NTNU – Trondheim
Norwegian University of
Science and Technology

Bivariate Bayesian Model Averaging and Ensemble Model Output Statistics

With a Case Study of Ensemble Temperature Forecasts in Trondheim

Jorinde Prokosch

Master of Science in Physics and Mathematics
Submission date: June 2013
Supervisor: Ingelin Steinsland, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Abstract

In this study a bivariate Bayesian model averaging (BMA) and Ensemble model output statistics (EMOS) technique for ensemble temperature forecasts are proposed to account for lead time dependencies between errors. Also univariate BMA and EMOS techniques are applied to generate calibrated normal predictive density functions. For univariate models, Maximum likelihood estimation (MLE) and minimum Continuous rank probability score (minCRPS) estimation are compared. In addition to the MLE, a sample method to simplify the minimum Energy score (minES) estimation is proposed for bivariate models. In a case study of 2-m surface temperature in Trondheim – Voll between year 2007 and 2011, using the European Center for Medium-Range Weather Forecasts (ECMWF) forecast ensembles, the BMA technique using minCRPS estimation shows the most calibrated and sharpest post-processed probabilistic forecasts. The bivariate EMOS model using minES estimation gives the best score and shows that there is lead time dependencies between errors.

Sammendrag

I dette studiet foreslår vi bivariate versjoner av de to teknikkene Bayesian model averaging (BMA) og Ensemble model output statistics (EMOS) for ensemble-temperaturvarsler, der vi har tatt hensyn til ledetid-avhengighet i feilen. Vi bruker også univariate BMA og EMOS teknikker for å generere kalibrerte normalfordelte prediktive tetthetsfunksjoner. For univariate modeller sammenligner vi metodene Maximum likelihood estimering (MLE) og minimum Continuous rank probability score (minCRPS). I tillegg til at vi bruker MLE, foreslår vi en sampling-metode for å forenkle minimum Energy score (minES) estimeringen for bivariate modeller. I et casestudie ser vi på temperatur i Trondheim – Voll for årene 2007 – 2011, der vi bruker European Center for Medium-Range Weather Forecasts (ECMWF) sine varsel-ensembler. Her viser vi at BMA teknikken med minCRPS estimering gir de best kalibrerte og skarpeste post-prosesserte sannsynlighetsvarslene. Den bivariate EMOS modellen med minES estimering gir den beste scoren og viser at det er ledetid-avhengighet i feilen.

Preface

This report is the result of the course TMA4500 Master's thesis for the program Industrial Mathematics at the Norwegian University of Science and Technology (NTNU), Department of Mathematical Science. This work was carried out in the spring of 2013.

The work of this Master's thesis has been a continuation of my Specialization project from the fall of 2012. For completeness, some of the theoretical background are therefore included in this thesis. The software MATLAB was used for the analysis.

I would especially like to thank my supervisor, Professor Ingelin Steinsland who has helped and supported me throughout this master thesis. She always used her time for discussions and provided me with invaluable guidance which has contributed much in this master.

Jorinde Prokosch
Trondheim, June 2013

Contents

Glossary of notations

Notation	Meaning
t	Time for which a forecast is issued.
l	Lead time.
n	One given day.
y_{t+l}	Observational data at issue time t and lead time l.
$x_{m,t,l}$	Ensemble member number m at issue time t and lead time l.
$\bar{x}_{t,l}$	Ensemble mean at issue time t and lead time l.
$y_{t-k,l}$	Training data at issue time $t-k$ and lead time l
w	Weight.
T	Total number of times a forecast has been issued for one lead time, l.
M	Total number of members of an ensemble forecast.
K	Total number of training days.
N	Total number of days.
$p(.)$	Probability density function (pdf).
$F(.)$	Cumulative distribution function (cdf).
η	Mean.
μ	Predictive mean.
τ^2	Variance.
σ^2	Predictive variance.

