probability, y_i

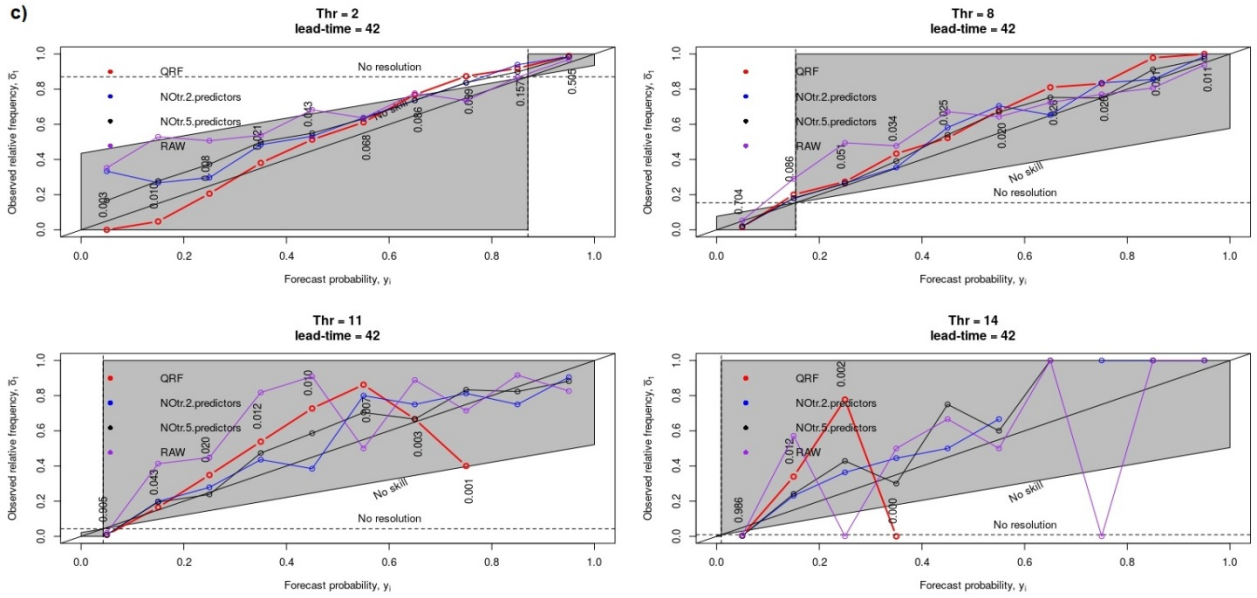


Figure 5: As Figure 4 but for summer.

