NTNU

Norwegian University of
Science and Technology

Statistical Postprocessing of Ensemble Forecasts of Wind

Siri Sofie Eide

Master of Science in Physics and Mathematics

Submission date:	June 2016
Supervisor:	Ingelin Steinsland, MATH
Co-supervisor:	John Bjørnar Bremnes, Meteorologisk Institutt

Norwegian University of Science and Technology
Department of Mathematical Sciences

Abstract

This thesis considers methods and models for postprocessing ensemble forecasts of wind. Based on Bayesian model averaging (BMA), several different extensions are proposed and tested. Firstly, historical observations of wind speed are included in the model as forecasts, both as a climatology and as an ensemble. Secondly, an extension to the BMA in which thin plate regression splines over both forecast wind speed and forecast wind direction are used in the modelling of the expectation of the predictive probability density functions (PDFs) is tested. Each method is assessed mainly using the continuous rank probability score (CRPS), but certain aspects of the forecasts, such as their performance for stronger winds, are assessed using the Brier score and the quantile score. We identify a shortcoming of the BMA involving bias in the forecasting of stronger winds, and an amendment to the method is proposed. This extension is shown to produce better forecasts and goes a long way towards solving the problem with bias in the forecasts of stronger wind.

Sammendrag

Denne masteroppgaven undersøker metoder og modeller for postprosessering av ensemblevarsler for vind. Den tar utgangspunkt i Bayesian model averaging (BMA), men flere ulike utvidelser av metoden blir foreslått og testet. Først blir gamle observasjoner av vindhastighet inkludert i modellen, både i form av en klimatologi og som et ensemble. Deretter testes en forlengelse av BMA der thin plate regression splines over både varslet vindhastighet og varslet vindretning brukes i modelleringen av forventningen til sannsynlighetsvarselet. Hver metode evalueres hovedsakelig ved bruk av continuous rank probability score (CRPS), men visse aspekter, som deres evne til å varsle sterk vind, evalueres også ved bruk av Brier score og kvantilscore. Vi identifiserer et problem med BMA som har å gjøre med bias ved varsling av sterkere vind, og foreslår en endring i metoden. Det vises at denne endringen fører til bedre varsler og langt på vei løser problemet med bias ved varslingen av sterkere vind.

Preface

This master's thesis was written as part of the course *TMA4905 - Statistics, Master's Thesis* under the program of Industrial Mathematics in the Department of Mathematical Science of the Norwegian University of Science and Technology (NTNU). The work was carried out at the Norwegian Meteorological Institute (MET) in the spring of 2016.

I would like to thank my supervisors, Ingelin Steinsland at NTNU and John Bjørnar Bremnes at MET, for their invaluable guidance and unfaltering support. They have always been there, answering my questions and helping me sort through my thoughts, and without them there would be no master's thesis.

These months also would not have been the same without all the lovely people here at MET who have given me inspiration and motivation, shared puns and jokes about work-life, and who have kept me more or less sane, despite instilling in me an irrational fear of owl paraphernalia that will stay with me until the bitter end.

Siri Sofie Eide
Oslo, June 2016

Contents

1. Introduction

Knowing what weather we can expect to see hours, days or weeks into the future is always helpful in making informed decisions. Sometimes, this can mean knowing whether or not to bring an umbrella, what jacket to wear, or what type of ski wax to use. But for some this type of information is of greater importance, such as in civil protection, aviation, farming, construction work or professional sports where an unexpected change in the weather conditions can have massive consequences. Fog can cause the cancellation of a biathlon, freezing rain has been known to shut down entire cities and oil rigs are occasionally evacuated if particularly strong winds are expected.

If, for example you were an oil rig engineer and your job involved hanging off the side of an oil rig by a rope, and for safety reasons you were not allowed to do your job if wind speeds exceeded 13 m/s (BBC, 2005). If the weather forecast told you it would be 12 m/s tomorrow, would you go to work?

These are some examples of why it is essential that weather forecasts are precise and reliable. Traditionally, weather forecasting has been done in a deterministic way, deterministic meaning that everything that happens is viewed as the inevitable result of preceding events. Deterministic forecasting models are largely based on mathematical representations of the dynamics and physics of the atmosphere, and can often produce sufficient forecasts of the weather up to 2 weeks into the future (Kalnay, 2003).

However, a number of factors can result in forecast busts, whereby what is forecast bears little relation to what is observed. Sometimes we don't have complete knowledge of the initial conditions, or of the physical relationships we wish to model. And even when we do, numerical or human error can cause small inaccuracies that completely change the resulting forecast.

This is why probabilistic forecasting has in recent years grown increasingly popular. A probabilistic forecast gives an indication of the uncertainty of the future weather. If rather than saying that it *will* rain tomorrow, we say that there is an 80 % chance of rain tomorrow, we have accounted for the possibility that this might not happen, even though we believe quite strongly that it will.

The Norwegian website for weather forecasting *yr.no*[1], incorporates probability in their long term forecasts for cloud cover, temperature, precipitation and wind. Figure 1.1a shows how they use colored labels to indicate degrees of certainty. A legend below forecasts relays the degree of certainty associated with each label, green meaning rather certain, yellow meaning somewhat certain and red meaning uncertain. For those interested, the website also explains that forecasts labeled as green are usually correct at least 70 % of the time; yellow forecasts will consistently be correct somewhere between 50 % and 70 % of the time; while the forecasts labeled as red are normally correct less than 50 % of the time (Yr NRK, 2013).

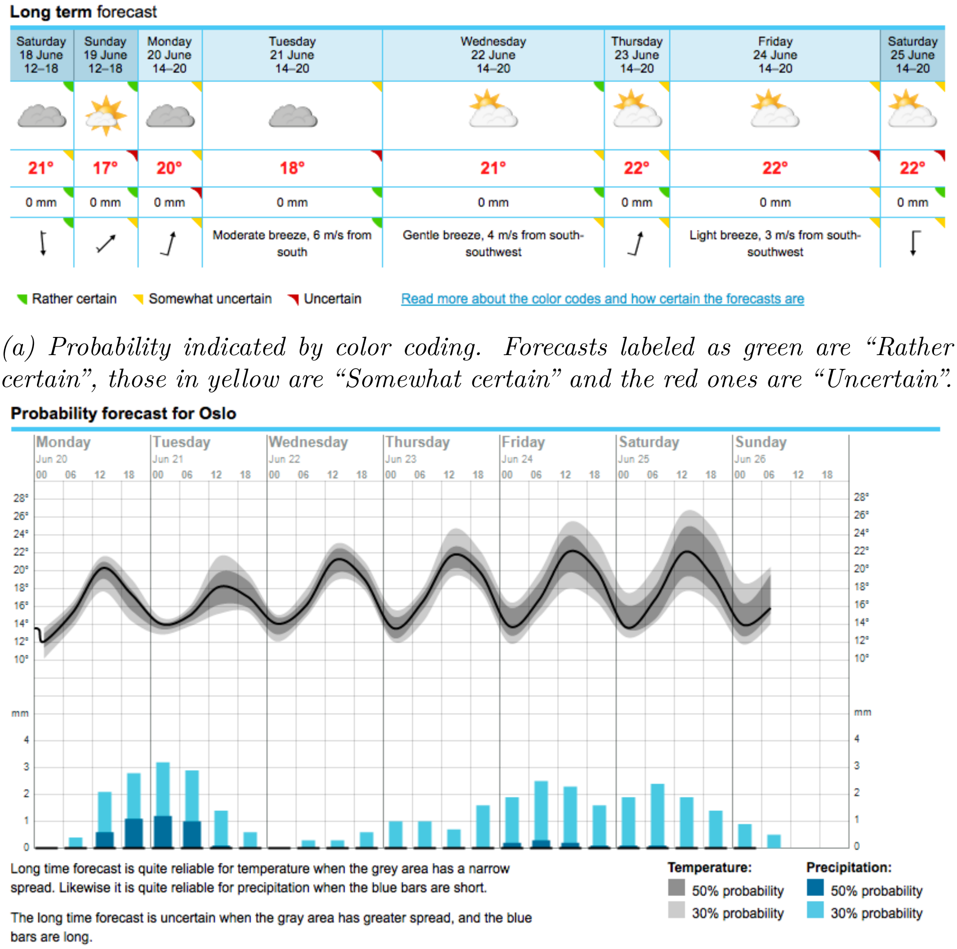
Figure 1.1b illustrates a different way of including uncertainty in the forecasts, with forecast intervals spanning 50 % and 80 % probability. There are many alternative ways of presenting probabilistic forecasts to a public that might not have any experience with probability or statistics. However, it remains a challenge and an area of ongoing research to find the best methods of communicating probabilistic forecasts in a concise and informative way.

There are many ways to compute probabilistic forecasts. One approach is to use ensembles. An ensemble consists of a number of members – unique forecasts created either by means of different models or with the same model using slightly different initial conditions (Kalnay, 2003). The models themselves are usually deterministic, but the resulting ensemble forecasts can be viewed as representative of an underlying probability density function, although postprocessing is often necessary, as the raw ensemble forecasts can be both biased and over- or underdispersed (Feldmann, 2012).

Bias in an ensemble means that it on average over- or underforecasts a particular parameter or parameters. If the tendency is to overforecast this is called a positive bias and underforecasting implies a negative bias. Dispersion has to do with the spread of the forecasts. If the observed quantity frequently lies outside the range of the ensemble forecasts this is a sign of underdispersion.

The work done by Leith (1974) is considered the start of ensembles as we know them, although they didn't gain much attention until the 1990s, when the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction started developing their own ensemble forecasting systems. The most basic forms of post-processing of ensemble forecasts include adding or multiplying by a constant to remove a systematic bias, and performing linear regression.

[1] A joint service by the Norwegian Meteorological Institute (MET) and the Norwegian Broadcasting Corporation (NRK).



Long term forecast

Saturday 18 June 12–18	Sunday 19 June 12–18	Monday 20 June 14–20	Tuesday 21 June 14–20	Wednesday 22 June 14–20	Thursday 23 June 14–20	Friday 24 June 14–20	Saturday 25 June 14–20
21°	17°	20°	18°	21°	22°	22°	22°
0 mm	0 mm	0 mm	0 mm	0 mm	0 mm	0 mm	0 mm
↓	↗	↑	Moderate breeze, 6 m/s from south	Gentle breeze, 4 m/s from south-southwest	↑	Light breeze, 3 m/s from south-southwest	↓

Rather certain   Somewhat uncertain   Uncertain   Read more about the color codes and how certain the forecasts are

(a) *Probability indicated by color coding. Forecasts labeled as green are "Rather certain", those in yellow are "Somewhat certain" and the red ones are "Uncertain".*

Probability forecast for Oslo

Long time forecast is quite reliable for temperature when the grey area has a narrow spread. Likewise it is quite reliable for precipitation when the blue bars are short.

The long time forecast is uncertain when the gray area has greater spread, and the blue bars are long.

Temperature:
- 50% probability
- 30% probability

Precipitation:
- 50% probability
- 30% probability

(b) *Probability indicated by 50 % and 80 % forecast intervals around the forecasts.*

Figure 1.1: Screenshots of long term forecasts for Oslo from yr.no issued on 17th June 2016, for the period 18th-25th June 2016, illustrating how a probabilistic forecast might be presented to the public.

Many years of research in this field have given us much more sophisticated methods, like ensemble model output statistics (EMOS) as described by, amongst others Thorarinsdottir and Gneiting (2010) and Baran and Lerch (2015a), Bayesian model averaging, developed by Raftery et al. (2005) and non-homogeneous regression (Gneiting et al., 2005) and (Thorarinsdottir and Johnson, 2012).

In this thesis I evaluate and propose postprocessing techniques for ensemble forecasts of wind speed and direction, based on Bayesian model averaging. The forecasts come from ECMWF. ECMWF is an international meteorological organization founded in 1975 and based in Reading, England. Norway is one of its 21 European member states. Validating wind speed observations were provided by the Norwegian Meteorological Institute (MET).

The rest of this thesis is organized as follows: Chapter 2 gives an introduction to forecasting, the postprocessing methods considered and proposed, and the means of evaluation used to assess the forecasts. In Chapter 3 the data used in the study are presented and explored. Chapter 4 gives a closer look at how the methods introduced in Chapter 2 are applied to the data. Assessment of the different methods and other findings are presented in Chapter 5. Chapter 6 ends the thesis with a discussion of the results obtained and possible future work.

2. Background

This chapter contains a brief introduction to the most important properties of probabilistic forecasts and to Bayesian model averaging. Two graphical tools that are useful in evaluating forecasts, namely the verification rank histogram and the pit histogram are presented, as well as three of the most commonly used scoring rules: the continuous ranked probability score, the quantile score and the Brier score. The concept of skill scores is also introduced.

2.1 Forecasts and probability

The most commonly used type of weather forecast is based on Numerical Weather Prediction (NWP). In NWP observations of the current state of the atmosphere are fed into computers, which in turn use mathematical models to forecast the future weather (NOAA-NCEI , U.S. Dept. of Commerce). In order to get good forecasts it is crucial that the observations that make up the initial conditions for the models are correct.

Errors in the initial conditions lead to errors in the forecasts, and tiny errors are impossible to avoid. The forecasts are therefore never perfect, and the discrepancies between forecasts and observations tend to become larger as the time from when the forecasts are initialized to the time for which they apply increases (Wilks, 2006).

This brings us to the concept of lead time. A lead time is exactly this, the interval between the moment at which the forecast is created and the time for which it applies. It is generally easier to make specific forecasts for shorter lead times (what will the temperature be in 2 hours, and will it be raining?) than for longer lead times (what will the temperature be at 4 o'clock 10 days from now, and will it be raining?).

For extremely long lead times it becomes impossible to make specific forecasts based on mathematical models, and the safest thing is to use climatology as a forecast. A climatology is essentially what is "normal", or the "average" weather over a certain period. The length of this period must be chosen in a way that makes sense according to the context in which the climatology is to be used.

While the climatology tends to be the best forecast when lead times are long, the best forecast for extremely short lead times can often be a persistence forecast. A persistence forecast is simply using the current weather as the forecast. The idea is that no big changes are expected to occur in the immediate future. Neither the climatology nor the persistence forecast is based on mathematical models, but rather on observations.

The forecasts considered in this thesis are probabilistic, but what is the difference between a deterministic and a probabilistic forecast? Where deterministic forecasts tell us exactly what the weather will be like in the future, a probabilistic forecast gives us either a single probability or a probability density function (PDF) related to the event we are forecasting. The probability or PDF $p^l_{s,t}(y)$ is assigned to the event y occurring at site s at time t, with lead time l, i.e. l hours in advance. In postprocessing of ensembles, $p^l_{s,t}(y)$ is a function of the ensemble forecasts, $f^{l,i}_{s,t}$, $i = 1, \cdots, M$ where M is the number of ensemble members.

2.2 Bayesian Model Averaging (BMA)

The application of Bayesian model averaging to ensemble forecasts was proposed by Raftery et al. (2005) and is a method used to generate calibrated and sharp predictive probability density functions (PDFs) from ensemble forecasts.

A PDF being calibrated means that there is statistical consistency between the predictive distributions and the validating observations (Baran and Lerch, 2015b) or in other words that it is reasonable to believe that the validating observations could have been drawn from the predictive PDFs (Gneiting, 2014). Raw ensemble forecasts often suffer from bias and underdispersion, making them uncalibrated.

Sharpness is a measure of the concentration of the PDF (Gneiting, 2014), and is independent on the validating observations. The sharper the predictive distribution, the higher is the certainty with which we can forecast, as long as the distribution is calibrated.

BMA aims to estimate the predictive probability density function of the weather quantity Y based on an ensemble forecast $f_1, ..., f_M$, where f_m is the forecast of ensemble member m.

We assume that each forecast f_m corresponds to a component PDF $g_m(y|f_m; \theta_m)$, where θ_m are parameters to be estimated. Further, we express the predictive PDF of the weather quantity Y as the sum of the component

PDFs associated with each ensemble member,

$$p(y|f_1, ..., f_M; \theta_1, ..., \theta_M) = \sum_{m=1}^{M} w_m g_m(y|f_m; \theta_m). \tag{2.1}$$

Here w_m are weights based on the predictive performance of forecast f_m, with $\sum w_m = 1$. If some of the ensemble members are exchangeable, they are given equal weight, and a single set of corresponding parameters. This is usually the case when the only difference between one ensemble member and the next is a small, random perturbation of the initial conditions in the model.

So how is the distribution of the component PDFs chosen? There is no single correct answer to this question. The choice is based on the physical properties of the quantity Y and on the data at hand. Raftery et al. (2005) studied surface temperatures, and proposed the use of normal distributions for this type of data. For precipitation, Sloughter et al. (2007) showed that a mixture of a discrete component at zero and a gamma distribution could be used. For wind speed, a strictly positive quantity, various distributions have been applied. Previously many have favored the Weibull distribution, as discussed by Tuller and Brett (1984), but in connection with BMA, the gamma distribution has more commonly been used, e.g. by Sloughter et al. (2010).

As this thesis considers wind speed, i.e. a quantity that can not take negative values, a gamma distribution will be assumed for all predictive PDFs. The PDF of the gamma distribution with shape parameter α and scale parameter β is

$$g(y) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} y^{\alpha-1} \exp(-y/\beta) \tag{2.2}$$

when y is non-negative and $g(y) = 0$ otherwise. Here $\Gamma(\alpha)$ is the gamma function evaluated in α. Conditioning on the forecasts we get the following expression for the component PDFs

$$g_m(y|f_m) = \frac{1}{\beta_m^{\alpha_m}\Gamma(\alpha_m)} y^{\alpha_m-1} \exp(-y/\beta_m) \tag{2.3}$$

where α_m and β_m are parameters that need to be estimated. The gamma

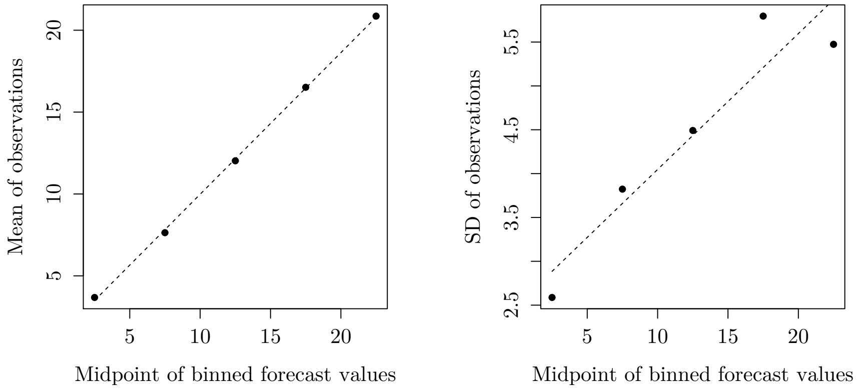


Figure 2.1: Forecast values divided into bins, and means and standard deviations for gamma distribution fits to the observed wind speeds, conditional on the forecast being within a certain bin, as done by Sloughter et al. (2010).

distribution has mean $\mu = \alpha\beta$ and variance $\sigma^2 = \alpha\beta^2$. Assuming a linear relationship between the ensemble forecasts and the observations, the mean and standard deviation of each component distribution can be expressed as

$$\mu_m = b_{0m} + b_{1m} f_m \tag{2.4}$$

$$\sigma_m = c_{0m} + c_{1m} f_m \tag{2.5}$$

and α_m and β_m are easily calculated. Following the method of Sloughter et al. (2010), the assumption of linearity is examined by dividing the forecast values into bins and plotting forecast values, represented by the midpoint of each bin against the mean and standard deviation for a gamma distribution fit to the observed wind speeds, conditional on the forecast being within that bin. These plots are presented in Figure 2.1. The assumption seems to hold for the mean. For the standard deviation, however, it is perhaps a little less convincing, but the assumption does not seem completely unreasonable.

The common way to estimate b_{0m} and b_{1m} is through linear regression. To facilitate estimation the parameters c_{0m} and c_{1m} are taken to be the same for all ensemble members, and are replaced by c_0 and c_1. These parameters and the weights w are estimated through a variant of the expectation–maximization (EM) algorithm. Further details about the method, and examples of its use can be found in (Sloughter et al., 2010) and (Fraley et al., 2010).

2.3 Thin plate regression splines

In the standard BMA discussed in the previous section, the expectation is modelled as a linear function of forecast speed. In this thesis, an extension to BMA is proposed in which two-dimensional thin plate regression splines, are used to model the potentially more complex relationship between the observed wind speed and forecast wind speed and direction. Therefore, a short introduction to thin plate splines and thin plate regression splines is given in this Section.