Table 1: *Summary of notations used in this study.*

1. Introduction

Will you need an umbrella tomorrow? Can you go windsurfing next weekend? It has always been of great interest to forecast the weather. Weather forecasts are of importance in that for example airlines get information about the weather conditions in order to schedule flights, farmers can plan the planting and harvesting of their crops, and electricity suppliers can make decisions related to electricity pricing. However, how good is today's forecast likely to be? Since the early 1990s, probabilistic forecasts have been increasingly used for weather predictions at many weather centres [22]. Probabilistic forecasts provide a probability distribution and give an estimate for how accurate the forecast is. Deterministic forecasts, on the other hand, do not account for risk or uncertainty. Still, would you bring an umbrella if the forecasts says that there is a 30% chance of rain today? Should we warn a city if there is a 5% chance for storm to occur tomorrow? Probabilistic forecasts are of importance to better understand and evaluate uncertainty when making decisions.

Current research have shown that including spatial dependencies between different observation sites give significantly better performance than univariate models [3]. Also bivariate models for wind vectors, where the spatial relationship between the components is taken into account, have shown improvements [27, 28]. However, to our knowledge, there has been few studies which accounts for lead time dependencies between errors. A lead time is the delay between the issue time of e.g. a weather forecast and the time for when the weather forecast says it will occur. An issue time is the time when the forecasts are made. Nowadays, lead times are considered to be time independent when forecasting a weather event. However, according to Palmer [24] many economical losses could be saved in making better decisions related to the weather. Forecasts without time dependency can, for example, lead to forecasting smaller chances for frost

tomorrow than there actually are. This can have huge economical consequences for e.g. winter road maintenance. Between 1992 and 2004 Snoqualmie Pass on Interstate Highway 90 (I-90) was closed 120 hours per year on average because of road maintenance, resulting in an annual loss of at least 17.5 million dollars [4]. This motivates us to develop a model that accounts for lead time dependencies between errors.

Different forecasts are made from different forecasters. However, which forecaster should we trust? The forecaster saying that there is a 30% chance for rain tomorrow? Or the one which says that the chance for rain tomorrow is only 5%? According to Gneiting et al. [12], the goal is to make the probability forecasts as sharp as possible subject to its calibration. Calibration refers to the reliability of the forecast in that there is statistical consistency between the probabilistic forecast and the corresponding observations. We have a perfectly calibrated ensemble forecast when a weather event that is predicted to occur with probability P actually is observed with a relative frequency P in the long run. Sharpness refers to the concentration of the predictive distributions, meaning that under the condition that all forecasts are calibrated, we define the sharpest to be the best. In other words, the sharper a calibrated predictive distribution, the less uncertainty and the better its performance.

Probabilistic weather forecasts are often based on models which create a collection of M forecasts considered at the same time. These multiple forecasts are so called ensemble forecasts and can for instance be obtained by running a model several times with different initial conditions, or by running different model physics [14]. Each single forecast in the ensemble forecast is referred to as an ensemble member. Let us assume that these ensemble members are predicted by M different weather forecast providers. By comparing the forecasts a forecaster then can tell you how likely it is that a particular weather event will occur. If the forecasts vary a lot, the forecaster knows that the weather is uncertain. In contrast, if the forecasts are similar, they will have more confidence in predicting a particular event. To measure the performance of each provider in forecasting, all providers maintain their "place" in the ensemble each time they make new forecasts. However, in this study we have only one provider, the European Center for Medium-Range Weather Forecasts (ECMWF), who supplies all the M ensemble members by using one specific model.

This model is run several times with different initial conditions. Thus, the ensemble members are exchangeable in that the ordering of the ensemble members does not matter [5]. In other words, ensemble member m issued today is not correlated with ensemble member m issued yesterday.

In order to get more information from an ensemble forecast, probabilistic forecasts in form of predictive probability density functions (pdf) have to be generated. Furthermore, studies have shown that ensemble forecasts often tend to be underdispersive in that the observed value far too often lies outside the ensemble range [6, 14, 17, 19]. In order to address these issues, post-processing techniques can be applied. The most common approaches are Bayesian model average (BMA) [26] and Ensemble model output statistics (EMOS) [16]. These models convert ensemble forecasts into calibrated and sharp probability forecasts. BMA makes use of mixture distributions, in which each ensemble member corresponds to its own probability component. EMOS is based on multiple linear regression. Both methods can be applied to a number of different weather variables like temperature, precipitation, air pressure or wind speed for univariate models. In this thesis we propose extended versions of bivariate BMA and EMOS models for ensemble temperature forecasts to accounts for lead time dependencies between errors.