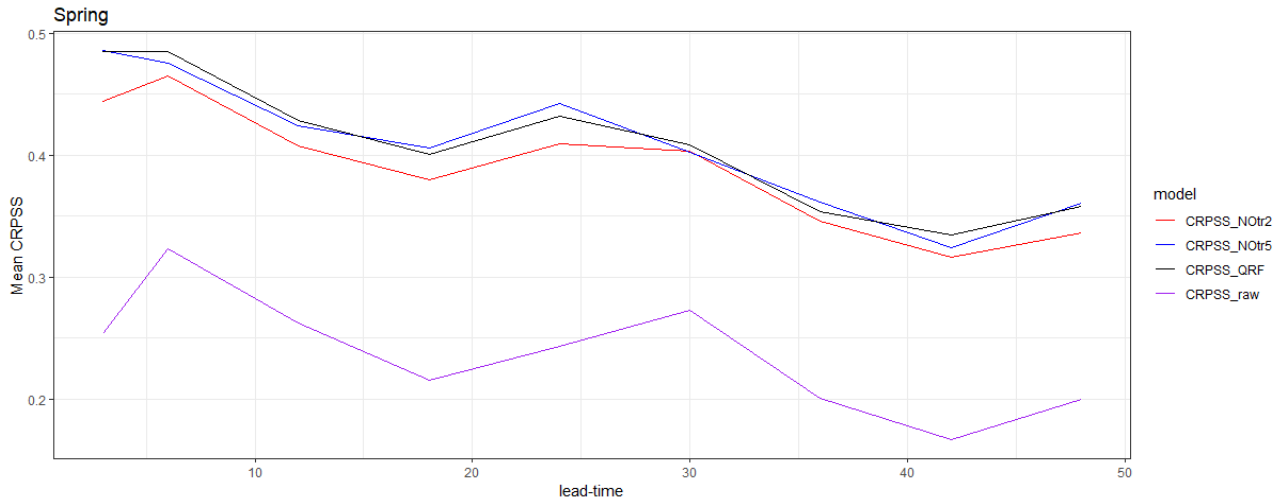


Figure 6: Mean CRPSS comparison between raw and calibrated forecasts, using NOtr_2, NOtr_5 and QRF for spring.