2.3.1 Thin plate splines

Wood (2003) has shown that the thin plate regression spline is by certain definitions optimal. Thin plate regression splines are based on thin plate splines, a concept introduced by Duchon (1977). They can be thought of as the two-dimensional analogue of the cubic spline in one dimension (Belongie, 2000).

The name "thin plate spline" refers to the spline's likeness to a thin metal sheet, which can be bent but also has a certain rigidity. In the mathematical sense, this rigidity corresponds to a tuning parameter that controls the smoothness of the spline.

Say you want to estimate a smooth function $g(\mathbf{x})$ from observations of response variable y_i and vectors of covariates $\mathbf{x}_i$, $(i = 1, ..., n)$, such that

$$y_i = g(\mathbf{x}_i) + \epsilon_i \tag{2.6}$$

where every vector $\mathbf{x}_i$ is of length d $(d \leq n)$ and ϵ_i is a random error term. The thin plate spline smoothing estimate $\hat{f}$ of g is the function that minimises the penalized least squares function

$$||\mathbf{y} - \mathbf{f}||^2 + \lambda J_{md}(f) \tag{2.7}$$

where $\mathbf{y}$ is the vector of y_i, $f = (f(x_1), f(x_2), ..., f(x_n))^T$ and $|| \cdot ||$ is the Euclidean norm. The penalty function $J_{md}(f)$ measures the "wiggliness" of

f. Here m is the order of differentiation and can be any integer satisfying $2m > d$, although it is often chosen such that $2m > d + 1$, as this gives results that are more "visually smooth".

The smoothing parameter $\lambda \in [0, \infty)$ controls the tradeoff between goodness of fit and smoothness of f. The wiggliness penalty is defined for any number d of predictor variables and order of differentiation satisfying the condition $2m > d$ as

$$J_{md} = \int \cdots \int_{\Re^d} \sum_{\nu_1+\cdots+\nu_d=m} \frac{m!}{\nu_1! \cdots \nu_d!} \left(\frac{\partial^m f}{\partial x_1^{\nu_1} \cdots \partial x_d^{\nu_d}} \right)^2 dx_1 \cdots dx_d. \quad (2.8)$$

In 2 dimensions using second derivatives (i.e. $d = 2$ and $m = 2$), which will be used in this thesis, Equation (2.8) becomes

$$J_{22} = \int \int \left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 dx_1 dx_2. \quad (2.9)$$

It can be shown that the function $\hat{f}(\mathbf{x})$ minimizing (2.7) can be expressed in the form

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \delta_i \eta_{md}(||\mathbf{x} - \mathbf{x}_i||) + \sum_{j=1}^{M} \alpha_j \phi_j(\mathbf{x}), \quad (2.10)$$

where $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ are unknown parameter vectors that need to be estimated. The vector $\boldsymbol{\delta}$ is also subject to the constraint $\mathbf{T}'\boldsymbol{\delta} = \mathbf{0}$ where $T_{ij} = \phi_j(\mathbf{x}_i)$. The functions ϕ_j are unpenalized, linearly independent polynomials of degree less than m. In total there are $M = \binom{m+d-1}{d}$ of these functions. The ϕ_j span the space of functions for which J_{md} is zero, meaning the functions that are considered "completely smooth".

The remaining basis function used in (2.10), η_{md}, is a function of the Euclidean distance r between any two $\mathbf{x}$ and $\mathbf{x}_i$, and is defined as

$$\eta_{md}(r) = \begin{cases} \frac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d} \log(r) & d \text{ even}, \\ \\ \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} r^{2m-d} & d \text{ odd}. \end{cases} \quad (2.11)$$

Defining a penalty matrix $\mathbf{E}$ such that $E_{ij} \equiv \eta_{md}(||\mathbf{x}_i - \mathbf{x}_j||)$, the minimization problem can be rewritten as

$$\text{minimise } ||\mathbf{y} - \mathbf{E}\boldsymbol{\delta} - \mathbf{T}\boldsymbol{\alpha}||^2 + \lambda\boldsymbol{\delta}'\mathbf{E}\boldsymbol{\delta} \text{ subject to } \mathbf{T}'\boldsymbol{\delta} = \mathbf{0} \qquad (2.12)$$

with respect to $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$.

2.3.2 Thin plate regression splines

The thin plate spline is in many ways optimal, the only problem is that it comes with a high computational cost, as the number of unknown parameters is the same as the number of unique predictor combinations. This is the problem thin plate regression splines seek to solve. The computational cost of fitting thin plate splines with n parameters is $\mathcal{O}(n^3)$. This can be drastically reduced by replacing the matrix $\mathbf{E}$ above with an eigen approximation, $\mathbf{E}_k$ of rank k $(k > M)$.

This is done by first decomposing $\mathbf{E}$ into $\mathbf{E} = \mathbf{UDU}'$, where $\mathbf{D}$ is a diagonal matrix of the eigenvalues of $\mathbf{E}$, arranged so that they are weakly decreasing in absolute value, in other words, $|D_{i,i}| \geq |D_{i+1,i+1}|$, and the columns of $\mathbf{U}$ are the corresponding eigenvectors.

The eigen approximation $\mathbf{E}_k$ can now be written as $\mathbf{E}_k = \mathbf{U}_k\mathbf{D}_k\mathbf{U}_k'$, where $\mathbf{D}_k$ denotes the top left $k \times k$ submatrix of $\mathbf{D}$, and $\mathbf{U}_k$ is made up by columns of the eigenvectors corresponding to the eigenvalues in $\mathbf{D}_k$, i.e. the first k columns of $\mathbf{U}$.

By writing $\boldsymbol{\delta} = \mathbf{U}_k\boldsymbol{\delta}$, $\boldsymbol{\delta}$ is restricted to the column space of $\mathbf{U}_k$ and Equation (2.12) turns into

$$\text{minimise } ||\mathbf{y} - \mathbf{U}_k\mathbf{D}_k\boldsymbol{\delta}_k - \mathbf{T}\boldsymbol{\alpha}||^2 + \lambda\boldsymbol{\delta}_k'\mathbf{D}_k\boldsymbol{\delta}_k \text{ subject to } \mathbf{T}'\mathrm{U}_k\boldsymbol{\delta}_k = \mathbf{0} \qquad (2.13)$$

with respect to $\boldsymbol{\delta}_k$ and $\boldsymbol{\alpha}$ (Wood, 2006). By using $\mathbf{E}_k$ instead of $\mathbf{E}$ the dimension is reduced from $n \times n$ to $n \times k$.

This introduction to thin plate splines and thin plate regression splines is based largely on (Wood, 2003), (Wood, 2006) and (SAS Institute Inc., 2015). For a fuller, more in-depth explanation of thin plate regression splines, which would lie outside the scope of this thesis, the reader is encouraged to consult either of the aforementioned articles, or (Wahba, 1990).

Methods for solving the minimization problem of Equation (2.13) will not be shown in this thesis, but if the reader is interested they can be found

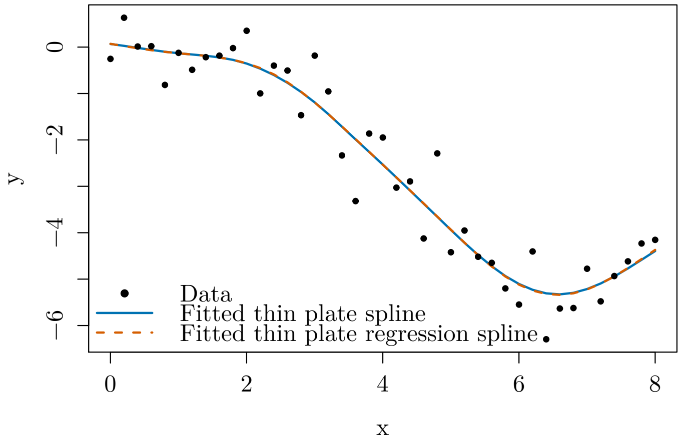


Figure 2.2: An example of a thin plate spline (blue line) and a thin plate regression spline (orange, dashed line) fitted to data.

in (Wood, 2006), where it is eventually shown that the computational cost of fitting a thin plate regression spline can be reduced to $\mathcal{O}(n^2k)$.

Figure 2.2 is an illustration of what it can look like when thin plate splines and thin plate regression splines are fitted to data. The data in this example were generated from

$$y_i = \frac{1}{500}(4x_i^4 - 36x_i^3 + x_i^2) + \epsilon_i \tag{2.14}$$

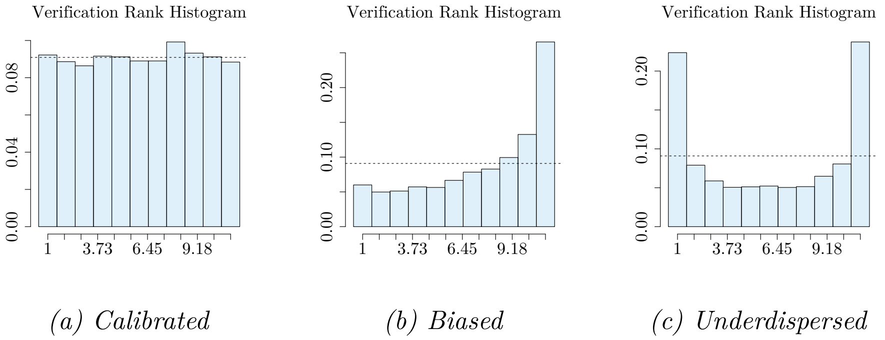
where $x_i = 0, 0.2, ..., 7.8, 8$ and $\epsilon_i \sim N(0, 0.2)$. The difference between the thin plate spline and the thin plate regression spline, which is much more computationally efficient, is hardly noticeable.

2.4 Validation

2.4.1 Verification Rank Histogram

The verification rank histogram, or Talagrand diagram (Talagrand et al., 1997), is a tool used to assess the calibration of an ensemble forecast. It is made by first recording the rank of every observation relative to its corresponding forecasts, i.e. if the observation has a value that is lower than every forecast, the observation has rank 1, if it is greater than 1 of the forecasts it has rank 2, and so on, so that if it is greater than all M forecasts, it has rank $M + 1$. This collection of ranks is then plotted in a histogram.

If the forecasts are representative of the true PDF of Y, the rank of observation y_m has a discrete uniform distribution, which corresponds to a histogram that is almost completely flat. Figure 2.3 shows a verification rank histogram in which the ensemble is calibrated, as well as histograms illustrating two of the most common problems in ensemble forecasting: bias and underdispersion.



(a) Calibrated *(b) Biased* *(c) Underdispersed*

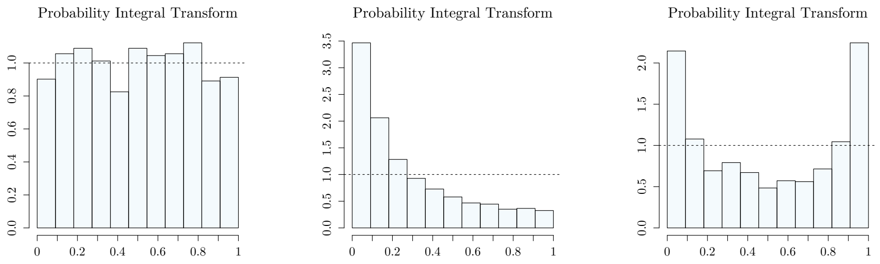
Figure 2.3: Three examples of verification rank histograms that are (a) calibrated, (b) biased towards underestimation (c) underdispersed, meaning that the observation is too often outside of the ensemble range.

2.4.2 PIT Histogram

The PIT histogram, or probability integral transform histogram, is also used to assess the calibration of forecasts. In many ways it is similar to the verification rank histogram. However, the PIT histogram is calculated based on continuous probability functions and not single value forecasts like those of the raw ensemble. It is essentially a histogram of the fitted CDF of each forecast evaluated in its validating observation.

Figure 2.4 is the PIT histogram equivalent of Figure 2.3, showing PIT histograms of forecasts that are calibrated, biased and underdispersed.

Like the verification rank histogram, the optimal PIT histogram is one that is flat. It is however possible to achieve flat PIT histograms with uncalibrated forecasts, as shown by Hamill (2000). A flat PIT histogram is therefore necessary but not sufficient in showing that the forecasts are calibrated.



(a) Calibrated *(b) Biased* *(c) Underdispersed*

Figure 2.4: Three examples of PIT histograms from forecasts that are (a) calibrated, (b) biased towards overestimation (c) underdispersed.

2.4.3 CRPS

Using the notation of Grimit et al. (2007), the continuous rank probability score, or CRPS, is defined as

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} (F(y) - \mathbb{1}\{y \geq x\})^2 dy \tag{2.15}$$

where F is the cumulative distribution function of the forecast associated with the validating observation x. The CRPS measures the difference between the CDF of the forecast and that of the observation, which takes the form of a step function. This is illustrated in Figure 2.5 by the shaded region between the two CDFs. The score is negatively oriented, i.e. the lower a score, the better, with 0 being the very best. The CRPS is expressed in the same unit as the observed variable (Grimit et al., 2007). For deterministic forecasts it reduces to the mean absolute error.

Clearly a forecast PDF is rewarded for sharpness, as the shape of its CDF is closer to that of the step function if it is sharp. It is also evident that a forecast is punished for being uncalibrated, as this increases the distance between the two graphs.

It has been shown that the CRPS is equivalent to the integral of the quantile score (QS) over all probability levels, or of the Brier score (BS) over all thresholds (e.g. by Gneiting and Ranjan (2011)). Both of these scoring rules will be introduced in the following subsections.

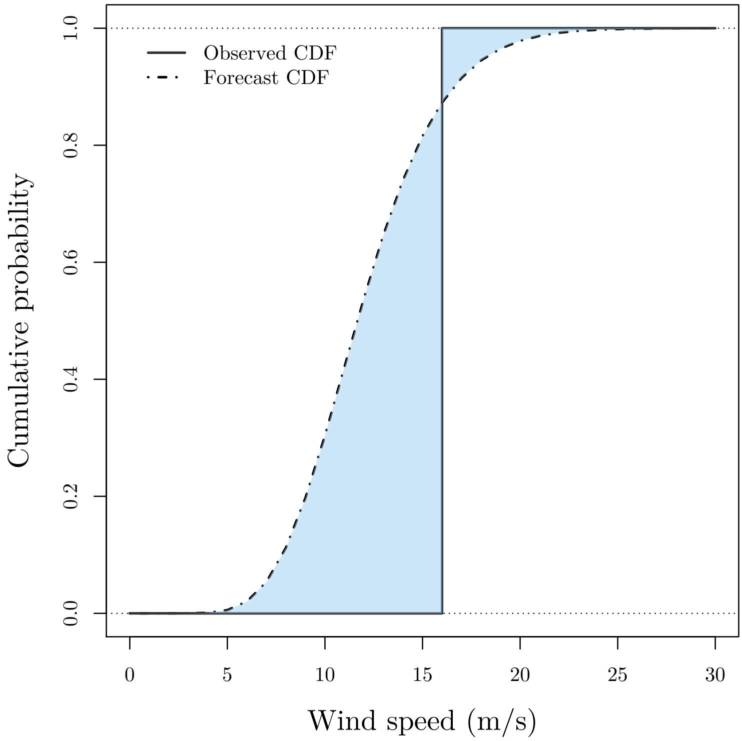


Figure 2.5: An example of the fitted CDF of a forecast (dashed curve) and a the CDF of the corresponding observed wind speed (step function).

2.4.4 Quantile Score

The quantile score is a scoring rule used to evaluate quantile forecasts, forecasts of quantiles of a predictive probability distribution. Quantile forecasts can be used for forecast intervals, e.g. a 90 % probability of observing a value between $q_{0.05}$ and $q_{0.95}$, or for giving some sort of upper or lower limits, e.g. 95 % chance of a value of observing a value lower than $q_{0.95}$.

Given N observations y_n and quantile forecasts $q_{\tau,n}$ for the τ-quantile, $\tau \in [0,1]$, the average quantile score (QS) is defined (e.g. (Bentzien and Friederichs, 2014b)) as

$$QS = \frac{1}{N} \sum_{n=1}^{N} \rho_\mathcal{T}(y_n - q_{\mathcal{T},n}), \tag{2.16}$$

where $\rho_\mathcal{T}$ is the check loss function defined as

$$\rho_\mathcal{T}(v) = \begin{cases} \tau|v| & \text{if } v \geq 0, \\ (1-\tau)|v| & \text{if } v < 0. \end{cases} \tag{2.17}$$

In other words, the absolute value of the difference between the observation and the quantile forecast is multiplied either by τ or by $1-\tau$ depending on the sign of the difference, thus over-forecasting and under-forcasting is penalized asymmetrically. The perfect QS is a QS of 0 (Bouallègue et al., 2015). For a more thorough explanation of the quantile score, see the aforementioned references or Bentzien and Friederichs (2014a).

2.4.5 Brier Score

The Brier score is a measure of the forecast error in probability space, and is defined as

$$BS = \frac{1}{N} \sum_{n=1}^{N} (y_n - p_n)^2 \tag{2.18}$$

where N is the total number of observations, y_n is the value of the n'th observation and p_n is the predicted probability associated with that observation (Wilks, 2006, p. 284). The BS can be used with categorical quantities or with continuous quantities by looking at them as binary with respect to some threshold.

A BS of 0 indicates a perfect model, and a BS of 1 means that the model is wrong 100 % of the time. These are the extreme cases. In general, the BS should be as close to 0 as possible.

2.4.6 Skill Scores

In some cases, when looking at a score in isolation, it can be hard to judge whether it is “good” or “bad”. For instance, the best CRPS one can achieve is 0, but what is a bad CRPS? What is the value above which we classify CRPS as bad?

With such a score it is much more informative to look at it in comparison to something else. This is the purpose of the skill scores. A skill score is usually defined in the following way, as by Wilson and Nurmi (2011):

$$\text{skill score} = \frac{\text{score for the forecast} - \text{score for the standard forecast}}{\text{perfect score} - \text{score for the standard forecast}}. \tag{2.19}$$

Here "the forecast" is the forecast which we want to evaluate, and "the standard forecast" is some reference forecast, like a climatology or an operational forecast. Specifically, the continuous rank probability skill score (CRPSS), which has a perfect score of 0, can be written as

$$CRPSS = 1 - \frac{\overline{CRPS_{forecast}}}{\overline{CRPS_{ref}}}, \tag{2.20}$$

where the bars signify averages. The Brier skill score (BSS) can be written as

$$BSS = 1 - \frac{BS_{forecast}}{BS_{ref}}, \tag{2.21}$$

and the quantile skill score (QSS) as

$$QSS = 1 - \frac{QS_{forecast}}{QS_{ref}}, \tag{2.22}$$

as both the BS and the QS have a perfect score of 0. The skill scores have a range from $-\infty$ to 1, with 1 being the skill score of the perfect forecast.

3. Data and exploratory analysis

This chapter gives a brief presentation of the data used in the thesis and where it comes from. It explores trends in observations and the quality of ensemble forecasts.

3.1 Data

This thesis considers forecasts of wind speed and direction at 204 locations in Norway from the 51 members of the ECMWF ENS ensemble, and measurements of maximum 10-minute average wind speed the last hour.

The data set containing the observations covers the period between 1^{st} January 2006 and 31^{th} December 2015. Ensemble forecasts were made available for the period from 24^{th} November 2013 to 22^{th} January 2016. All in all this results in a set of observations and corresponding forecasts for the period between 24^{th} November 2013 and 31^{th} December 2015, i.e. a little more than 2 years of data. All of the forecasts considered in this thesis were generated at 00 UTC with lead times from 0 to 114 hours at intervals of 6h.

The 51 members of the ECMWF ensemble can be considered exchangeable, although one member, the control member, is slightly more reliable as it has not been perturbed, i.e. it "utilises the most accurate estimate of the current conditions and the currently best description of the model physics" (ECMWF, 2016). The remaining 50 members are created with initial conditions perturbed around the intitial conditions of the control member in pairwise symmetrical perturbations.

Wind direction is given in degrees between $0°$ and $360°$, and indicates the direction the wind is coming from, with

$$0° = 360° = \text{north (N)},\ 90° = \text{east (E)},\ 180° = \text{south (S)},\ 270° = \text{west (W)},$$

and so on. A visual representation of this is found in Figure 3.1, which is taken from Pidwirny (1999).

Figure 3.2 shows all the observation sites for which forecasts are available. The site in red is Ytterøyane fyr, located in Sogn og Fjordane at

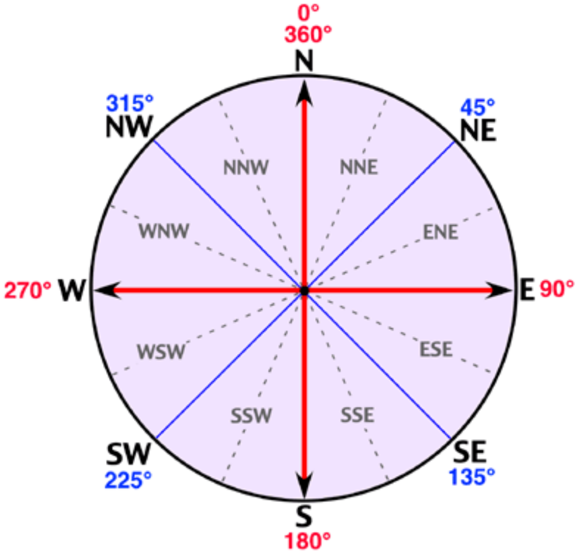


Figure 3.1: A wind compass showing how different angles correspond to different wind directions.

61°34′18″N 4°40′54″E. This site will be given special attention as it is representative of many areas with similar weather conditions on the west coast of Norway. As observations from some sites are either lacking or unreliable, only sites with more than 500 observations (circled in dark blue in Figure 3.2), will be used in this thesis.

Lead time	Correlation
6 h	0.729
72 h	0.639
114 h	0.535

Table 3.1: Correlation between the ensemble forecasts and their validating observations for lead times $l = 6, 72, 114$.

As for the forecasts, only 3 lead times will be considered: 6, 72 and 114 h. As previously mentioned, it tends to be easier to make specific forecasts for shorter lead times, while growing uncertainty makes forecasting more difficult for longer lead times. The correlations presented in Table 3.1, between the ensemble forecasts for different lead times and their validating observations, reflect this fact. In fact the correlation seems to decrease

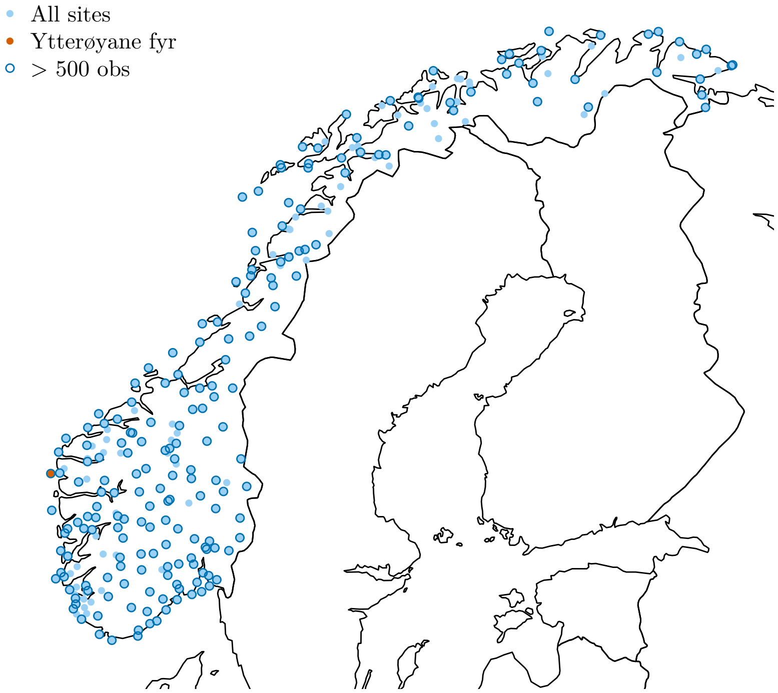


Figure 3.2: The location of the sites from which we have observations and corresponding forecasts.

linearly with lead time. This is why one short lead time and one long have been selected to be studied in this thesis. One moderately long lead time has also been chosen, as forecasts a few days into the future are often those of the greatest interest. It is also interesting to see for which lead times postprocessing of the ensemble forecasts has the greatest effect.

As all the forecasts considered in this thesis were initialized at 00 UTC, each lead time corresponds to one specific hour of the day; $l = 6$ h corresponds to 6 a.m., $l = 72$ h corresponds to 12 a.m. and $l = 114$ h corresponds to 6 p.m. UTC.

3.2 Exploratory analysis

Table 3.2 shows the minimum and maximum forecast and observed wind speed as well as the mean CRPS of the ensemble at each of the three lead times over all 204 sites. Observed wind speeds are recorded to an accuracy of 0.1 m/s. Somewhat surprisingly the forecasts with lead time $l = 72$ h have the lowest CRPS. This is also the lead time with the lowest maximum observed wind speed, which might indicate that the ensemble is less accurate when it comes to forecasting stronger winds.

	Forecast		Observed		
Lead time	Min.	Max.	Min.	Max.	Ensemble CRPS
6 h	0.001	26.461	0.0	38.1	1.822
72 h	0.002	30.350	0.0	37.4	1.769
114 h	0.002	30.024	0.1	43.5	1.907

Table 3.2: Lowest and highest wind speed forecast by the ensemble at lead times $l = 6, 72, 114$*, lowest and highest wind speed observed and CRPS of the ensemble forecasts. All 204 sites are used.*

Figure 3.3 shows the average and maximum observed and forecast wind speed between 24^{th} November 2013 and 31^{th} December 2015 at each observation site, and there are some clear patterns. Along the coast and in areas with higher altitude the wind tends to be stronger than in the lowlands. In fact, the highest average observed wind speed is more than 7 times as high as the lowest. The forecast wind speeds are also generally lower than those observed, especially in mountain areas in the west, away from the coast.

The discrepancy between observed and forecast wind speeds can also be seen in Figure 3.4 which shows 30 observations plotted alongside empirical quantiles of the $l = 72$ h ensemble forecasts. The observations almost always lie outside the ensemble range, and this seems to be especially true for the observations greater than 15 m/s.

The tendency to underforecast is confirmed by Figure 3.5, which shows rank histograms of the raw ensemble forecasts for all sites at lead times 6, 72 and 114 h. Clearly the forecasts suffer from a strong negative bias, as the observations are very often larger than every single ensemble member. They are also underdispersed and need some kind of calibration. For longer lead times the forecasts are naturally less accurate which accounts for a higher degree of dispersion and a lower bias.

As for the direction forecasts, no validating observations are available, making it hard to assess the quality of the forecasts. It is however possible

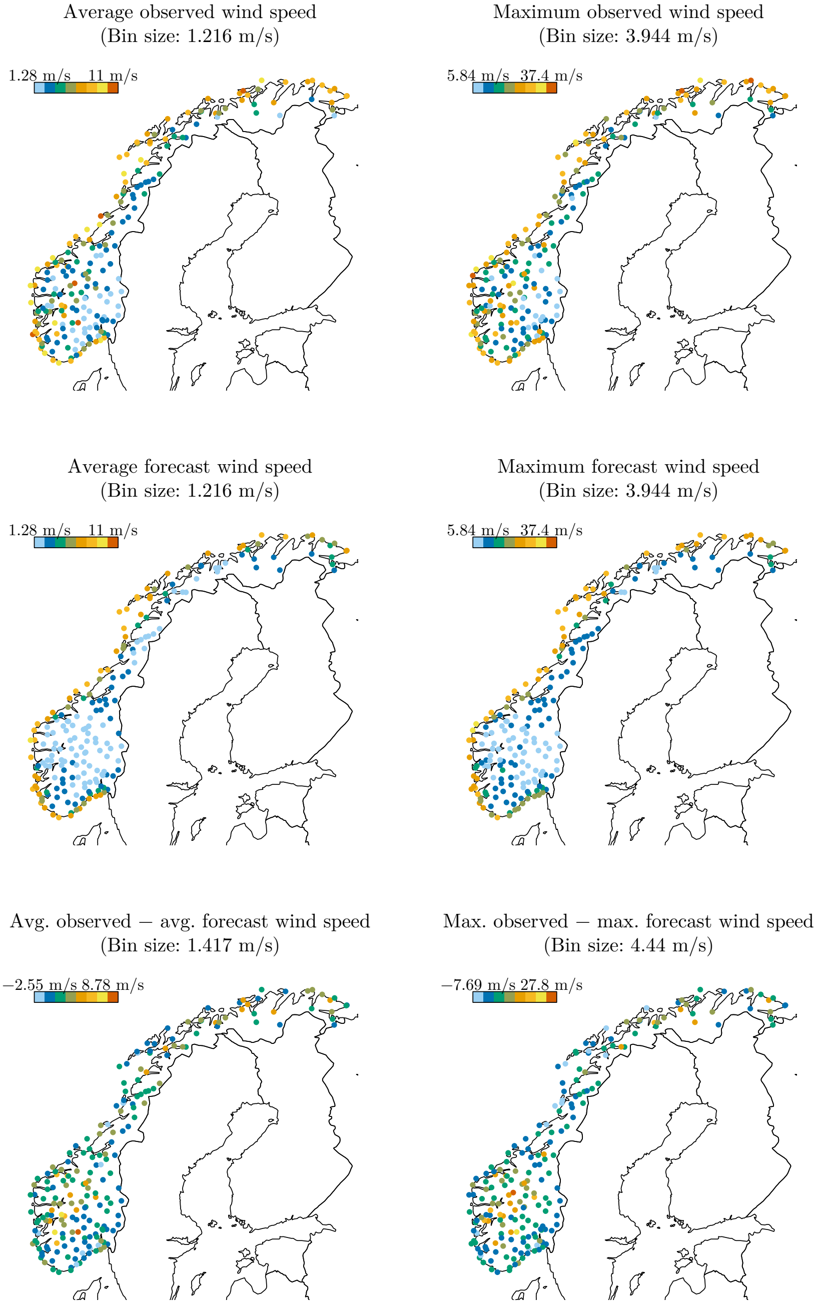


Figure 3.3: The average and maximum wind speed observed and forecast at each site with lead time 72 h.

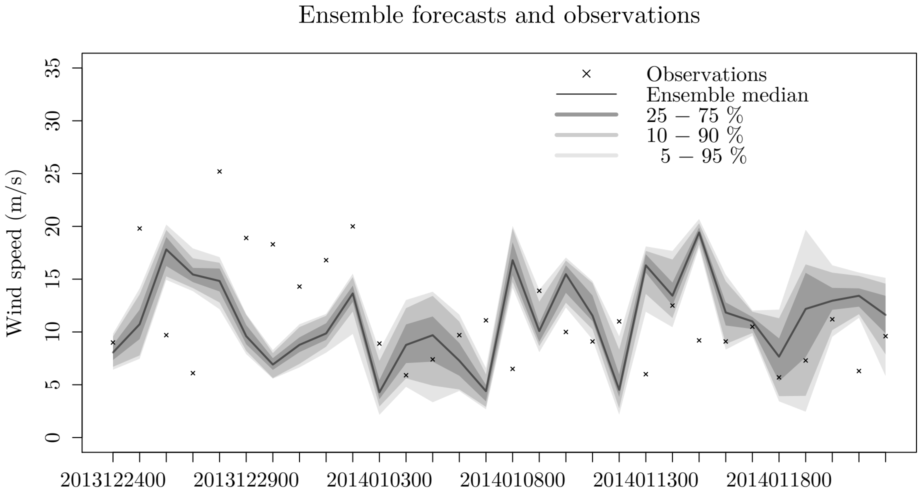


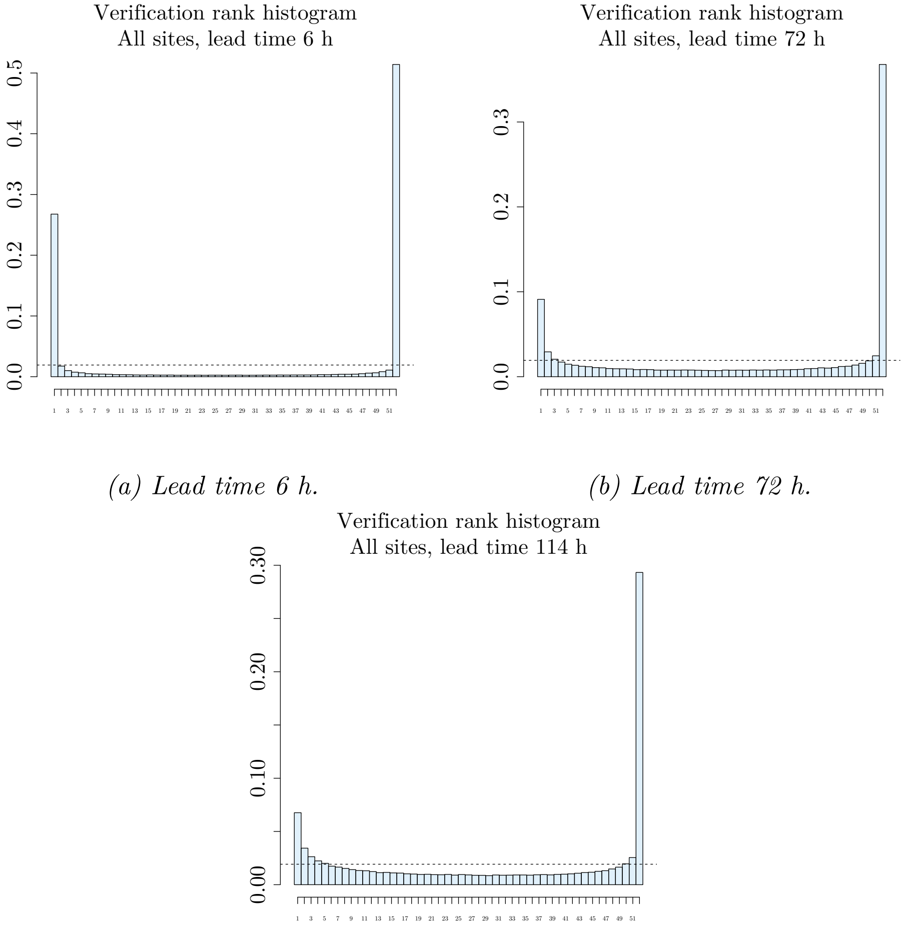
Figure 3.4: Median and 50 % (dark grey), 80 % (medium grey) and 90 % (light grey) forecast intervals for the raw ensemble forecasts with lead time $l = 72$ h, and validating observations as points.

to look for patterns and trends. Figure 3.6 shows 4 wind rose diagrams[1], diagrams where observed wind speeds are binned according to forecast wind direction and plotted as so-called paddles. The thin lines near the center represent low wind speeds and thicker paddles represent higher wind speeds. The length of each paddle illustrates the proportion of the observations that fall within that bin. The wind directions (N,S,E,W) indicate what direction the wind comes from.

The diagrams in Figure 3.6 were made using the forecast wind directions of the control member of the ensemble with lead time $l = 72$ h, and observed wind speed. Three of the sites for which the diagrams were made are sites where high wind speeds are frequently observed, namely Ytterøyane fyr, Røldalsfjellet – Elvershei and Hasvik – Sluskfjellet. The last site, Oslo – Blindern was included for comparison because this is a site where particularly high wind speeds are never observed. The maximum observed wind speed at each site is 25.2 m/s, 37.4 m/s, 34.3 m/s and 11.4 m/s respectively.

At all four sites wind is much more frequently forecast coming from certain directions. At Ytterøyane fyr, meridional winds (winds from the north or from the south) are overrepresented, and at Røldalsfjellet – Elvershei it

[1] Created using the `windRose` function in the R package *openair*.



(a) Lead time 6 h.

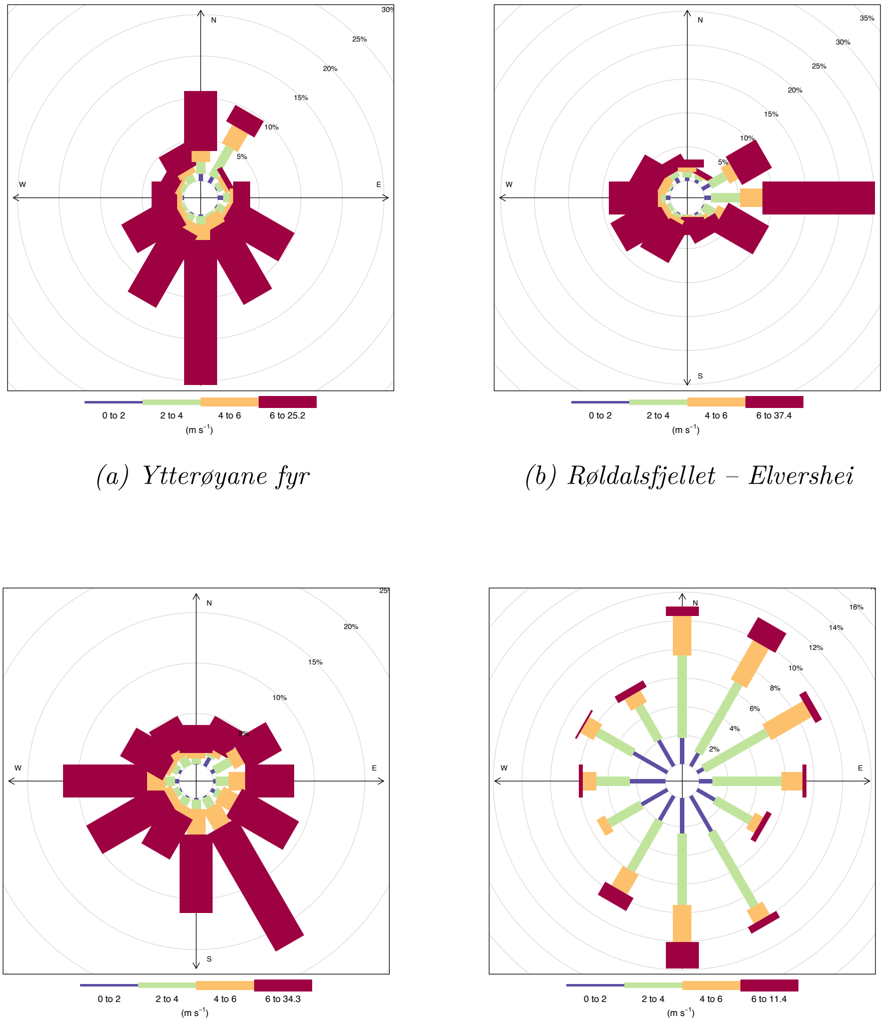
(b) Lead time 72 h.

(c) Lead time 114 h.

Figure 3.5: Rank histogram for all sites, for lead times 6, 72 and 114 hours. The histograms indicate bias and underdispersion.

is zonal winds (winds from the west or from the east) that dominate. At both of these sites there is a clear prevailing wind direction. This is likely to be related to the location and the topography of the sites, i.e. the model terrain and proximity to features such as mountains, fields, rivers and lakes or the ocean.

The third diagram in Figure 3.6, for Hasvik – Sluskfjellet, is a bit special, as it is westerly and southeasterly winds that dominate. In the fourth diagram, for Oslo – Blindern, there is a more even distribution of directions,



(a) Ytterøyane fyr

(b) Røldalsfjellet – Elvershei

(c) Hasvik – Sluskfjellet

(d) Oslo – Blindern

Figure 3.6: Wind rose diagrams for wind direction forecasts of the control member of the ensemble with $l = 72$ h and observations of wind speed for sites Ytterøyane fyr, Røldalsfjellet – Elvershei, Hasvik – Sluskfjellet and Oslo – Blindern. Data from 24th November 2013 to 31st December 2015.

although stronger winds are rarely observed when zonal wind is forecast, and very low wind speeds are rarely observed when wind from the northeast is forecast.

If the forecasts are representative of the reality, one thing is certain: the relationship between wind speed and direction is tremendously different from one site to the next.

4. Method and models

4.1 BMA model for wind forecasts

The basic BMA approach as described in Section 2.2 simply uses every member of the ECMWF ensemble with equal weight. The model, following Equation (2.1), can be written as

$$p(y|f_0, ..., f_{50}; \theta) = \sum_{m=0}^{50} wg(y|f_m; \theta) \tag{4.1}$$

where f_0 is the control member and $f_1, ..., f_{50}$ are the other ensemble members. Hereafter this model is referred to as BASICBMA.

4.2 The control member

As previously mentioned in Section 3.1, it is possible to think of all 51 members of the ensemble as exchangeable, but as the control member has the optimal initial conditions it might also be a good idea to treat it as separate from the other members. Figure 4.1 shows the mean absolute error (MAE) of each ensemble member with respect to observed wind for different lead times. For all lead times the first member, which is the control member, has the smallest MAE. For longer lead times it is markedly smaller than the rest.