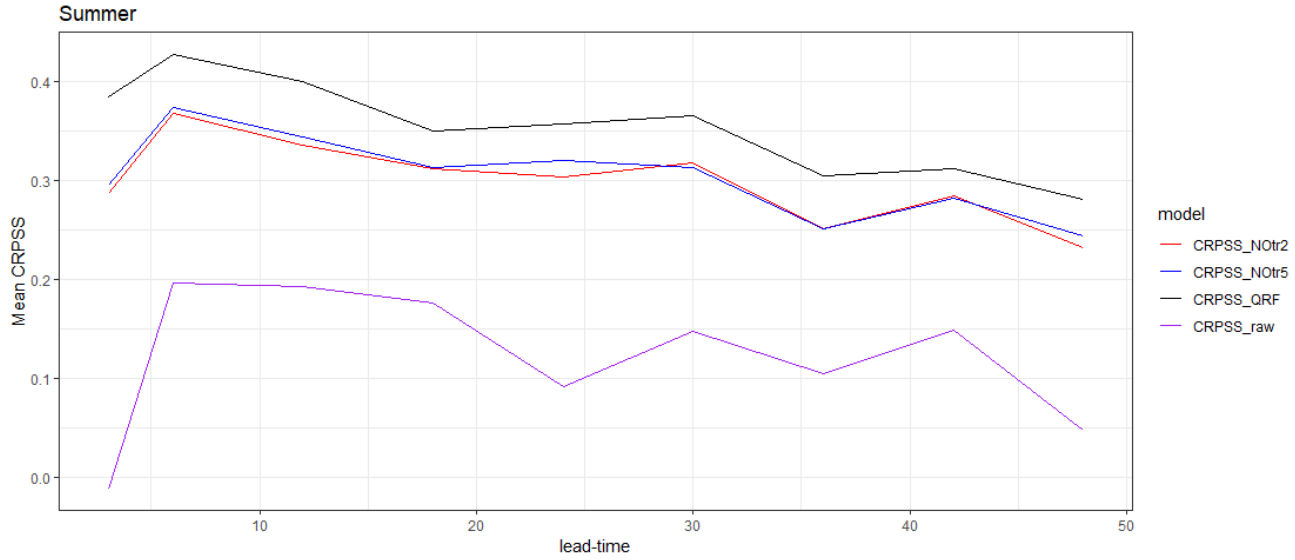


Figure 7: **As Figure 6 but for summer.**

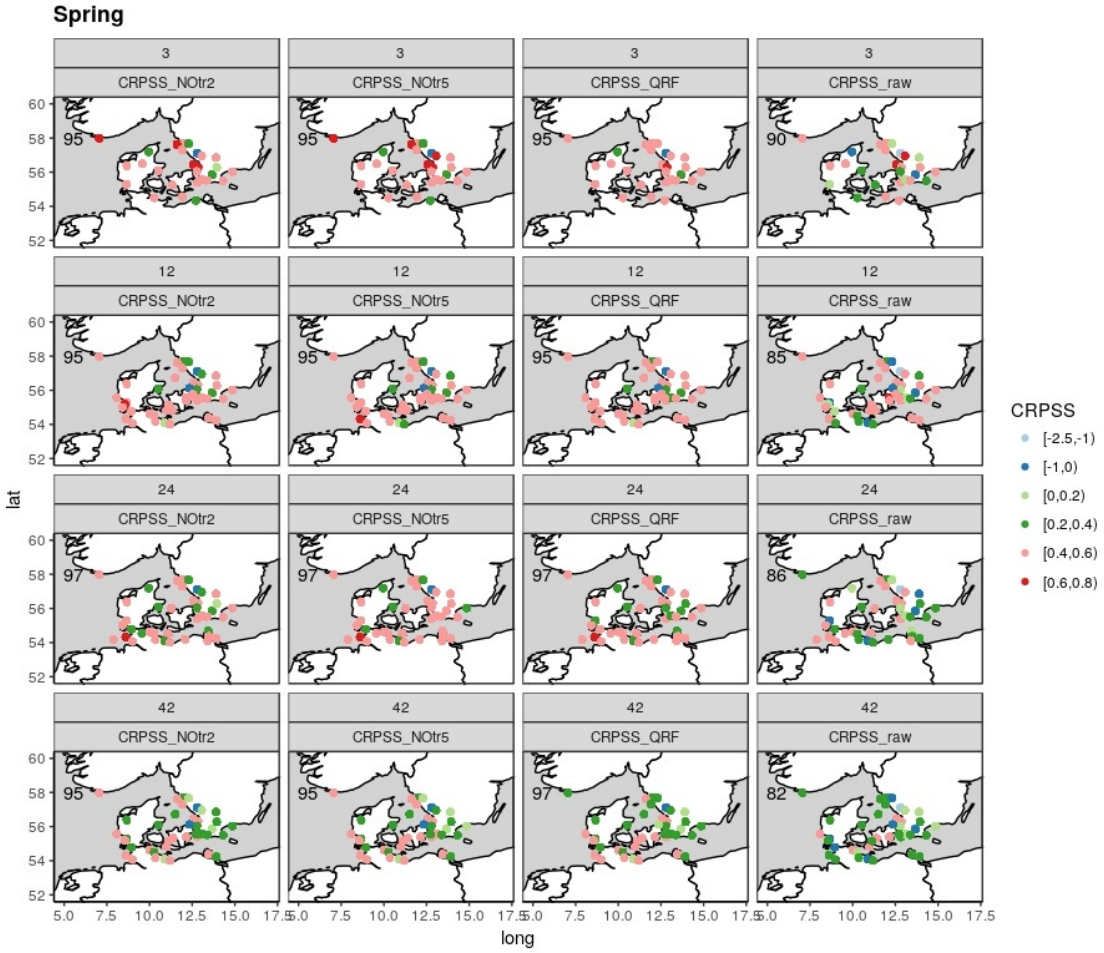


Figure 8: **Spatial comparison of CRPSS between NOtr_2, NOtr_5, QRF and raw forecasts, based on the seasonal station climatology for spring and specific lead-times. The number in the upper left corner of each map indicates the percentage of stations having positive CRPSS values.**