Is this difference big enough to solicit separate modelling of the bias and a different weight? To investigate this, two simple models are tested, one in which all ensemble members are treated as exchangeable (BASICBMA) and one in which the expectation of the control member is modelled separately. The latter is hereafter referred to as CMSBMA. In both cases lead time $l = 72$, and the site Ytterøyane fyr is used.

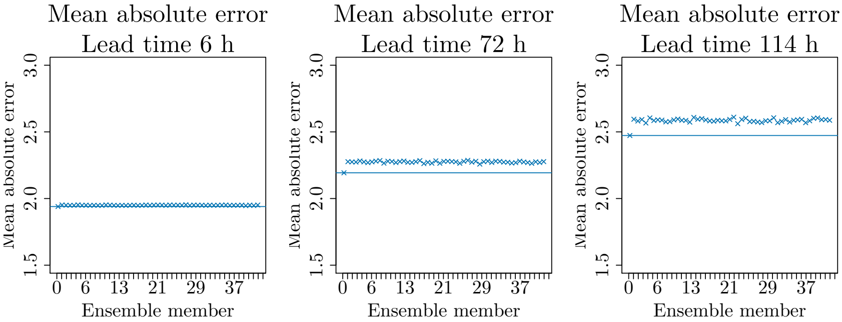


Figure 4.1: Mean absolute error each ensemble member in the ensemble with respect to observed wind speed. The horizontal line is the MAE of the control member.

4.3 Climatology and historical observations

As mentioned in Chapter 3, making specific forecasts gets harder for increasing lead times, and if we go far enough into the future it becomes impossible. For these lead times the best forecast tends to be the climatology.

In order to examine whether inclusion of a climatology might help the forecasts all observations from between 1$^{\text{st}}$ January 2006 and 23$^{\text{rd}}$ November 2013 are used to calculate a climatology. A climatology is constructed for day d by averaging over all observations from within $d \pm 30$ days (regardless of year). This means for example that for d = 3$^{\text{rd}}$ March, the assigned climatology is the average of all observations between 2$^{\text{nd}}$ February and 2$^{\text{nd}}$ April of 2006, 2007, $\cdots$, 2013.

As an experiment historical observations are included as a second ensemble, such that the observations from hour h on day d of 2006, 2007, $\cdots$, 2012 are used as a 7 member ensemble forecast for hour h of day d of 2014 and 2015.

Two BMA approaches will be tested: one in which a climatology is used as a single member ensemble (Equation (4.2)), and one in which historical observations are used as a 7 member ensemble in addition to the 51 member ECMWF ensemble (Equation (4.6)). Using the following abbreviations: cm = the control member of the ECMWF ensemble, ens = other members of the ECMWF ensemble, cl = climatology, and ho = historical observation, the first model can be written as

$$\begin{aligned}p(y|f^{\text{cm}}, f_1^{\text{ens}}, ..., f_{50}^{\text{ens}}, f^{\text{cl}}; \theta^{\text{cm}}, \theta^{\text{ens}}, \theta^{\text{cl}}) = \\ w^{\text{cm}} g^{\text{cm}}(y|f^{\text{cm}}; \theta^{\text{cm}}) + \sum_{m=1}^{50} w^{\text{ens}} g^{\text{ens}}(y|f_m^{\text{ens}}; \theta^{\text{ens}}) + w^{\text{cl}} g^{\text{cl}}(y|f^{\text{cl}}; \theta^{\text{cl}})\end{aligned} \tag{4.2}$$

where $\theta = (\alpha, \beta)^T$ and the component PDF for the control member

$$g^{\text{cm}}(y|f^{\text{cm}}) = \frac{1}{\beta_{\text{cm}}^{\alpha_{\text{cm}}} \Gamma(\alpha^{\text{cm}})} y^{\alpha^{\text{cm}}-1} \exp(-y/\beta^{\text{cm}}) \tag{4.3}$$

(and similarly for g^{ens} and g^{cl}) where α^{cm} and β^{cm} are found from their relationship with μ^{cm} and σ^{cm}, which are given by

$$\mu^{\text{cm}} = b_0^{\text{cm}} + b_1^{\text{cm}} f^{\text{cm}} \tag{4.4}$$

$$\sigma^{\text{cm}} = c_0 + c_1 f^{\text{cm}} \tag{4.5}$$

where b_0^{cm} and b_1^{cm}, c_0 and c_1, are estimated in accordance with the method described in Section 2.2. This model is hereafter referred to as CLIMBMA. Analogously, the second model becomes

$$\begin{aligned}p(y|f^{\text{cm}}, f_1^{\text{ens}}, ..., f_{50}^{\text{ens}}, f_1^{\text{ho}}, ..., f_7^{\text{ho}}; \theta^{\text{cm}}, \theta^{\text{ens}}, \theta^{\text{ho}}) = \\ w^{\text{cm}} g^{\text{cm}}(y|f^{\text{cm}}; \theta^{\text{cm}}) + \sum_{m=1}^{50} w^{\text{ens}} g^{\text{ens}}(y|f_m^{\text{ens}}; \theta^{\text{ens}}) + \sum_{n=1}^{7} w^{\text{ho}} g^{\text{ho}}(y|f_n^{\text{ho}}; \theta^{\text{ho}}),\end{aligned} \tag{4.6}$$

which is hereafter referred to as OBSBMA.

The models are fitted to data from Ytterøyane fyr. Figure 4.2 shows observations, the fitted climatology and historical observations for lead time 114 h. As a tendency to underforecast for strong winds has already been discussed, the climatology cannot be expected to improve forecasts for shorter lead times. It is rather for longer lead times this might have a positive effect. The historical observations however might help increase the spread of the forecasts for all lead times.

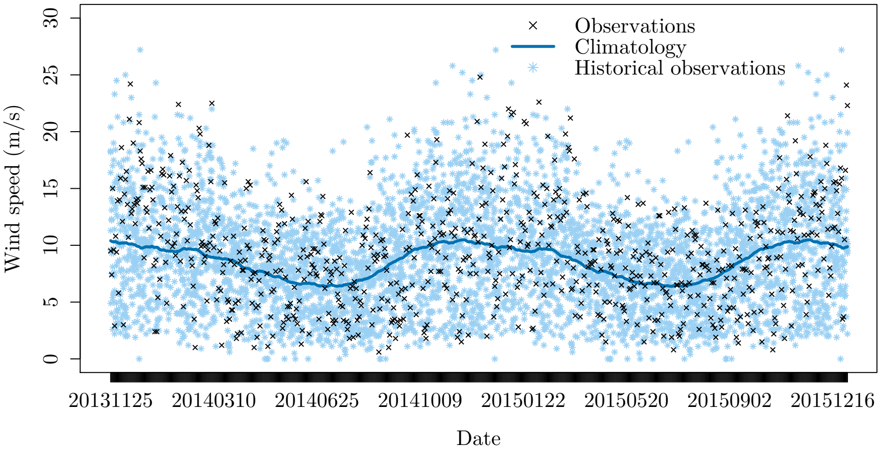


Figure 4.2: Observations, climatology and historical observations for lead time 114 at Ytterøyane fyr between 25th November 2013 and 31st December 2015.

4.4 BMA with wind direction in expectation

In addition to forecasts of wind speed the ECMWF ensemble forecasts wind direction. In an attempt to make use of these data, thin plate regression splines are used to bring forecast wind direction into the modelling of the expectation of the predictive PDFs in the BMA. Thus, the expression for the expectation of component PDF m becomes

$$\mu_m = s(f_m, d_m), \tag{4.7}$$

where f_m and d_m are the forecast wind speed and wind direction respectively of ensemble member m, and s is a smooth based on thin plate regression splines, as defined in Section 2.3. Wind direction is a circular quantity, meaning that direction $0° = 360°$. Ideally, the fitted spline should also be the same for these values. Unfortunately, this is not something the thin plate regression splines can easily handle. Therefore, for the fitting of the splines, the observations corresponding to direction forecasts between 0° and 90°, and between 270° and 360° were copied and given the same forecast wind direction, only in $[360°, 450°]$ and $[-90°, 0°]$, in order to make up for this shortcoming.

The estimation of σ, as defined in Section 2.2, is unchanged. A gamma distribution is still used for the predictive PDFs, and the relationship between μ and the distribution remains the same. Keeping the notation from the previous section and denoting direction by d, the full model now becomes

$$\begin{gathered} p(y|(f^{\text{cm}}, d^{\text{cm}}), (f_1^{\text{ens}}, d_1^{\text{ens}}), ..., (f_{50}^{\text{ens}}, d_{50}^{\text{ens}}); \theta^{\text{cm}}, \theta^{\text{ens}}) = \\ w^{\text{cm}} g^{\text{cm}}(y|(f^{\text{cm}}, d^{\text{cm}}); \theta^{\text{cm}}) + \sum_{m=1}^{50} w^{\text{ens}} g^{\text{ens}}(y|(f_m^{\text{ens}}, d_m^{\text{ens}}); \theta^{\text{ens}}). \end{gathered} \tag{4.8}$$

where g^{cm} and g^{ens} are found from Equation (2.3), and the only thing that has changed is the modelling of μ^{cm} and μ^{ens}. This model is hereafter referred to as DIRBMA.

4.5 Training and test scheme

Before we can make predictions we must select the data to which the model shall be fitted. There are many ways to do this, but with BMA it is common to use a sliding window training scheme where the treatment of the forecast for any given day comes from a model based on the k last days. This, however, requires recomputation of the model every day, something one might want to avoid, if possible.

To see whether or not it is necessary to fit a new model every day, and determine the number of training days to use, 5 different training schemes are tested on the last 375 days of the data at site Ytterøyane fyr with lead time 72 h. First a training period of 30 days is used, with models fitted daily and weekly. Then a training period of 60 days is used, again with models fitted daily and weekly. Finally one model is fitted in which 376 days are used for training and this same model is used for the entire test period.

4.6 Software

For all the programming executed in connection with this thesis the programming language R was used. As BMA is a commonly used postprocessing technique in weather forecasting, there are several very good R packages dedicated to it, such as *ensembleBMA* by Fraley et al. (2015). However, as this thesis discusses certain modifications to the method, the preexisting packages could not be used, and functions for fitting, forecasting and assessing the probability forecasts of BMAs had to be written. The function for fitting a BMA is enclosed in Appendix A. In the modified method where the modelling of the expectation was done with thin plate regression splines, the function `gam` from the R package *mgcv* was used.

5. Results

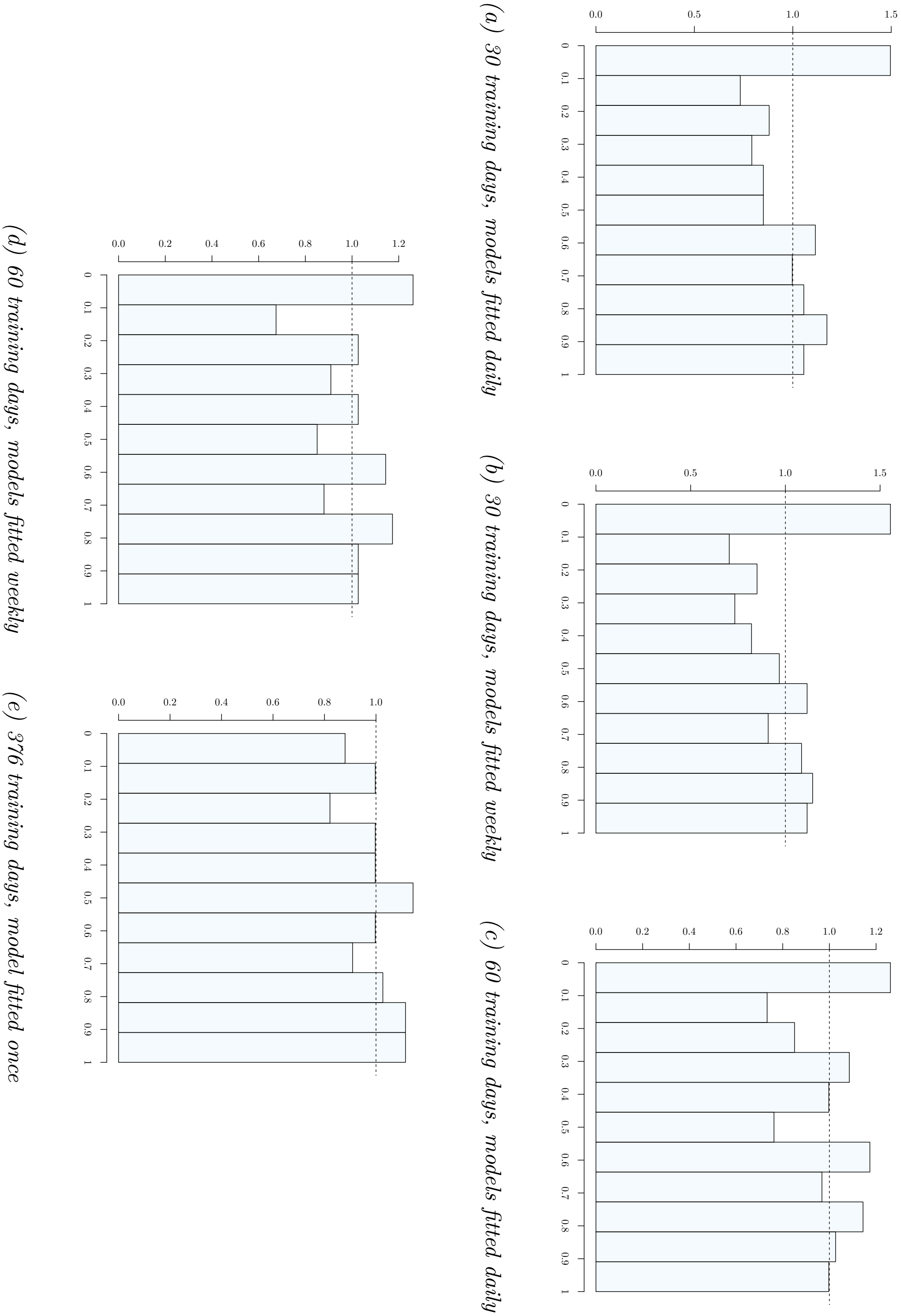
5.1 Training scheme

Figure 5.1 shows PIT histograms for the different training schemes introduced in Section 4.5 using the BASICBMA model described in Section 4.1. The number of days used in the evaluation is 375. This is a considerable number, but it would still be unwise to infer too much from these plots alone. However, it looks like there is no striking advantage to fitting a new model every day compared to once a week. It also seems like increasing the number of training days might have a positive effect. Out of the five schemes, the last one, where 376 training days and only one model is used, looks like it might be the one that is the most calibrated.

The PIT histogram is a useful, but far from perfect tool, and problems with the model can go undetected, as previously mentioned in subsection 2.4.2. In Figure 5.2 the forecasts have been divided into 4 groups according to their median, from the lowest to the highest, and a PIT histogram is plotted for each group in each model. This is done in order to examine the behaviour of the model in extreme events. The skewness exhibited by many of the plots will be discussed in Section 5.2. All that will be mentioned here is that none of the models stand out as being much better than the others.

For a more tangible, quantitative comparison, the CRPSS for each case relative to the raw ensemble is shown in Table 5.1, as well as BSS at thresholds 5 m/s and 20 m/s. What is perhaps the most striking is the fact that only one of the models has a positive CRPSS. It seems clear that the difference between daily and weekly models is small and that what really matters is the number of training days.

Because results are shown to be as good as or better than those of the other training schemes, the scheme that fits the model just once using 376 training days will be used throughout the rest of the thesis.



(a) 30 training days, models fitted daily

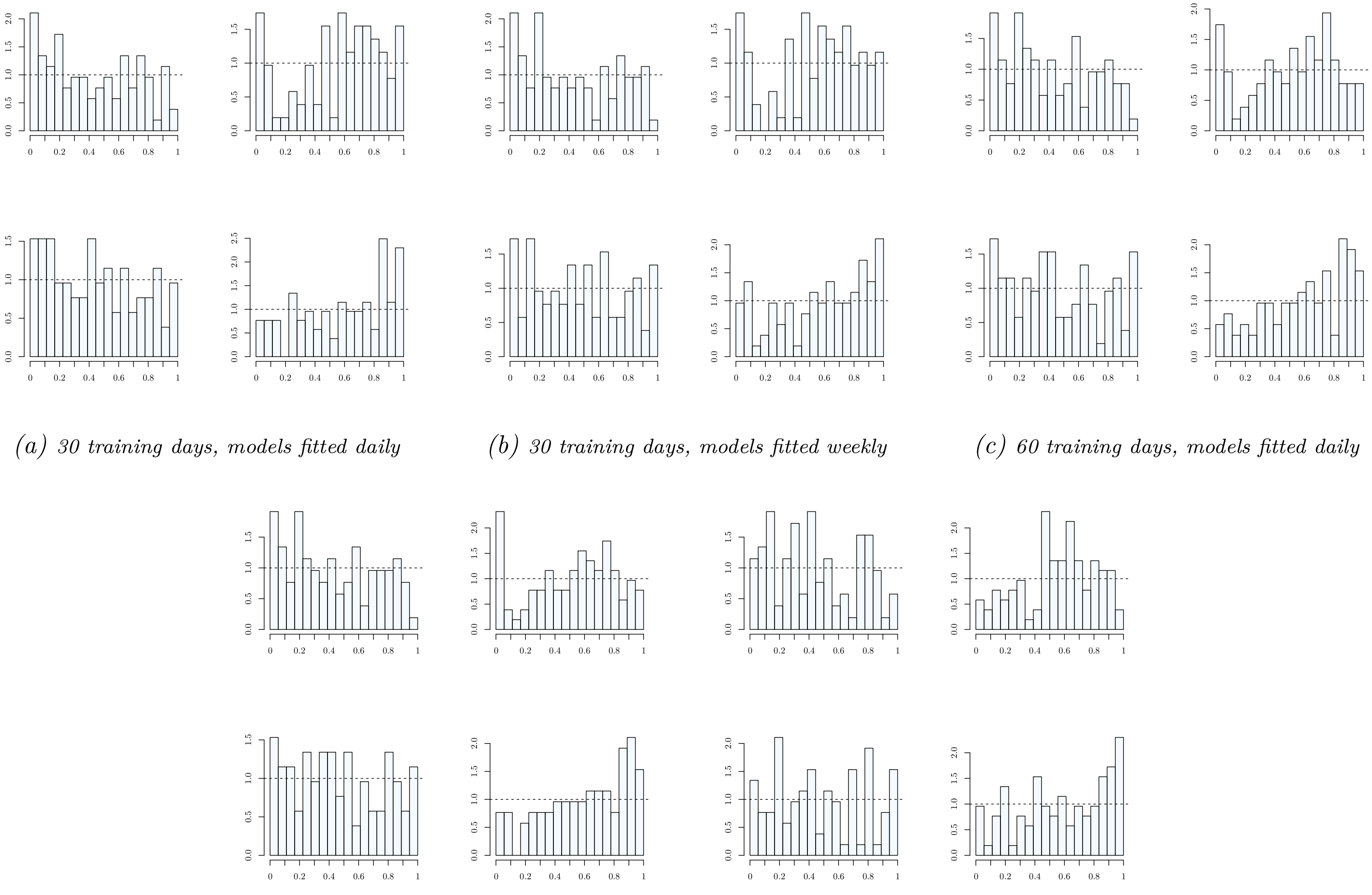
(b) 30 training days, models fitted weekly

(c) 60 training days, models fitted daily

(d) 60 training days, models fitted weekly

(e) 376 training days, model fitted once

Figure 5.1: PIT histograms for different training schemes evaluated on a period of 375 days for site Ytterøyane fyr.



(a) 30 training days, models fitted daily

(b) 30 training days, models fitted weekly

(c) 60 training days, models fitted daily

(d) 60 training days, models fitted weekly

(e) 376 training days, model fitted once

Figure 5.2: PIT histograms for each training scheme, by forecast median. In each plot: top left shows 0%-25 %, top right shows 25%-50%, bottom left shows 50%-75%, bottom right shows 75%-100%

BMA model fitted	Training days	CRPSS	BSS, 5 m/s	BSS, 20 m/s
daily	30	-0.041	-0.094	-0.086
weekly	30	-0.044	-0.000	-0.029
daily	60	-0.014	-0.071	-0.036
weekly	60	-0.021	0.002	-0.026
once	376	0.053	0.000	0.000

Table 5.1: CRPSS and BSS of the BMA models compared to the raw ensemble, evaluated on a period of 375 days using various training schemes.