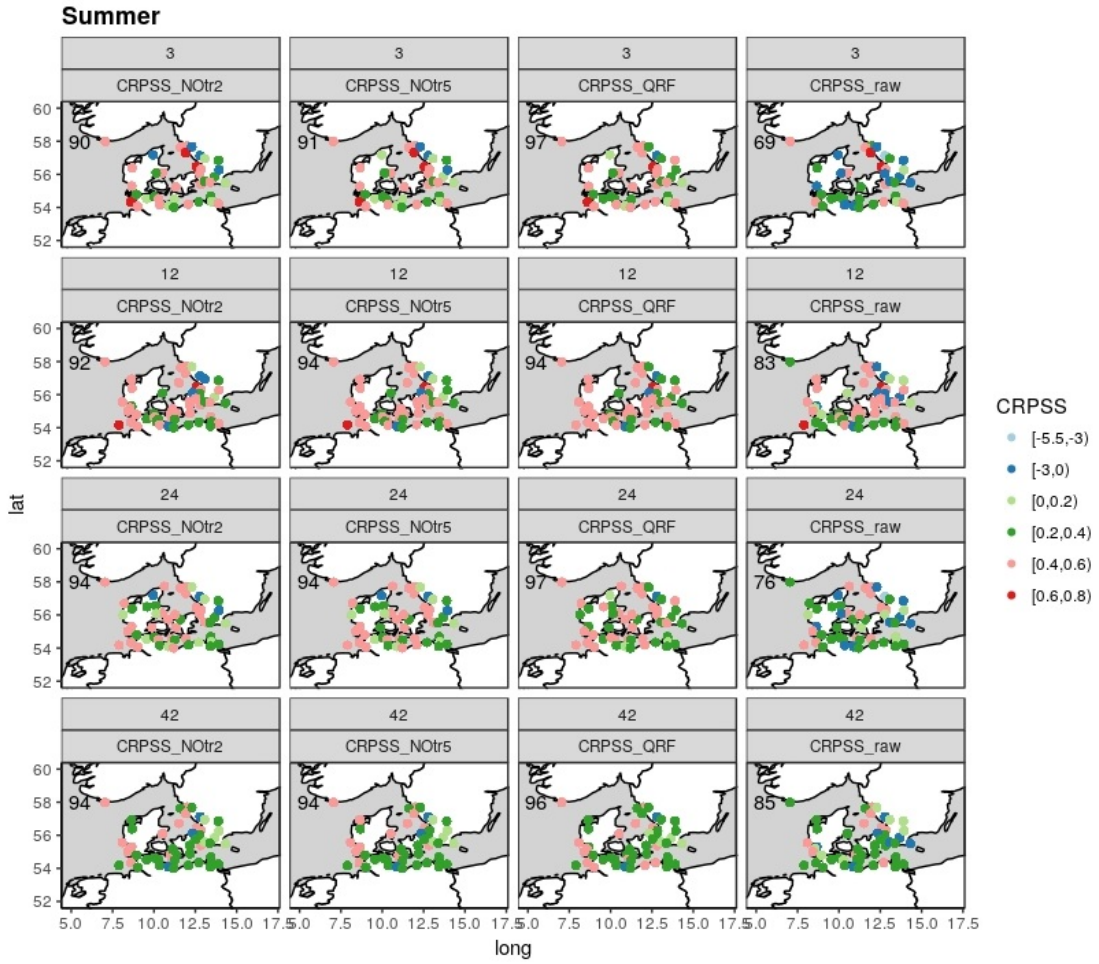


Figure 9: As Figure 8 but for summer.

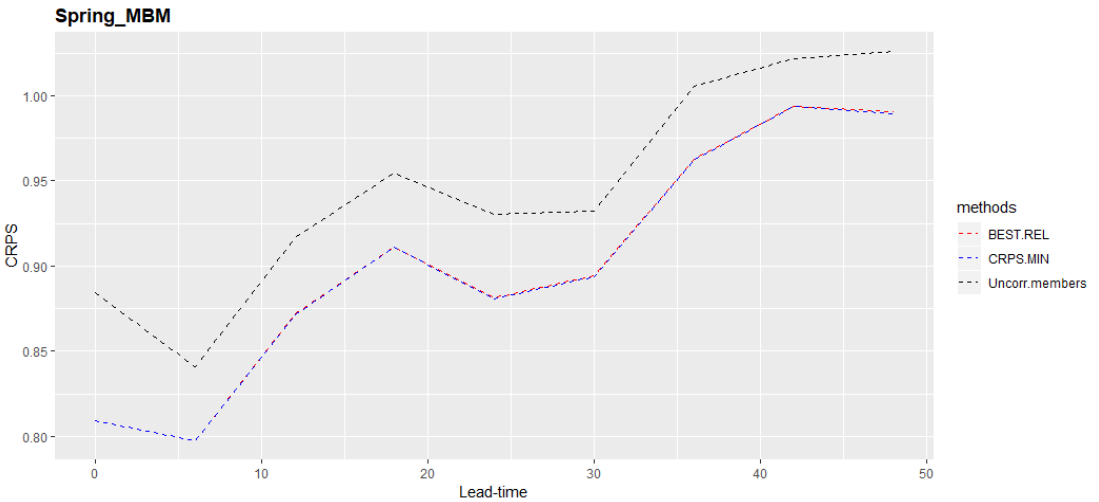


Figure 10: CRPS comparison between the uncorrected RAW data and the corrected data based on the two MBM approaches and wind gust as a second predictor for spring.

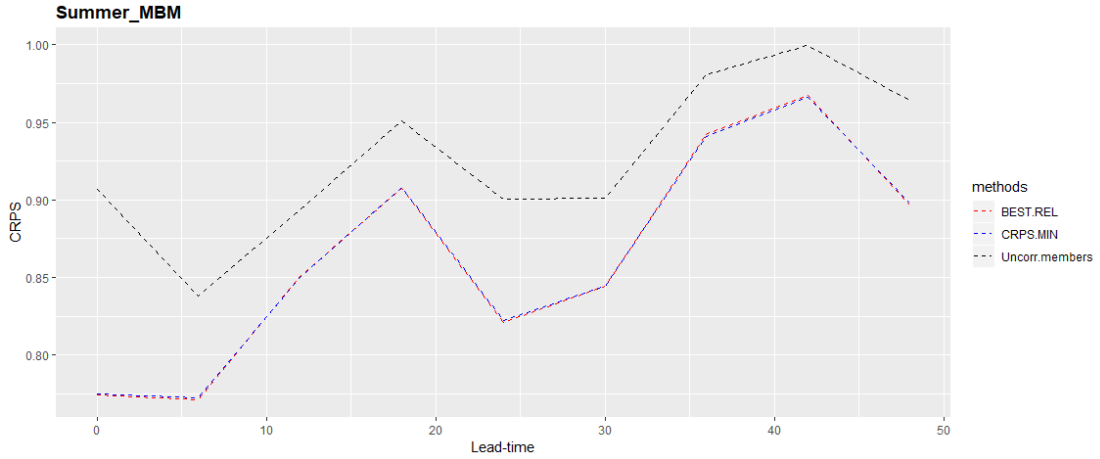
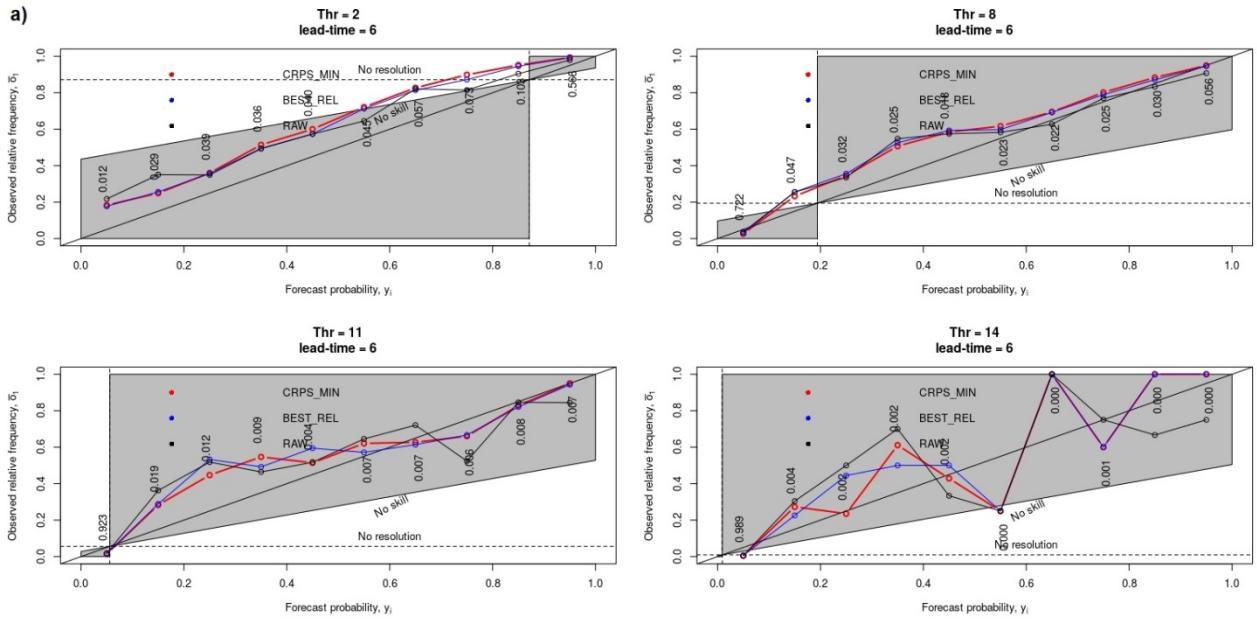