5.2 Bias in forecasts of strong wind

Figure 5.3 shows observations from 11th January 2015 to 11th March 2015 as points with median and 50 %, 80 % and 90 % forecast intervals for the raw ensemble forecasts and a fitted BASICBMA. It is easy to see that the use of BMA leads to an increase in the dispersion of the ensemble, meaning that the model is better able to forecast those observations that are outside the range of the raw ensemble, such as, for example, between the 12th and the 14th of February 2015. Whether it also reduces the bias is harder to see with the naked eye, though the root mean square error of the median of the raw ensemble for this period is 3.73 and for the BASICBMA it is 3.61.

Figure 5.4 shows PIT histograms of the BASICBMA forecasts at Yterøyane fyr with lead time $l = 72$ hours. The data are divided according to the sample quartiles of the medians of the predictive pdfs. Here the top left plot is a PIT histogram where only the forecasts whose median is in the lowest 25 % are included. In the top right plot forecasts with median in the 25 % to 50 % quartile are used. The bottom left plot shows the 50 % to 75 % quartile and the bottom right plot is a PIT histogram of the forecasts whose median is in the top 25 %.

In the interquartile range, panels (b) and (c) of Figure 5.4, no discernible bias is observed. However, in the cases of low and high forecast wind speeds, panels (a) and (d) of Figure 5.4 respectively, clear biases are observed. For low forecast wind speeds, the observed wind speeds are mostly lower, and for high forecast wind speeds the observed wind speeds tend to be even more “extreme”. A simple t-test confirms that these trends are indeed significant at a 0.1 % significance level.

To examine whether this might be the case regardless of lead time, analogue plots were created for lead times 6 hours and 114 hours, displayed in Figures B.1 and B.2, which can be found in Appendix B. In Figure B.1,

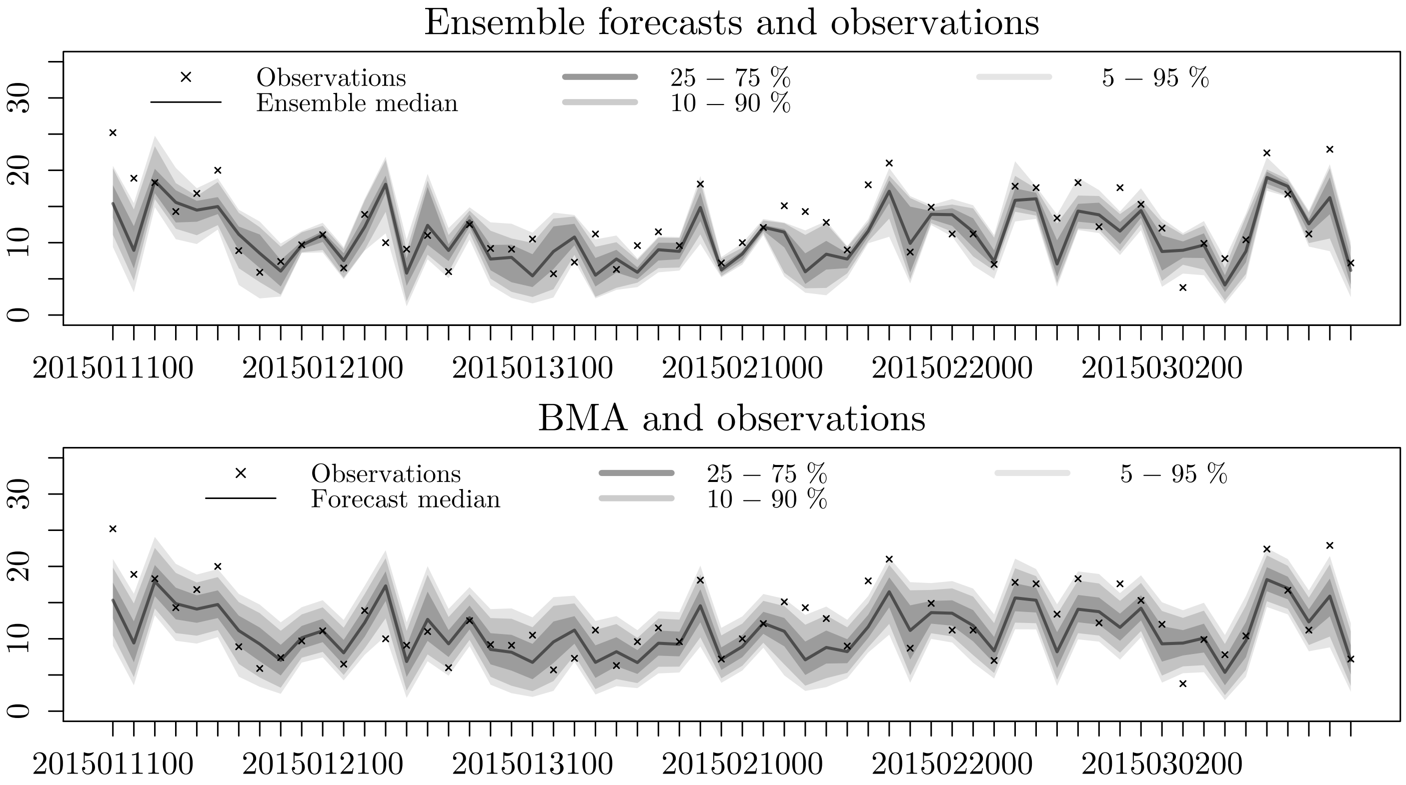
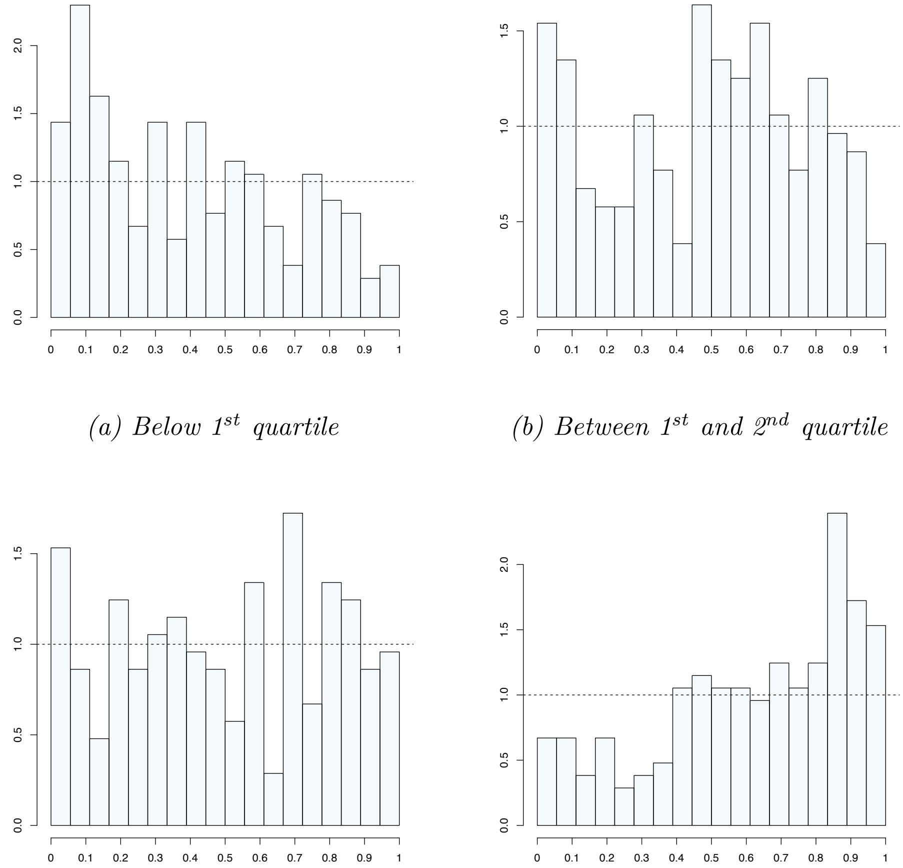


Figure 5.3: Sample quantiles of the raw ensemble forecasts (upper panel), and quantiles of the BASICBMA predictive PDF (lower panel), both with lead time $l = 72$, as well as observations from Ytterøyane fyr for the period from 11^{th} January 2015 to 11^{th} March 2015. The solid line shows the median while the shaded areas are the 50 % (dark grey), 80 % (medium grey) and 90 % (light grey) forecast intervals.



(a) Below 1^{st} quartile

(b) Between 1^{st} and 2^{nd} quartile

(c) Between 2^{nd} and 3^{rd} quartile

(d) Above 3^{rd} quartile

Figure 5.4: PIT histograms conditioning on the median of the BASICBMA predictive PDFs. Lead time $l = 72$ and site Ytterøyane fyr was used.

which shows $l = 6$ h, none of the plots clearly display any of the trends found in Figure 5.4. However, in Figure B.2, illustrating the forecasts at $l = 114$ h, these trends are present, especially for strong winds, suggesting that these may be particularly hard to predict far in advance.

So far only Ytterøyane fyr has been considered. To determine whether the problem is a common one or merely due to a really unfortunate choice of site, a number of additional sites were tested. As the problem is tied to the forecasting of strong winds, the sites that were examined were sites where strong winds are regularly observed. The findings, presented as PIT histograms for models fitted at the various sites with lead time 72 h, which can be found in Appendix C, were inconsistent, but several of them display the same tendency toward a bias in the forecasts of strong winds as the one seen in the plots for Ytterøyane fyr.

5.3 Exchangeability

In Table 5.2, different scoring rules are applied to models with all members and all but the control member of the ensemble respectively considered exchangeable, i.e. BASICBMA and CMSBMA. The skill scores (BSS and CRPSS) are calculated with the raw ensemble forecast as reference. The QS is included to assess the performance of the models for the higher quantiles, as this is where forecasting seems to be more challenging. The CMSBMA model is better for all scores. This suggests that the control member might in fact be different enough to solicit separate modelling of the expectation.

	CRPSS	BSS 5 m/s	BSS 20 m/s	QS 0.9	QS 0.95
BASICBMA	0.053	0.005	0.001	0.539	0.333
CMSBMA	0.054	0.006	0.002	0.539	0.327

Table 5.2: Scoring rules applied to models BASICBMA and CMSBMA. BSS is calculated with thresholds at 5 m/s and 20 m/s. BSS and CRPSS are relative to raw ensemble forecasts.

The weights for each model which are, as previously mentioned in Section 2.2, based on predictive performance, are shown in Table 5.3. For the model with all members treated as exchangeable, the weights are equal, $1/M = 0.0196$ for all members.

The model in which the control member is treated separately has two fitted weights, one for the control member and one for every exchangeable member. The fact that the weight assigned to the control member by the

	Ensemble member weights	
Model	Control member	Other members
BASICBMA	0.0196	0.0196
CMSBMA	0.0549	0.0189

Table 5.3: Weights of models with all 51 ensemble members considered exchangeable and with the control member treated as different

EM algorithm is considerably larger than that assigned to the exchangeable members supports the notion that the control member should be treated separately from the perturbed members.

5.4 Climatology and historical observations

Table 5.4 shows different skill scores applied to two models, CLIMBMA and OBSBMA (defined in Section 4.3). The scores show the improvement relative to the CMSBMA model in which neither climatology or historical observations have been used.

Lead time	Model	CRPSS	BSS 5	BSS 20	QSS 0.90	QSS 0.95
6 h	CLIMBMA	0.003	0.003	0.011	0.003	-0.009
6 h	OBSBMA	-0.000	0.000	0.000	-0.000	-0.000
72 h	CLIMBMA	0.001	0.000	-0.000	0.000	-0.000
72 h	OBSBMA	0.000	0.000	-0.000	-0.000	-0.000
114 h	CLIMBMA	-0.000	0.000	0.000	0.000	0.000
114 h	OBSBMA	-0.000	0.000	0.000	0.000	0.000

Table 5.4: CRPSS, BSS and QSS of models CLIMBMA and OBSBMA relative to CMSBMA, for lead times $l = 6, 72, 114$ h.

As the uncertainty of the ensemble forecasts increases for longer lead times, one might expect to see an improvement, especially for these lead times, when introducing a climatology. However, the results presented in Table 5.4 show that the effect both of adding the climatology and of adding an ensemble of historical observations to the probabilistic models is astoundingly underwhelming. To further emphasize the lack of influence given to these added variables, the weight assigned to each ensemble member in each model for lead time 114 is shown in Table 5.5.

Clearly, even with lead times as long as 114 hours, the ensemble fore-

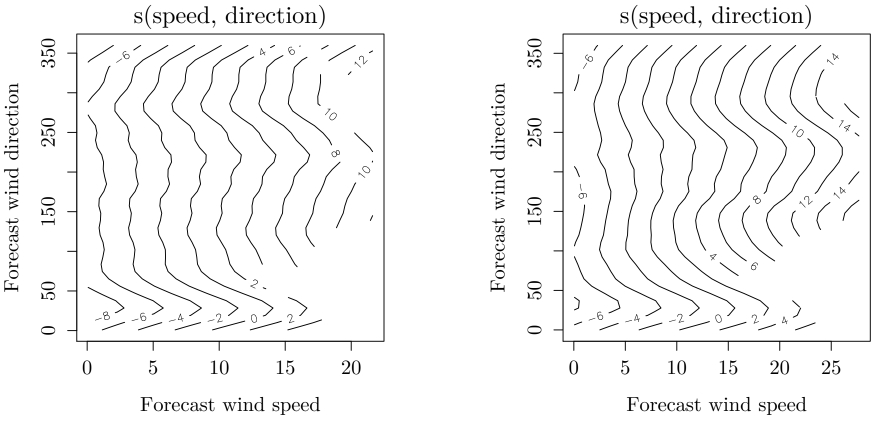
Member	Weight	Member	Weight
Control member	0.395	Control member	0.395
Regular members	0.012	Regular members	0.012
Climatology	4.159×10^{-38}	Historical observations	8.966×10^{-53}

Table 5.5: Weights assigned to each ensemble member in a model where climatology was added as a separate one-member ensemble, and a model where historical observations were added as a separate 7 member ensemble.

casts are still more informative than a climatology or a sample of historical observations. Consequently, neither of these ideas will be considered any further in this thesis.

5.5 Wind direction

A model where thin plate regression splines have been used to incorporate forecast wind direction, DIRBMA, specified in Section 4.4, is fitted to forecasts with lead time $l = 72$ h and observations from Ytterøyane fyr.



(a) Control member

(b) Other members

Figure 5.5: The thin plate regression splines fitted in the DIRBMA model for lead time $l = 72$ h at Ytterøyane fyr.

Figure 5.5 shows how the splines have been fitted to the forecasts in the DIRBMA model. The panel to the left shows the splines fitted to the control member forecasts, and the panel to the right is for the remaining members.

The smooths in both plots are approximately linear in forecast wind speed but non-linear in forecast wind direction, suggesting that the relationship between forecast wind speed and observed wind speed is slightly different for different forecast wind directions.

Figure 5.6 shows PIT histograms for this model akin to those in Figure 5.4 for CMSBMA. Hints of the trends observed in those plots can also be seen in Figure 5.6, but to a much lesser degree. This suggests that the use of forecast wind direction in the model has led to at least some improvement in the forecasts. To further investigate this, a comparison of scores for the model with the direction dependent expectation and one without are presented in Table 5.6.

	CRPSS	BSS 5 m/s	BSS 20 m/s	QS 0.9	QS 0.95
CMSBMA	0.054	0.006	0.002	0.539	0.327
DIRBMA	0.145	0.146	0.091	0.497	0.305

Table 5.6: CMSBMA and DIRBMA fitted to observations from Ytterøyane fyr and forecasts with lead time 72 h. The BSS is calculated for the two thresholds of 5 m/s and 20 m/s, and the QS is calculated for the 0.9 and the 0.95 quantile. The CRPSS and BSS are relative to the raw ensemble.

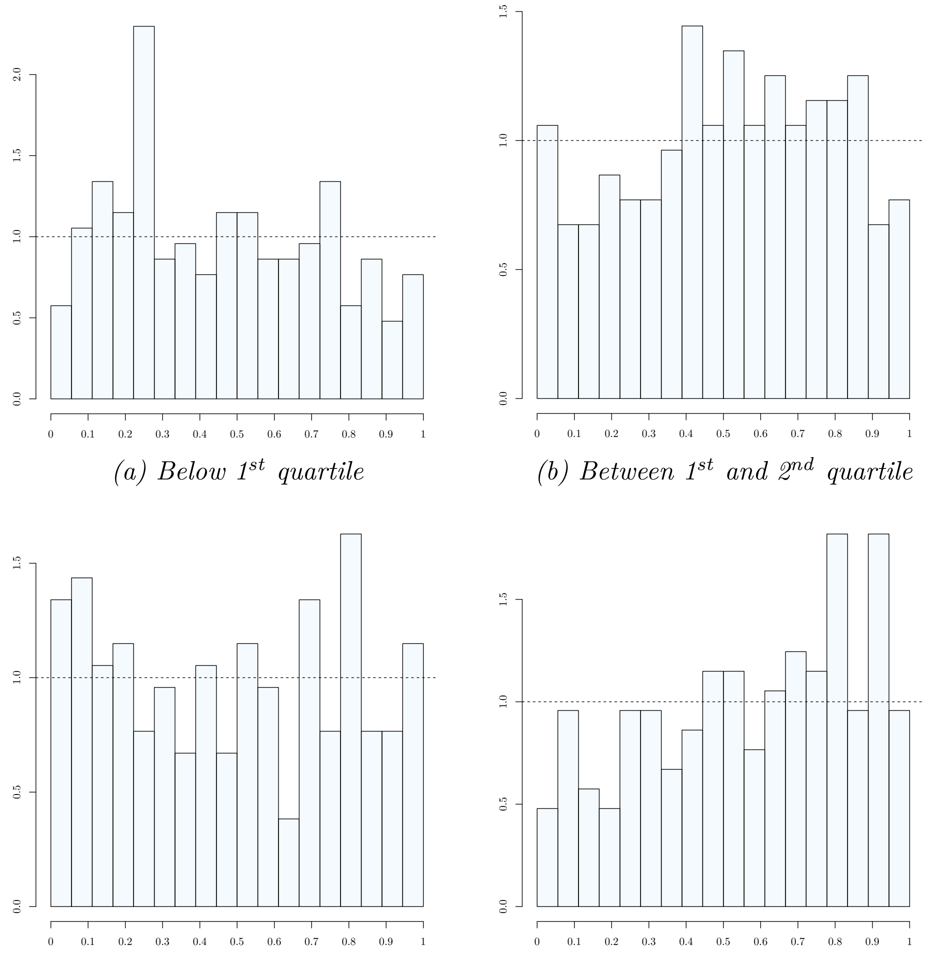
Firstly, the almost 3 fold increase in the CRPSS suggests that the wind direction dependent model, DIRBMA, performs better overall than the CMSBMA. The increase in the Brier skill scores for both thresholds suggests that the model's ability to forecast both low and high wind speeds is greatly improved by taking into account wind direction. The quantile score for the 0.9 and 0.95 quantiles also support the conclusions drawn from Figure 5.6, that the bias in the forecasts for higher wind speeds is weaker than in the CMSBMA.

In order to find out if this method is generally better than CMSBMA, similar models are fitted for all 204 sites and lead times $l = 6, 72$ and 114 h. Figure 5.7 shows a box-and-whisker plot[1] of the CRPSS of these models, as well as the CMSBMA models, both relative to the raw ensemble forecasts.

In the box-and-whisker plots the thick line is the median, the bottom and top of the box are the first and third quartiles, and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. The circles are outliers, i.e. data points outside the range of the whiskers.

Clearly the use of BMA in general leads to forecasts that, measured

[1] Created using the `boxplot` function in the *graphics* package in R.



(a) Below 1st quartile

(b) Between 1st and 2nd quartile

(c) Between 2nd and 3rd quartile

(d) Above 3rd quartile

Figure 5.6: PIT histograms conditioning on the median of the predictive PDFs from a model where forecast wind direction has been used in the modelling of the expectation. Lead time $l = 72$ and site Ytterøyane fyr was used.

in CRPS, perform much better than the raw ensemble forecasts. This is not surprising as the method removes bias and increases dispersion. The improvement seems to be especially large for shorter lead times.

For the majority of sites the DIRBMA had an even greater positive effect. However, there are certain outliers. Most of these are sites where observed wind speeds never exceed 15 m/s, and the vast majority of observations are lower than 5 m/s. This suggests that direction plays a more important role when the range of observed wind speeds is larger. The relationship between direction and speed also varies a lot from site to site, as it is closely connected to the topography of the area around the site.

Figure 5.8 shows the CRPSS of models for all sites and lead times plotted on a map. The score is divided into categories and color coded such that red points represent negative CRPSS and different shades of blue represent different levels of improvement, with darker blue indicating greater improvement. A comparison of Figure 5.8 with Figure 3.3 suggests that the sites with negative CRPSS correspond to those sites where the difference between average forecast wind speed and average observed wind speed is small.

The outliers seen in Figure 5.7 are also generally sites where the CRPS of the raw ensemble forecasts is quite low. This seems reasonable as one of the main problems with the ensemble is its tendency to forecast wind speeds that are too low. That this is less of a problem when the bulk of the observed wind speeds are lower than 5 m/s is not surprising.

Because the models' ability to forecast "extreme" values is of particular interest, Figures 5.9 and 5.10 show box-and-whisker plots of the Brier skill score of the models relative to the raw ensemble forecasts for thresholds of 5 and 20 m/s respectively, and a box-and-whisker plot of the quantile score for the 0.95-quantile is shown in Figure 5.11. In Figure 5.10 and Figure 5.11 the BSS and QS are only calculated for sites with more than 10 observations greater than 20 m/s.

The exclusion of sites that rarely or never see wind speeds over 20 m/s is done in order to avoid the situation where the ensemble gets a Brier score of 0 because speeds greater than 20 m/s have been neither observed nor forecast. In this situation the Brier score tells us nothing and results in Brier skill scores of $-\infty$.

In general, the Brier skill scores show good results with both methods. For a threshold of 5 m/s there is an improvement at most sites. There are cases of great improvement, and again there are outliers where the use of BMA has not had a positive effect. For a threshold of 20 m/s the largest improvement happens for a lead time of 6 h. For lead times of 72 h and 114

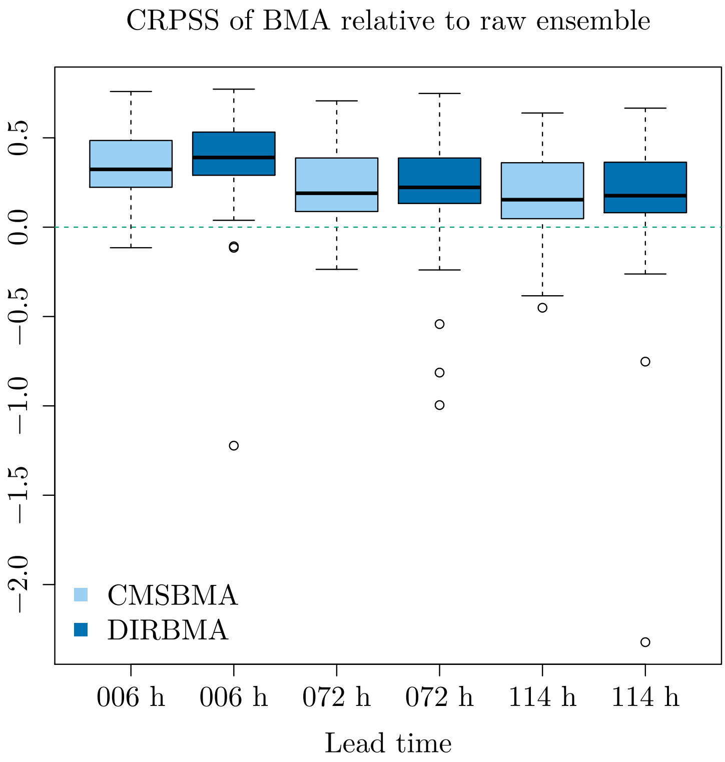


Figure 5.7: Box-and-whisker plots of the CRPSS of probabilistic forecasts from CMSBMA and DIRBMA relative to raw ensemble forecasts. Each data point is the CRPSS of a model for a specific site and lead time.

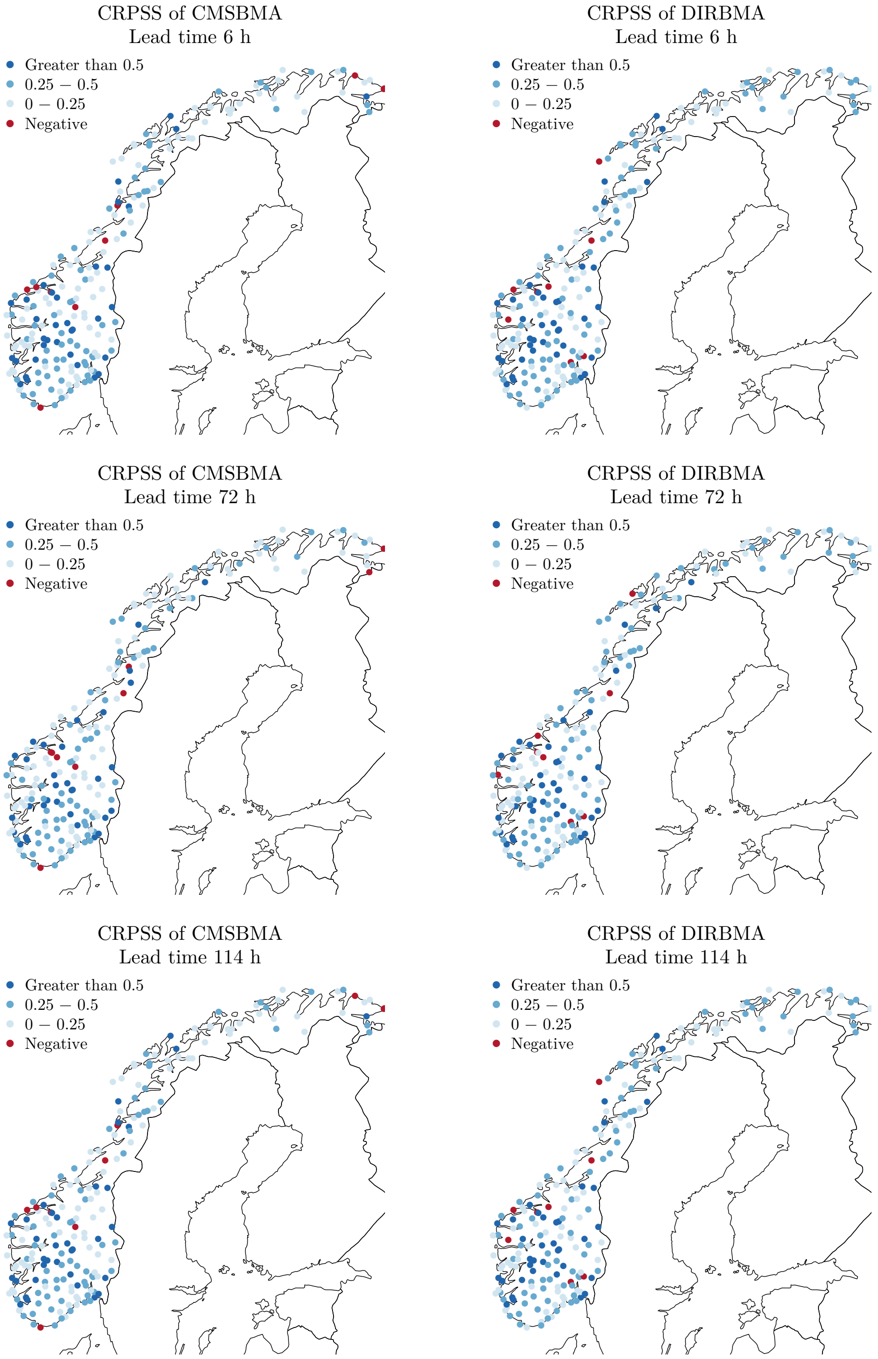


Figure 5.8: The CRPSS of probabilistic forecasts from CMSBMA and DIRBMA relative to raw ensemble forecasts at different sites for different lead times color coded and plotted on a map.

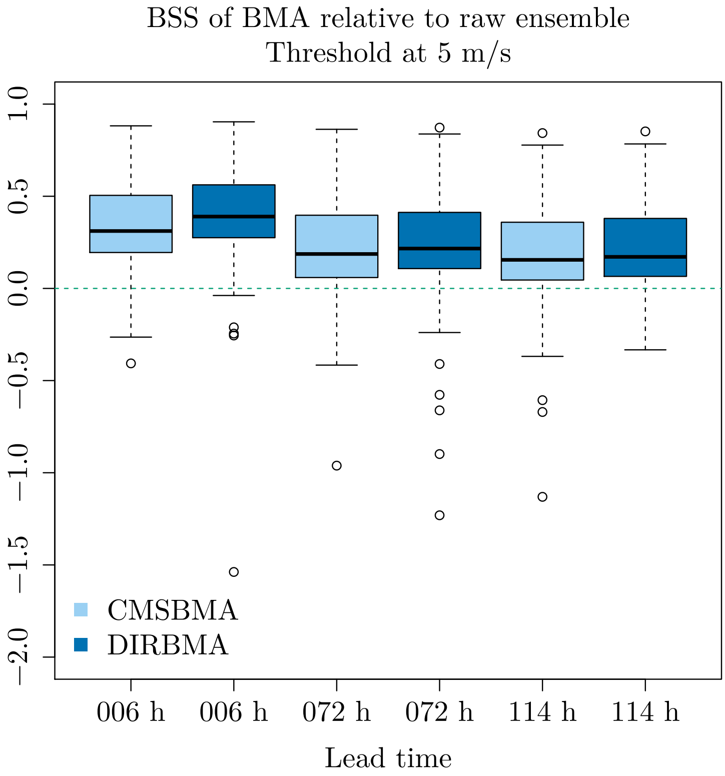


Figure 5.9: Box-and-whisker plots of the Brier skill score of probabilistic forecasts from CMSBMA and DIRBMA relative to raw ensemble forecasts where a threshold of 5 m/s has been used. Each data point is the BSS of a model for a specific site and lead time.

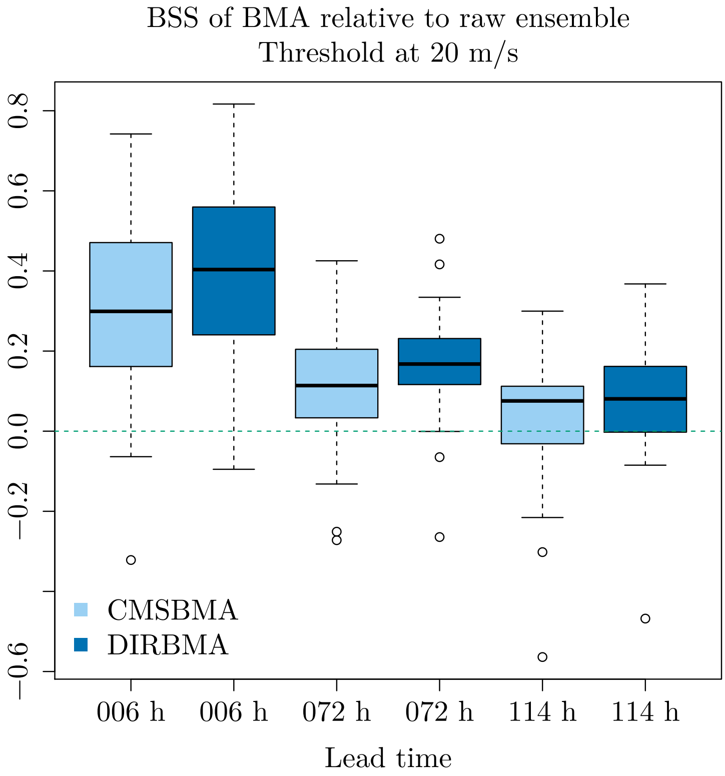


Figure 5.10: Box-and-whisker plots of the Brier skill score of probabilistic forecasts from CMSBMA and DIRBMA relative to raw ensemble forecasts where a threshold of 20 m/s has been used. Each data point is the BSS of a model for a specific site and lead time. Only sites with more than 10 observations of wind speeds greater than 20 m/s are used.

h the use of BMA seems to have less of an effect, but still generally leads to substantial improvements on the forecasts of strong wind.

Figure 5.11 shows the relative improvement in quantile score for the 0.95-quantile for DIRBMA compared to CMSBMA. Again there is an improvement at most sites, indicating that the use of forecast wind direction in the estimation of the expectation has a positive effect on the upper quantiles of the forecasts in most of the cases where CMSBMA might fall short.

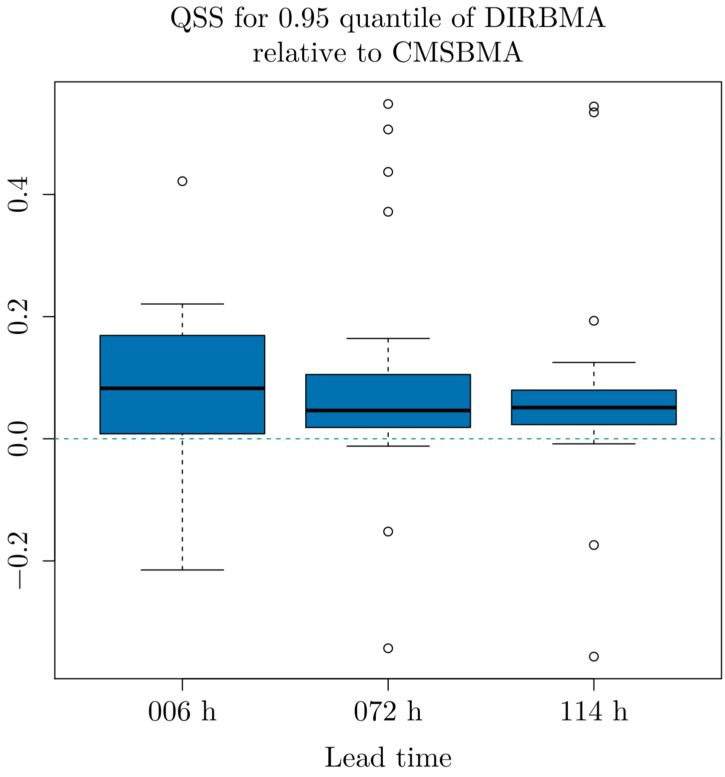


Figure 5.11: Box-and-whisker plots of the quantile skill score for the 0.95-quantile of probabilistic forecasts modelled using the DIRBMA model relative to probabilistic forecasts from the CMSBMA model. Each data point is the QSS of a model for a specific site and lead time. Only sites with more than 10 observations of wind speeds greater than 20 m/s are used.

6. Discussion and conclusion

In this thesis the postprocessing technique BMA has been applied to ensemble forecasts of wind speed and direction in order to obtain calibrated and sharp probabilistic forecasts for 375 days and 204 locations in Norway with 3 different lead times, $l = 6, 72, 114$.

The forecasts that were used came from the 51 member ECMWF ensemble, the first member of which – called the control member – has initial conditions that are the best estimate of the atmospheric state. The other members are perturbed around it in attempt to model the atmospheric uncertainty. This means that the control member, on average, produces better forecasts than the perturbed members. For this reason tests were done in which the control member was treated separately in the BMA. It was shown that doing this, and treating the other members as exchangeable, achieved better results than treating all 51 members as exchangeable.

Different numbers of training days were tested, as well as different frequencies of model fitting, and a training period of 376 days was chosen, as results revealed that a long training period was essential to fitting a good model.

The use of BMA generally had a very positive effect and led to improvement at most locations, and for all three lead times. However, a weakness was identified for longer lead times when dividing the probabilistic forecasts into groups according to forecast median. Although the models performed well in general, they struggled with forecasting stronger winds.

Several ideas were tested in order to try to remedy this weakness. Firstly, after applying the method to the 51 member ensemble, models were tested where climatology and historical observations respectively were included as separate ensembles. The inclusion of these data in the model had no discernible effect.

Secondly, in addition to the standard BMA, a modified version was used in which forecast wind direction was included in the modelling of the expectation of the predictive PDFs by means of thin plate regression splines. This method produced better models at most locations, and also to some degree amended the problem with forecasting high wind speeds.

At some locations the extended method had little or even a negative effect on the forecasts. This seemed to be linked in part to the range of wind speeds observed at these sites. That is, the sites where the method had no positive effect tended to be sites where particularly strong winds are rarely observed.

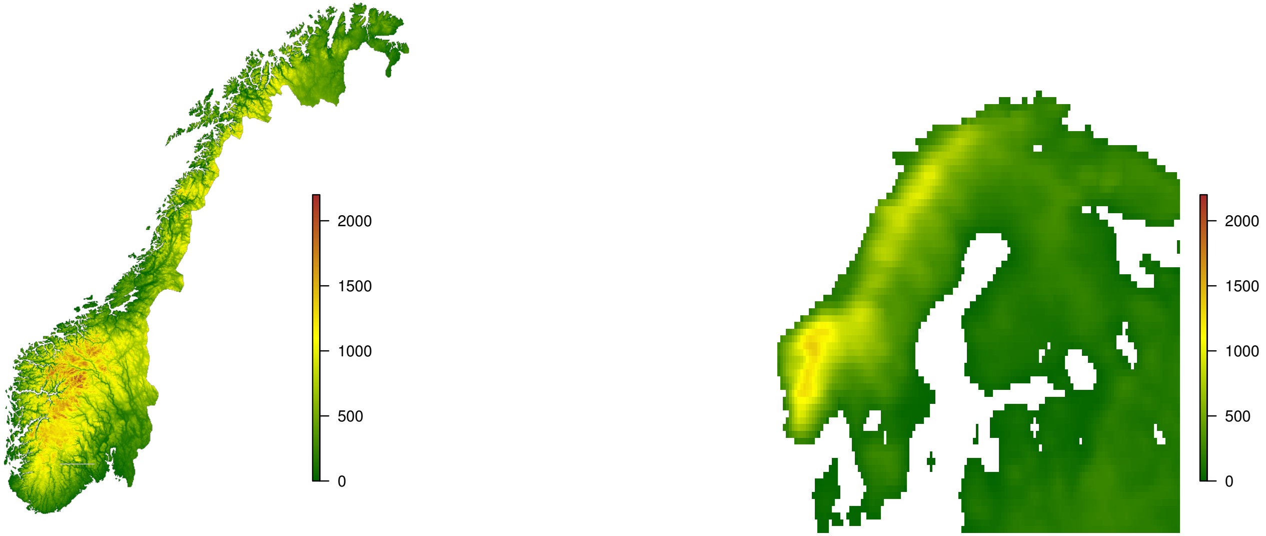
As no observations of wind direction were available, it is hard to tell if the shortcomings in these models were due to faulty forecasts or perhaps a lack of relationship between wind direction and wind speed for lower wind speeds. The raw forecasts of wind speed were biased and underdispersed, but the relationship between them and the observations was carefully examined and found to be linear before the forecasts were used in the fitting of any model. The relationship between the forecast wind direction and observed wind speed was not examined as thoroughly as it could have been, and validating observations might have provided additional insight into the quality of the wind direction forecasts.

Modelling wind in Norway is challenging as the country has an extremely varied and complex topography. Certain areas are dominated by high mountains, with narrow fjords and valleys that strongly affect both the wind speed and wind direction that are observed. These small scale features are not well represented in the ECMWF ensemble, which has a spatial resolution of approximately 32 km (Miller et al., 2010), meaning that within each 32 km grid square the topography is constant.

Figure 6.1 shows two topography maps of Norway. The one in Figure 6.1a has a spatial resolution of 100 m, while that in Figure 6.1b is the model topography used in the ECMWF ensemble. Clearly the latter is much coarser and does not does not include local topographical and coastal variations visible in the former. The difference is particularly striking in the west around the fjords and around the valleys stretching across the country.

Figure 3.3, in Section 3.2, showed that the sites that see the strongest winds are located along the coast and in areas with mountains, in other words, areas with complex topography. Furthermore, in Section 5.2, it was shown the BASICBMA model exhibited a negative bias for especially strong winds at many of these sites. This suggests that the model topography presented in Figure 6.1b is not sufficiently detailed, and that a higher resolution is likely to lead to better results in general.

As it stands, when the model wind is in a specific direction, this may be strongly correlated to high wind speeds in certain valleys (McNider and Pielke, 1984). For example, westerly winds from the ocean that are well forecast by the ECMWF ensemble can enter valleys that the ensemble,



(a) Map showing the topography of Norway, with a 100 m resolution, taken from kartverket.no, the website of Norway's national mapping agency.

(b) Map showing the model topography of Norway used by the ECMWF ensemble. (Approximately 32 km resolution.)