Figure 11: As Figure 10 but for summer.



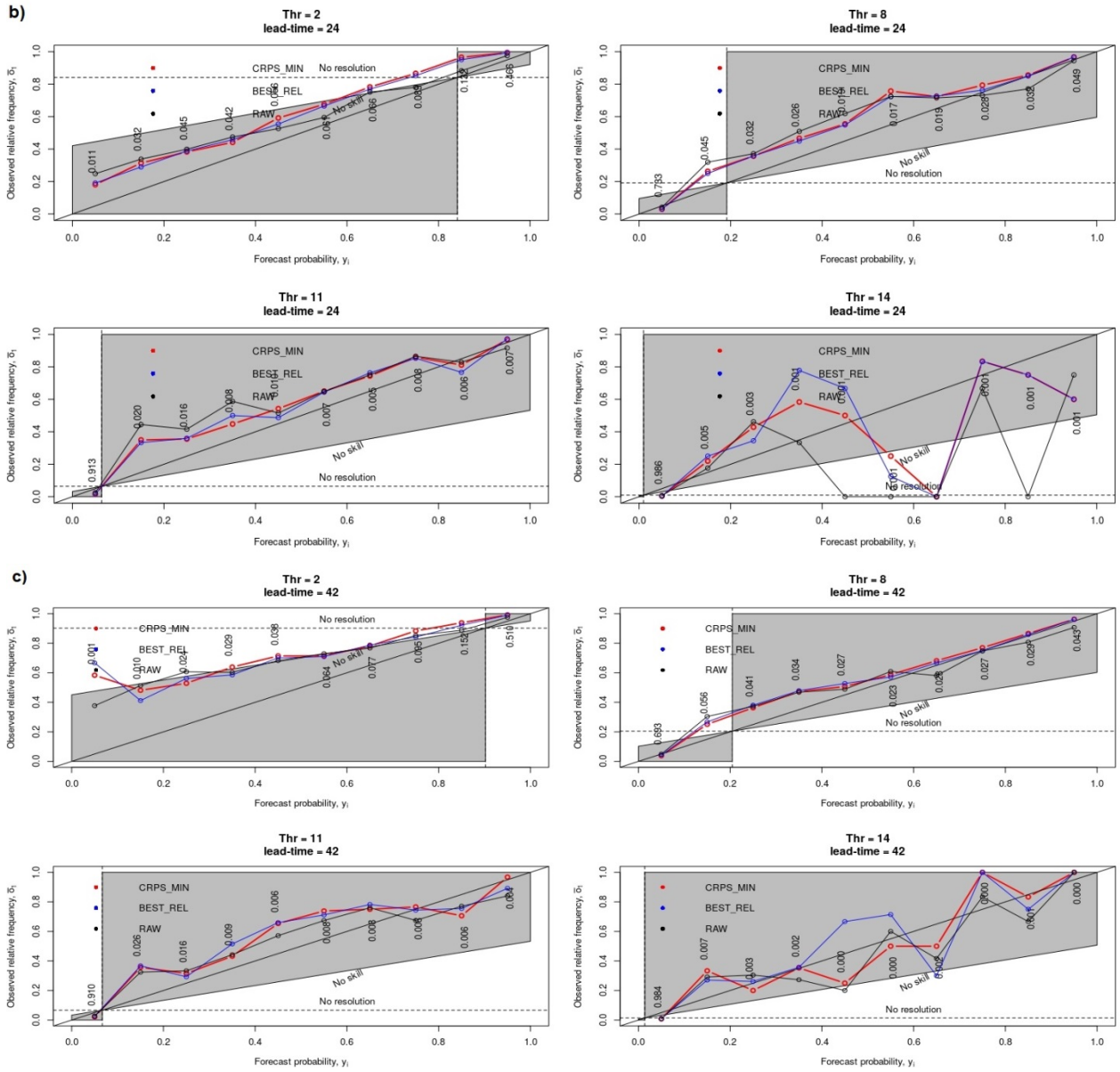