Figure 6.1: Topography maps of Norway in metres above sea level.

because of its low resolution, does not detect, and be accelerated. While this acceleration might not be well represented in the wind speed forecasts, the inclusion of forecast wind direction in the BMA could have a huge effect on the resulting forecasts.

In the case of Ytterøyane fyr, Figure 3.6 showed that the strongest winds are from the south. This could be a result of topographic blocking from the Norwegian mainland. Smith (1982) showed that when wind hits a mountain barrier, it is deflected to the left (in the northern hemisphere) and accelerated. Along the coast of Norway, this means that west of a mountain barrier, southerly winds will be observed. The ECMWF ensemble might be able to model the deflection but is unlikely to be able to model the acceleration, resulting in wind direction forecasts that are correct but wind speed forecasts that underestimate the actual wind speed. Also in these cases the inclusion of forecast wind direction in the BMA is likely to have a positive effect.

Conversely, in areas with unvaried topography, wind direction is unlikely to be related to wind speed and thus including forecast wind direction in the BMA will have less of an effect.

This being said, the main findings and conclusions of this master's thesis can be summarized as follows:

1. The control member should be considered unique and have its expectation modelled separately from the remaining ensemble members.

2. The use of climatology and historical observations does not appear to have any effect on forecast performance.

3. Forecast wind direction has been shown to be a good predictor for wind speed, especially in cases of high observed wind speeds.

Since the use of wind direction in the BMA had an overall positive effect, possible future amendments to the method would therefore involve examining the wind direction forecasts more closely before using them, and studying how their relationship with wind speed changes for different locations. Perhaps other spatial information pertaining to the individual sites, such as altitude and land area fraction, might shed more light on this.

Another method that could be worth investigating would be cyclic splines, i.e. splines that are specifically designed to handle cyclic variables. The choice of thin plate regression splines for the modelling of the expectation of the predictive PDFs of the BMA was by no means an obvious one, and there is no reason why other types of spline should not also be tested.

A further possibility that was considered, but not explored in this thesis was a varying-coefficient model (Hastie and Tibshirani, 1993) for the expectation, with regression coefficients varying with forecast wind direction.

To conclude, this thesis has shown how BMA can be a useful tool in forecasting wind speed, and in particular that forecast wind direction can be an important predictor. However, there are grounds for more research into methods and predictors that might further improve the forecasts of wind speed in Norway.

References

Baran, S. and Lerch, S. (2015a). Log-normal distribution based emos models for probabilistic wind speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, 141.

Baran, S. and Lerch, S. (2015b). Mixture emos model for calibrating ensemble forecasts of wind speed.

BBC (2005). 'It really is quite safe'. `"http://news.bbc.co.uk/2/hi/uk/4576351.stm"`.

Belongie, S. (2000). Thin plate spline. From *MathWorld*–A Wolfram Web Resource, created by Eric W. Weisstein. `http://mathworld.wolfram.com/ThinPlateSpline.html`.

Bentzien, S. and Friederichs, P. (2014a). Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140:1924–1934.

Bentzien, S. and Friederichs, P. (2014b). The quantile score and its decomposition. Vienna, Austria.

Bouallègue, Z. B., Pinson, P., and Friederichs, P. (2015). Quantile forecast discrimination ability and value. *Quarterly Journal of the Royal Meteorological Society*, 141:3415–3424.

Duchon, J. (1977). Splines minimizing rotation invariant semi-norms in sobolev spaces. In Schempp, P. D. W. and Zeller, P. D. K., editors, *Lecture Notes in Math: Constructive Theory of Functions of Several Variables, Oberwolfach 1976*, volume 571. Springer.

Feldmann, K. (2012). *Statistical Postprocessing of Ensemble Forecasts for Temperature: The Importance of Spatial Modeling.* PhD thesis, Ruprecht-Karls-Universitat Heidelberg.

Fraley, C., Raftery, A. E., and Gneiting, T. (2010). Calibrating multi-model forecast ensembles with exchangeable and missing members using bayesian model averaging. *Monthly Weather Review*, 138.

Fraley, C., Raftery, A. E., Sloughter, J. M., Gneiting, T., and University of Washington. (2015). *ensembleBMA: Probabilistic Forecasting using Ensembles and Bayesian Model Averaging.* R package version 5.1.1.

Gneiting, T. (2014). Calibration of medium-range weather forecasts. *Technical Memorandum*, No. 719.

Gneiting, T., Raftery, A. E., Westveld, III, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133:1098–1118.

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422.

Grimit, E. P., Gneiting, T., Berrocal, V. J., and Johnson, N. A. (2007). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132:2925–2942.

Hamill, T. M. (2000). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. 55(4):757–796.

Kalnay, E. (2003). *Atmospheric modeling, data assimilation, and predictability.* Cambridge University Press, Cambridge New York.

Leith, C. E. (1974). Theoretical skill of monte carlo forecasts. *Monthly Weather Review*, 102.

McNider, R. T. and Pielke, R. A. (1984). Numerical simulation of slope and mountain flows. 23(10):1441–1453.

Miller, M., Buizza, R., Haseler, J., Hortal, M., Janssen, P., and Untch, A. (2010). Increased resolution in the ecmwf deterministic and ensemble prediction systems. Technical Report 124, ECMWF.

Pidwirny, M. (1999). 7(n). Forces acting to create wind. `"http://www.physicalgeography.net/fundamentals/7n.html"`.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review.*

Sloughter, J. M., Gneiting, T., and Raftery, A. E. (2010). Probabilistic wind speed forecasting using ensembles and bayesian model averaging. *Journal of the American Statistical Association*, 105(489).

Sloughter, J. M., Raftery, A. E., Gneiting, T., and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using bayesian model averaging. *Monthly Weather Review*, 135.

Smith, R. B. (1982). Synoptic observations and theory of orographically disturbed wind and pressure. *J. Atmos. Sci.*, 39(1):60–70.

Talagrand, O., Vautard, R., and Strauss, B. (1997). Evaluation of probabilistic prediction systems. In *Proceedings, ECMWF Workshop on Predictability*, Reading, UK.

Yr NRK (2013). Slik utnytter du langtidsvarslene best. `"http://om.yr.no/forklaring/forsta-varslene/langtidsvarsel/"`.

Thorarinsdottir, T. L. and Gneiting, T. (2010). Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A*, 173.

Thorarinsdottir, T. L. and Johnson, M. S. (2012). Probabilistic wind gust forecasting using nonhomogeneous gaussian regression. *Monthly Weather Review*, 140:889–897.

Tuller, S. E. and Brett, A. C. (1984). The characteristics of wind velocity that favor the fitting of a weibull distribution in wind speed analysis. *Journal of climate and applied meteorology*, 23.

ECMWF (2016). Medium-range forecasts. `http://www.ecmwf.int/en/forecasts/documentation-and-support/medium-range-forecasts`.

NOAA-NCEI (U.S. Dept. of Commerce) (2016). Numerical weather prediction. `https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/numerical-weather-prediction`.

SAS Institute Inc. (2015). *SAS/STAT® 14.1 User's Guide: High-Performance Procedures.* Cary, NC: SAS Institute Inc.

Wahba, G. (1990). *Spline models for observational data.* Society for Industrial and Applied Mathematics, Philadelphia, Pa.

Wilks, D. S. (2006). *Statistical Methods in the Atmospheric Sciences.* Academic Press, second edition.

Wilson, L. and Nurmi, P. (2011). Skill scores. `http://www.eumetcal.org/resources/ukmeteocal/verification/www/english/msg/ver_cont_var/uos5/uos5_ko1.htm`.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65:95–114.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC Press.

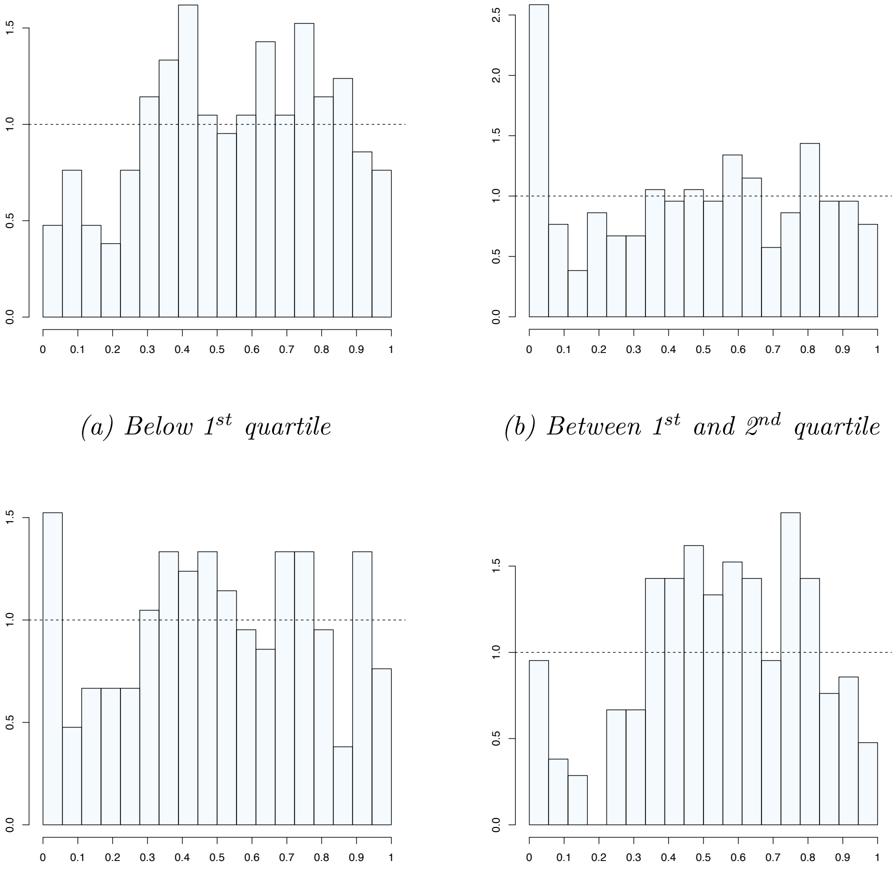
Appendices

Appendix A Function for fitting a BMA

```r
fit.bma <- function(forecast, y, direction = NULL, varCoefs = c(1,1),
                   exchangeable = NULL, oneBias = F){
  forecast <- as.matrix(forecast)
  M  <- ncol(forecast)
  n  <- length(y)

  # For observed winds == 0, add a small positive noise.
  y[y == 0] <- rgamma(sum(y==0), .1)

  if(!is.null(exchangeable) & length(exchangeable) != M){
    stop("Length of exchangeable does not match
         total number of members.")}
  if(is.null(exchangeable)) exchangeable <- 1:M
  uniqueX  <- unique(exchangeable)
  nUniqueX <- length(uniqueX)
  w        <- rep(1/M, M)

  ########### Estimate biasCoefs ###########
  # Standard method, no direction:
  if(is.null(direction)){
    cat("Estimating bias coefficients.")
    meth <- 1

    biasCoefs <- matrix(NA, nrow = 2, ncol = nUniqueX)
    mu <- matrix(NA, nrow = n, ncol = M)
    it <- 0
    for(ex in uniqueX){
      it <- it + 1
      x  <- as.vector(as.matrix(forecast[,exchangeable == ex]))

      biasCoefs[,it] <- as.numeric(lm(rep(y,
                                          sum(exchangeable == ex))
                                      ~ x)$coef)

      if (biasCoefs[1, it] <= 0) {
        biasCoefs[1, it] <- min(y)
        biasCoefs[2, it] <- sum((rep(y, sum(exchangeable == ex)) -
                                   min(y)) * x)/sum(x^2)
      }

      mu[, exchangeable == ex] <- biasCoefs[1,it] +
                                  biasCoefs[2,it] *
                                  forecast[,exchangeable == ex]

      cat(".")
    }
    rm(it)
  }

  # If direction is provided:
  if(!is.null(direction)){
    require(mgcv)
    cat("Estimating expectation.")
    if(nrow(forecast) == nrow(direction) &
         ncol(forecast) == ncol(direction)){
      meth       <- 2
      biasModels <- list()
      mu         <- matrix(NA, nrow = n, ncol = M)
      it         <- 0

      for(ex in uniqueX){
        it  <- it + 1
        dat <- data.frame(obs = rep(y, sum(exchangeable == ex)),
                          forec = as.vector(
                                  as.matrix(
                                    forecast[,exchangeable == ex])),
                          direc = as.vector(
                                  as.matrix(
                                    direction[,exchangeable == ex])))

        # In order to get similar values for 0 and 360 degrees
        # (direction North), some observations are copied and pasted
        # such that the model is fitted for directions in [-90,450]
        augment    <- function(u) {
          ul       <- u[u$direc < 90, ]
          ul$direc <- ul$direc + 360
          uu       <- u[u$direc > 270, ]
          uu$direc <- uu$direc - 360
          return(rbind(u, ul, uu))
        }
        dat                      <- augment(dat)
        biasModels[[it]]         <- mgcv::gam(obs ~ s(forec, direc),
                                             data=dat)
        mu[, exchangeable == ex] <- fitted(biasModels[[it]])[1:
                                    length(rep(y,
                                           sum(exchangeable == ex)))]

       cat(".")
      }
      rm(it)

    }
  }

  # Both methods risk fitting negative values of mu
  if(any(mu < 0)) mu[mu<0] <- runif(sum(mu<0), max = .1)
  cat(".done.\n")

  ########### Estimate varCoefs ###########
  cat("Estimating variance coefficients")
  getParam <- function(mu, varCoefs, forecast){
    sigma  <- varCoefs[1] + varCoefs[2]*forecast
    shape  <- as.matrix(mu^2/sigma^2)
    scale  <- as.matrix(sigma^2/mu)
    return(list(shape = shape, scale = scale))
  }

  tol      <- sqrt(.Machine$double.eps)
  d        <- 10
  ll_new   <- -Inf
  param    <- getParam(mu, varCoefs, forecast)
```

```
shape    <- param$shape
scale    <- param$scale

# Iterative fitting of variance coefficients
count    <- 0
while(d > tol){
  count  <- count + 1

  ## E-step
  nevner <- rowSums(t(w*t(dgamma(y, shape = shape, scale = scale))))
  z      <- t(w*t(dgamma(y, shape = shape, scale = scale)))/nevner

  ## CM-1
  w      <- colMeans(z)
  w      <- sapply(split(w, exchangeable), mean)[exchangeable]

  if(!count %% 50 | nUniqueX == 1){

    ## CM-2
    loglik = function(v, y, forecast, mu)
    {
      stopifnot(length(v) == 2)
      sigma <- (v[1])^2 + (v[2])^2*forecast
      shape <- mu^2/sigma^2
      scale <- as.matrix(sigma^2/mu)

      ll    <- sum(z*log(t(w*t(dgamma(y, shape = shape,
                                      scale = scale, log = FALSE))))
  )
      return(ll)
    }

    result   = optim(sqrt(varCoefs),
                     fn         = loglik,
                     method     = if(meth == 2){"BFGS"},
                     control    = list(fnscale = -1),
                     y          = y,
                     forecast   = forecast,
                     mu         = mu)

    varCoefs <- result$par^2

    param    <- getParam(mu, varCoefs, forecast)
    shape    <- param$shape
    scale    <- param$scale

    # Update loglikelihood:
    ll_new   <- result$value
    d        <- abs(ll_old-ll_new)/(1+abs(ll_new))
    ll_old   <- ll_new
    cat(".")
   }
}
cat("done.\n")

if(meth == 1){
  coefs <- list(biasCoefs      = as.matrix(biasCoefs),
                  varCoefs      = as.matrix(varCoefs),
                  weights       = w,
                  exchangeable  = exchangeable,
                  loglikelihood = ll_new,
                  nIter         = count,
                  method        = 1)
  }

  if(meth == 2){
    coefs <- list(biasModels    = biasModels,
                  varCoefs      = as.matrix(varCoefs),
                  weights       = w,
                  exchangeable  = exchangeable,
                  loglikelihood = ll_new,
                  nIter         = count,
                  method        = 2)
  }

  return(coefs)
}
```

Appendix B PIT histograms for other lead times



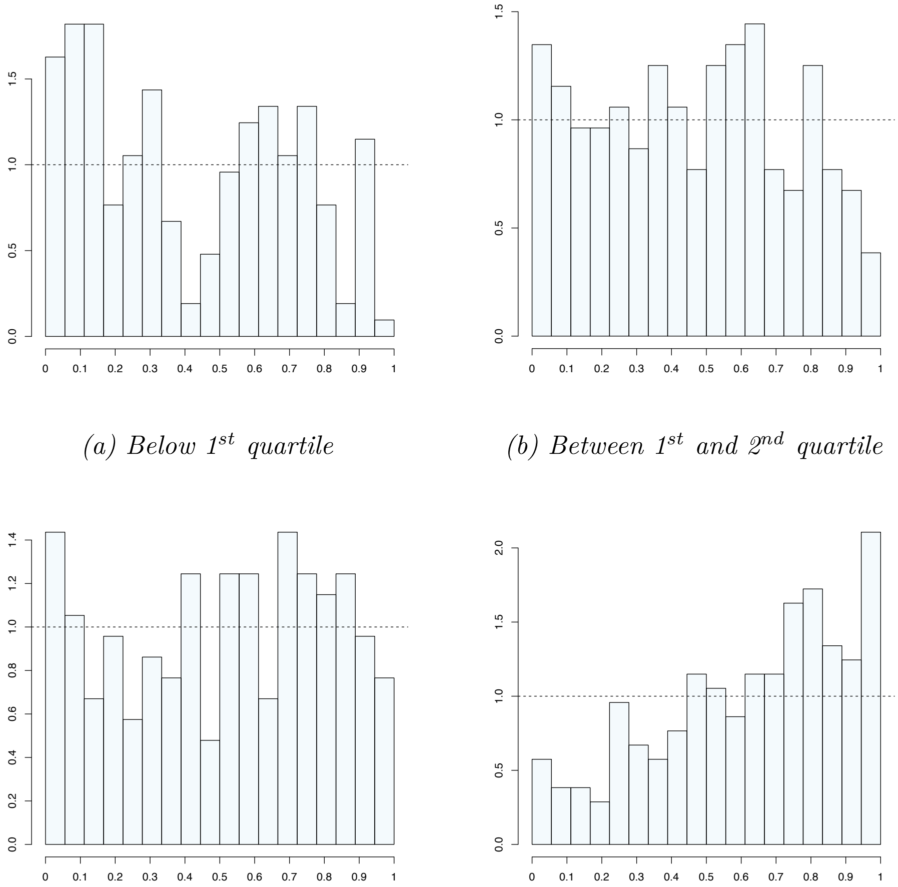
(a) Below 1st quartile

(b) Between 1st and 2nd quartile

(c) Between 2nd and 3rd quartile

(d) Above 3rd quartile

Figure B.1: PIT histograms conditioning on the median of the predictive PDFs. Lead time $l = 6$ *h and site Ytterøyane fyr was used.*



(a) Below 1st quartile

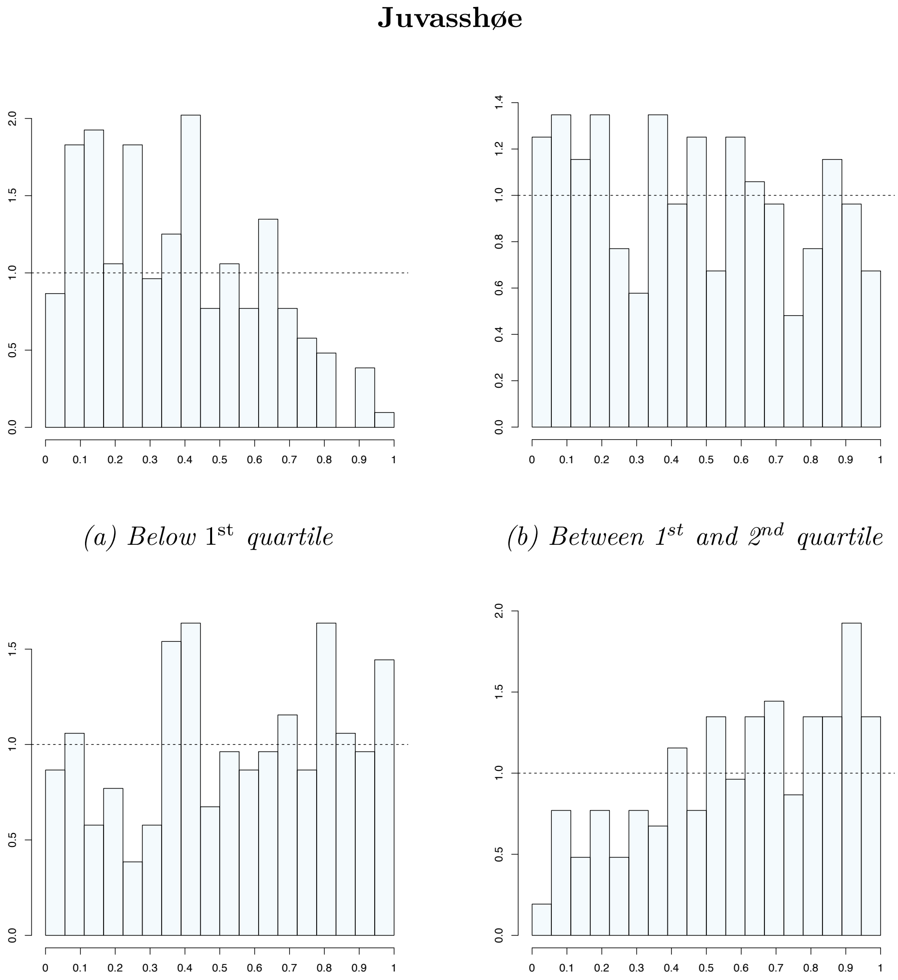
(b) Between 1st and 2nd quartile

(c) Between 2nd and 3rd quartile

(d) Above 3rd quartile

Figure B.2: PIT histograms conditioning on the median of the predictive PDFs. Lead time $l = 114$ h and site Ytterøyane fyr was used.

Appendix C PIT histograms for other sites



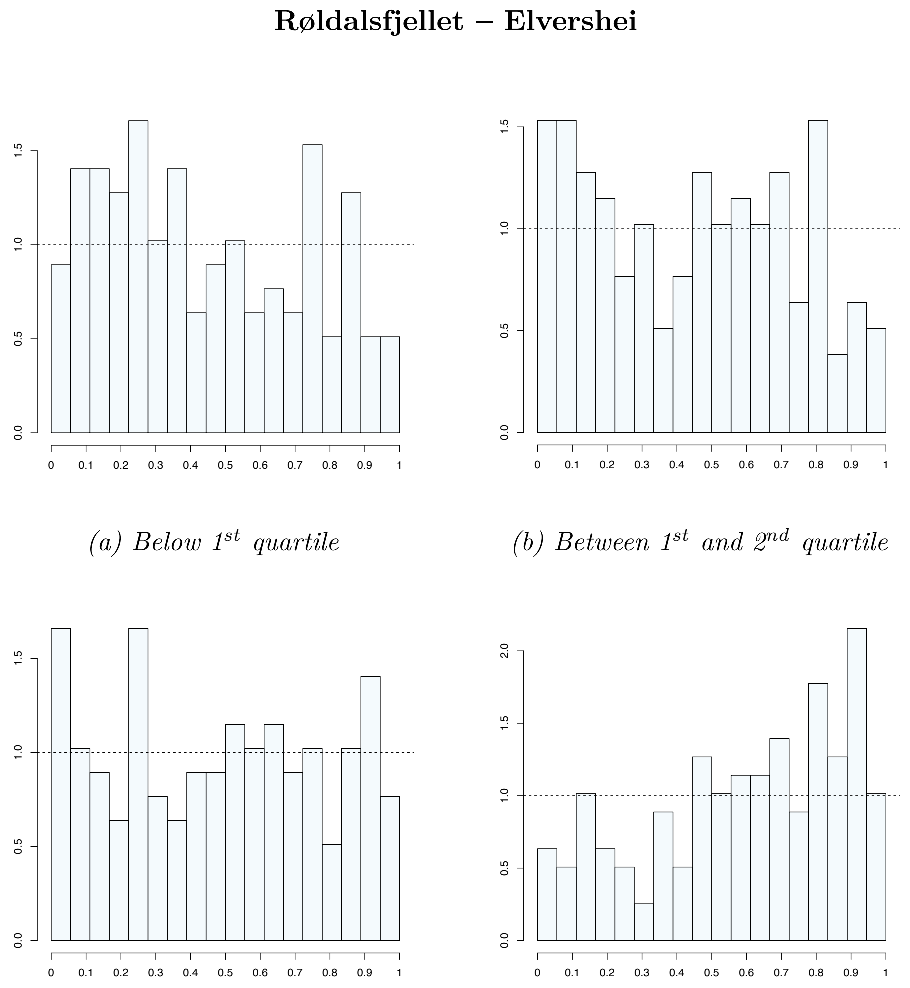
(a) Below 1^{st} *quartile*

(b) Between 1^{st} *and* 2^{nd} *quartile*

(c) Between 2^{nd} *and* 3^{rd} *quartile*

(d) Above 3^{rd} *quartile*

Figure C.1: PIT histograms conditioning on the median of the predictive PDFs. Lead time $l = 72$ *h and site Juvvasshøe.*



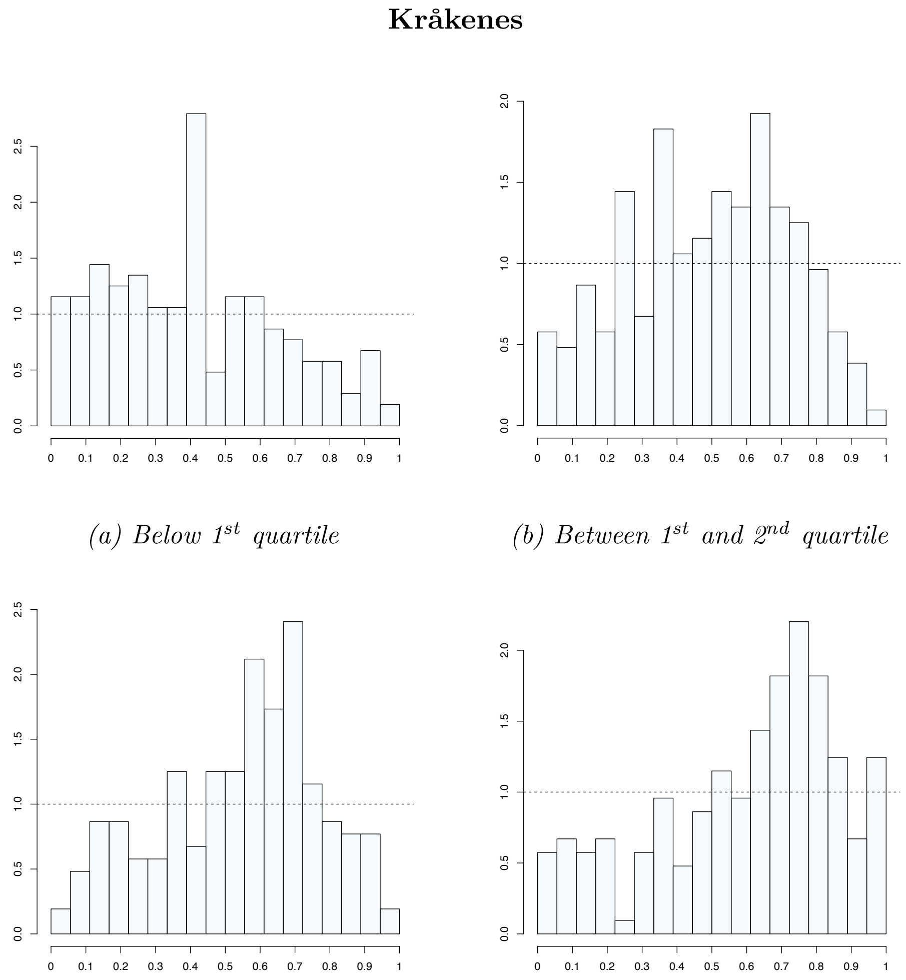
(a) Below 1st quartile

(b) Between 1st and 2nd quartile

(c) Between 2nd and 3rd quartile

(d) Above 3rd quartile

Figure C.2: PIT histograms conditioning on the median of the predictive PDFs. Lead time $l = 72$ h and site Røldalsfjellet – Elvershei.



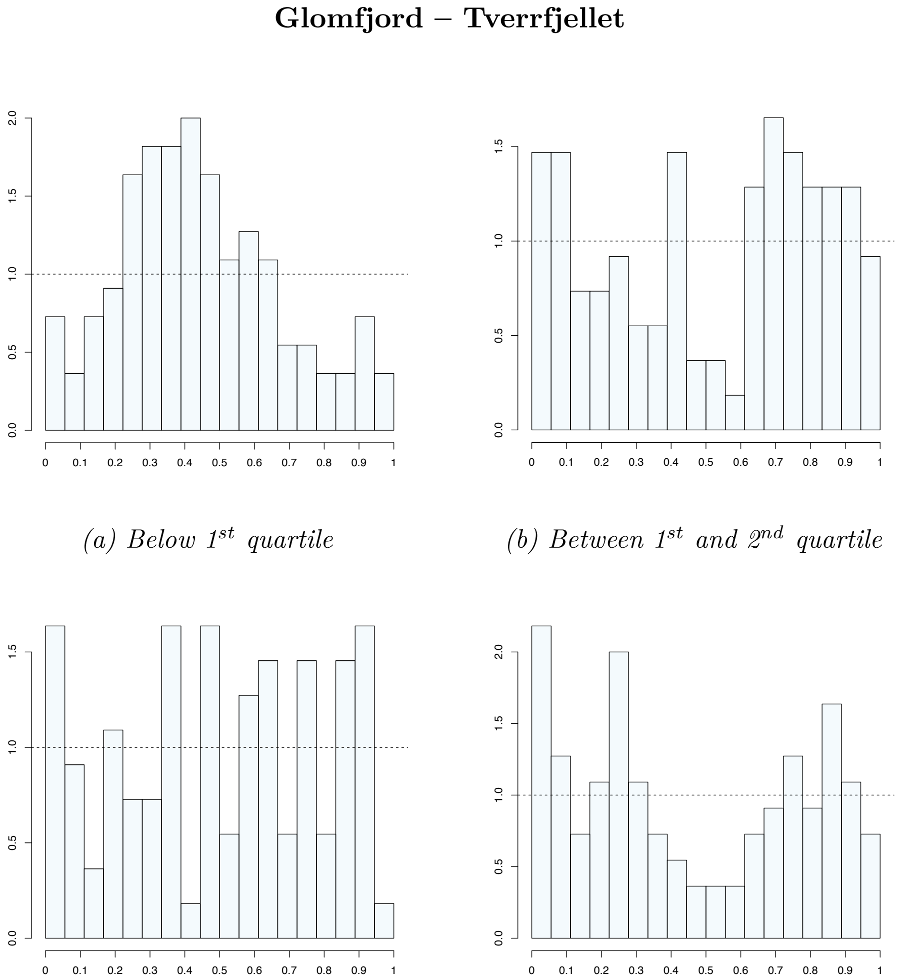
(a) Below 1st quartile

(b) Between 1st and 2nd quartile

(c) Between 2nd and 3rd quartile

(d) Above 3rd quartile

Figure C.3: PIT histograms conditioning on the median of the predictive PDFs. Lead time $l = 72$ h and site Kråkenes.



(a) Below 1st quartile

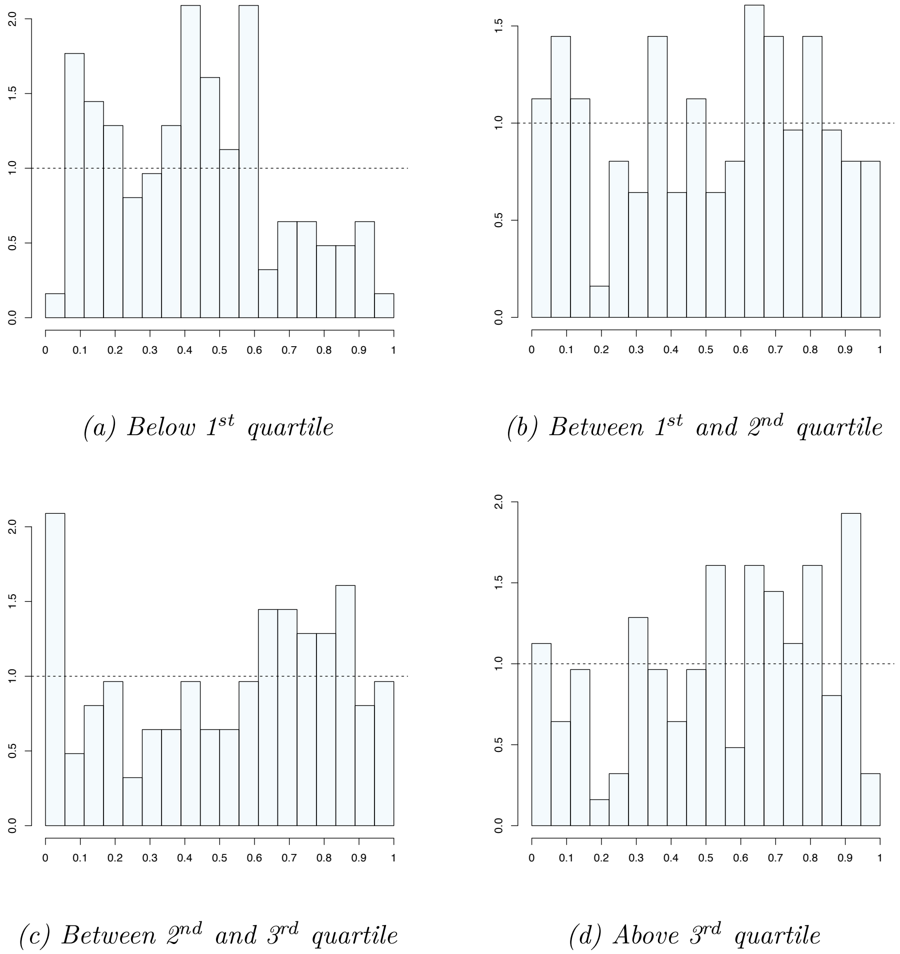
(b) Between 1st and 2nd quartile

(c) Between 2nd and 3rd quartile

(d) Above 3rd quartile

Figure C.4: PIT histograms conditioning on the median of the predictive PDFs. Lead time $l = 72$ h and site Glomfjord – Tverrfjellet.

Narvik – Fagernesfjellet



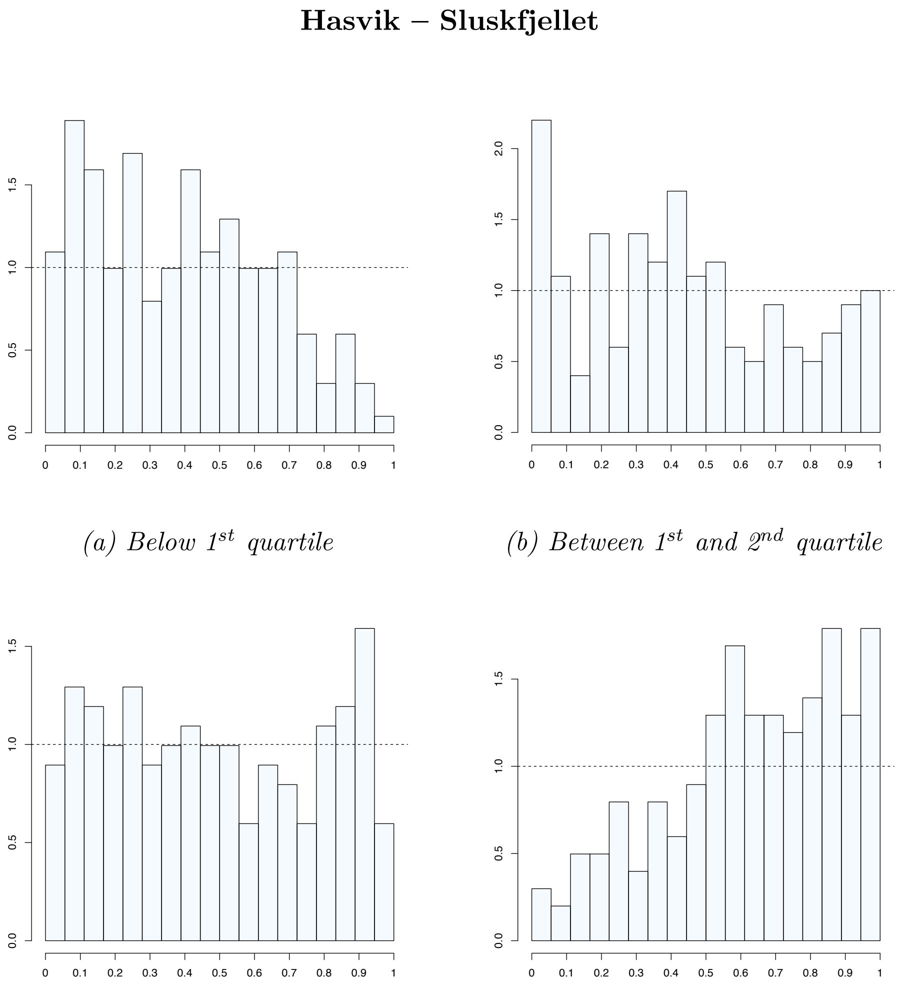
(a) Below 1st quartile

(b) Between 1st and 2nd quartile

(c) Between 2nd and 3rd quartile

(d) Above 3rd quartile

Figure C.5: PIT histograms conditioning on the median of the predictive PDFs. Lead time $l = 72$ h and site Narvik – Fagernesfjellet.



(a) Below 1st quartile

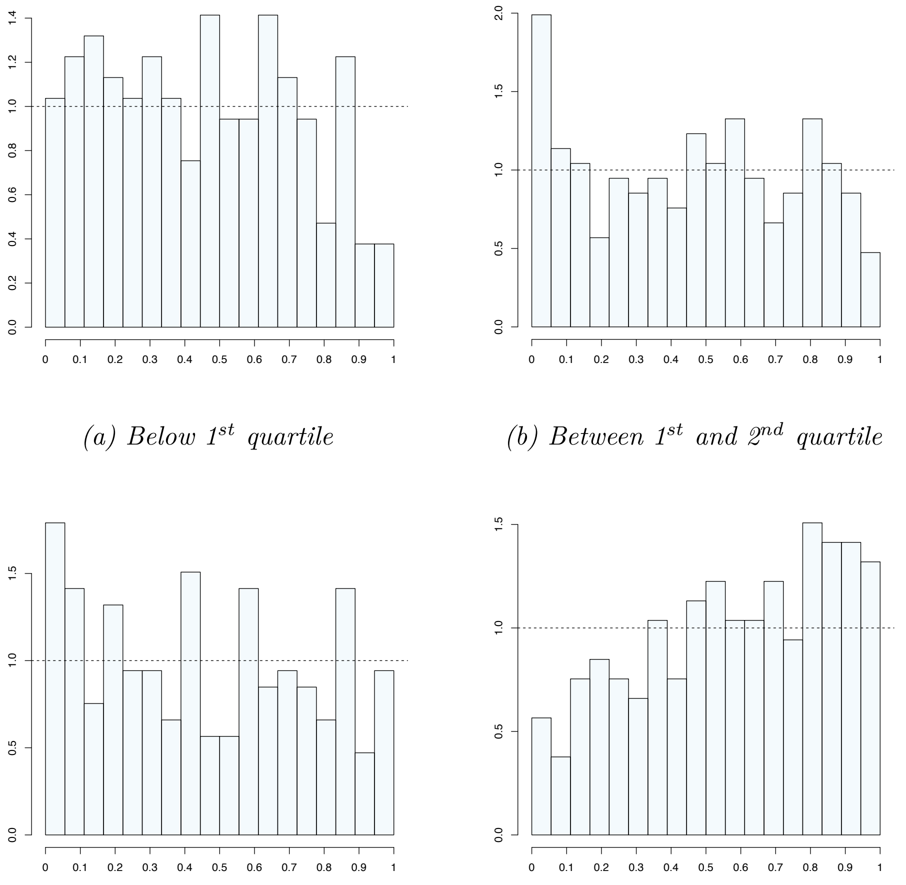
(b) Between 1st and 2nd quartile

(c) Between 2nd and 3rd quartile

(d) Above 3rd quartile

Figure C.6: PIT histograms conditioning on the median of the predictive PDFs. Lead time $l = 72$ h and site Hasvik – Sluskfjellet.

Hammerfest lufthavn



(a) Below 1st quartile

(b) Between 1st and 2nd quartile

(c) Between 2nd and 3rd quartile

(d) Above 3rd quartile

Figure C.7: PIT histograms conditioning on the median of the predictive PDFs. Lead time $l = 72$ h and site Hammerfest lufthavn.