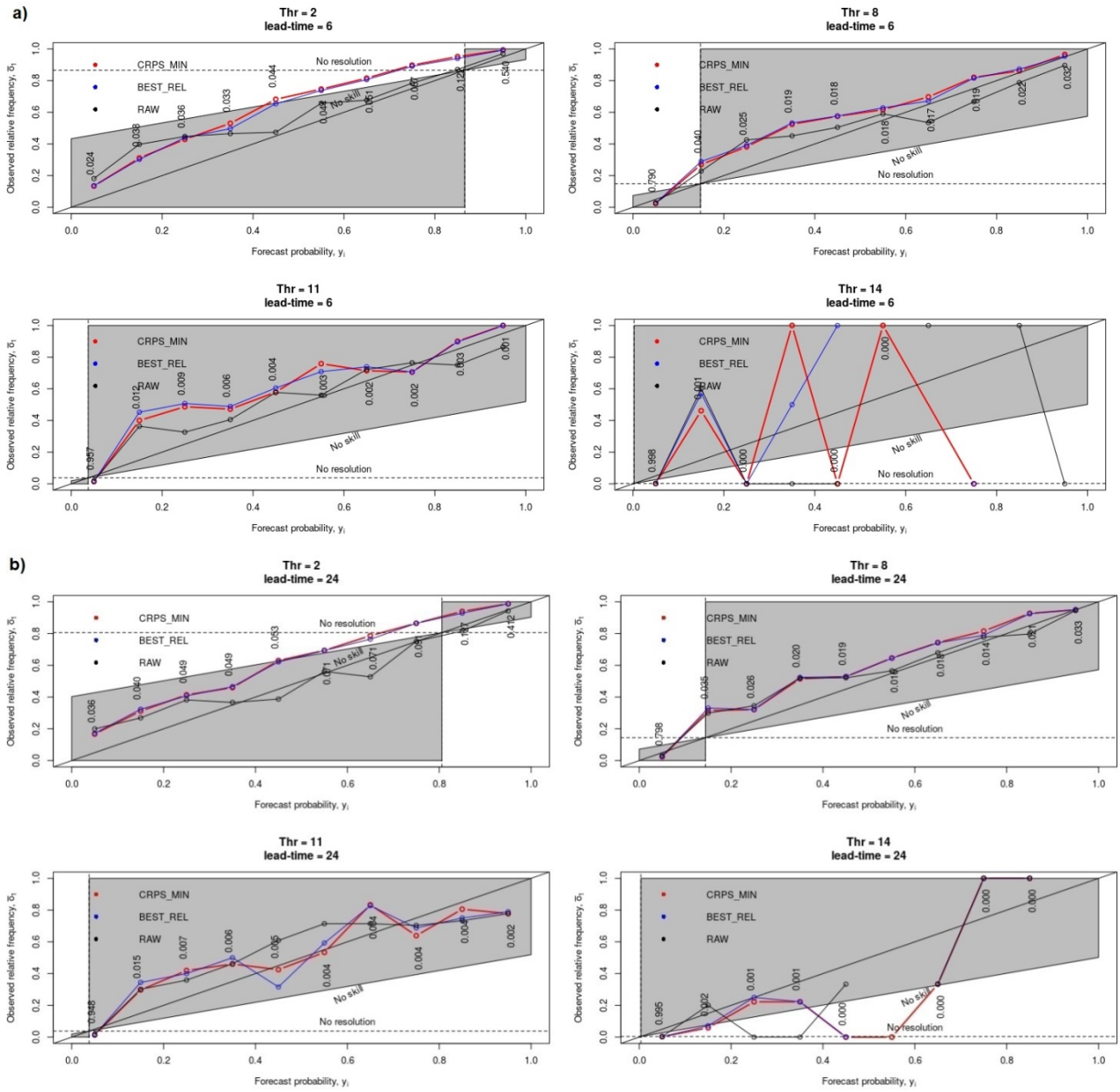


Figure 12: Reliability diagrams for wind speed greater than 2, 8, 11 and 14 m/s during the spring for the raw forecasts and the BEST REL and CRPS MIN MBM approaches for lead-times a) 6h, b) 24h and c) 42h.


a)
Thr = 2
lead-time = 6
Thr = 8
lead-time = 6
Thr = 11
lead-time = 6
Thr = 14
lead-time = 6
b)
Thr = 2
lead-time = 24
Thr = 8
lead-time = 24
Thr = 11
lead-time = 24
Thr = 14
lead-time = 24
CRPS_MIN
BEST_REL
RAW
No resolution
No skill
Observed relative frequency, $\bar{o}_i$
Forecast probability, y_i

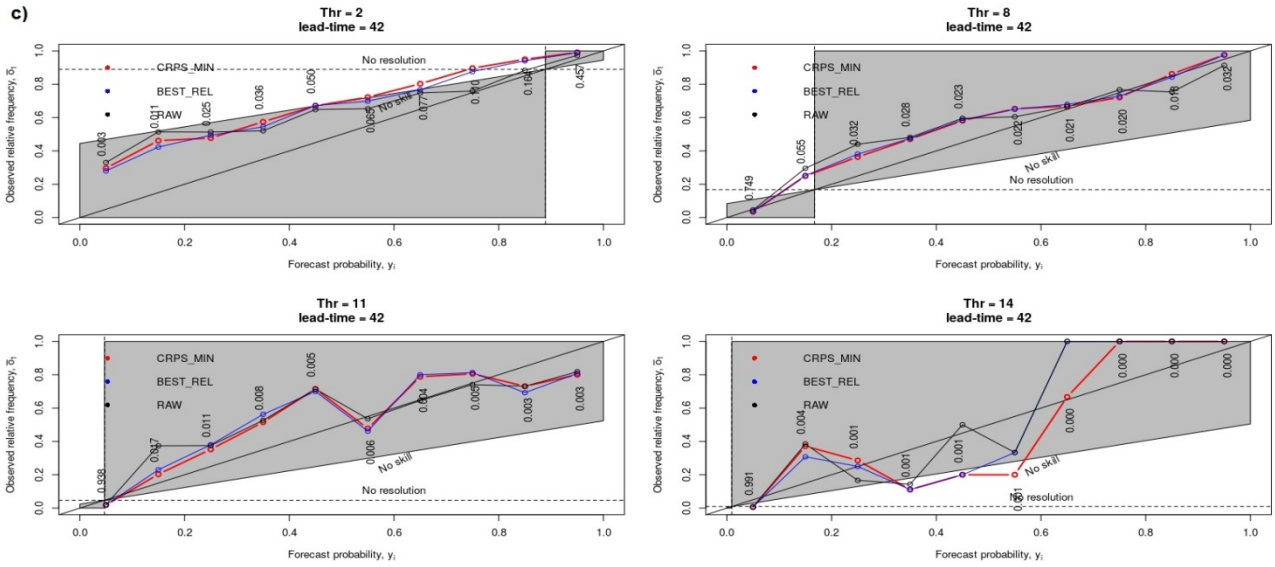


Figure 13: As Figure 12 but for summer.

Royal Netherlands Meteorological Institute

PO Box 201 | NL-3730 AE De Bilt
Netherlands | www.knmi.nl