A common technique for estimation of the BMA and EMOS model parameters is Maximum likelihood estimation (MLE) [26]. Additionally, Gneiting et al. [16] proposed the use of minimum CRPS (minCRPS) estimation, where the parameters found minimize the CRPS value for the training data. In order to find out which of these estimation methods lead to best post-processing performance, we consider both methods. However, minCRPS is an estimation method for univariate models. For bivariate models we therefore propose minimum ES (minES) estimation, where sampling is used to simplify the estimation process.

Scoring functions, such as mean absolute error (MAE) and root mean square error (RMSE) are often sufficient methods to evaluate the quality of the deterministic forecasts. These evaluation methods depend both on the deterministic forecast and the realization, and assess the quality of the predictions [11]. The score is negatively oriented, meaning the smaller, the better. However, in order to assess probabilistic forecasts we need to eval-

uate both the calibration and the sharpness [26]. An often used scoring rule for a joint assessment of calibration and sharpness is the continuous ranked probability score (CRPS) [7, 15, 20]. It is a widely used strictly proper scoring rule, meaning that the forecaster gets the best score by forecasting his or her true beliefs, although it may be possible to get the same score by using a different forecast [15]. As for scoring functions, scoring rules are negatively orientated. In order to assess the calibration of a forecast, we also consider probability integral transform (PIT) histograms [12] for predictive distributions.

In this Master's thesis we compare univariate BMA and EMOS post-processing techniques and propose an extended version for bivariate models. Both the univariate and bivariate BMA and EMOS models are applied to temperature data from Trondheim – Voll, which are provided by ECMWF. According to Raftery et al. [26], a normal distribution is considered to be appropriate for temperature data. We apply the methods on forecasts of 6-hourly intervals up to +42 hours. Hence, 8 different lead times are evaluated. For bivariate models only lead time 5 and 6 are considered. Additionally, two different estimation methods are compared: MLE and minCRPS for univariate model. For bivariate models we propose the minES method in addition to use MLE.

This thesis is organized as follows. In the next chapter we introduce the data used in the case study done. Chapter 3 we give a brief review of univariate post-processing techniques. Additionally, estimation and assessment methods are explained in this chapter. We describe an extended version of the BMA and EMOS models in Chapter 4. A simulation study is done in Chapter 5 to see if the parameters are consistent. In a case study in Chapter 6, we decide on the length of training period before we apply both the univariate and extended bivariate BMA and EMOS techniques to ECMWF data and make use of the assessment methods. The thesis is concluded with a discussion in Chapter 7, where we summarize the results and suggest possible ideas for further work.

2. Data and case

This chapter serves an overview of the data used in a case study in Chapter 6. In the first section, temperature observation data and ensemble forecasts available are introduced. Explanatory analysis of these are described in the second section.

2.1. Available observations and forecasts

In this study we consider the observation station, Voll, located in the Trondheim area in the Trøndelag district of central Norway, see Figure 2.1. The Trondheim region is situated close to the Trondheimsfjord and is characterized by lowland with small hills up to 500 meters above sea level. Voll is located in Trondheim municipality, 127 meters above sea level with coordinates (63.4106°N, 10.4536°E). It is situated 3.7 km away from centred Trondheim and was established in January 1923. Temperature observations are provided by the Norwegian Meteorological Institute (MET) between January 1st 2007 and December 31st 2011 through *eKlima* (http://eklima.met.no), where hourly data is available. The temperature is measured daily in 2-m height above the ground. In our analysis, 6-hourly intervals observations are considered, starting at 00:00 Coordinates Universal Time (UTC). UTC is the primary time standard by which the world regulates clocks and time.

Ensemble forecasts for temperature are obtained from the European Center for Medium-Range Weather Forecasts (ECMWF) between January 1st 2007 and December 31st 2011 through TIGGE (http://tigge-portal.ecmwf.int). ECMWF is an international meteorological organization, founded in England in 1975. They develop numerical methods for medium-range weather forecasting. Ensemble forecasts are created by running numerical weather

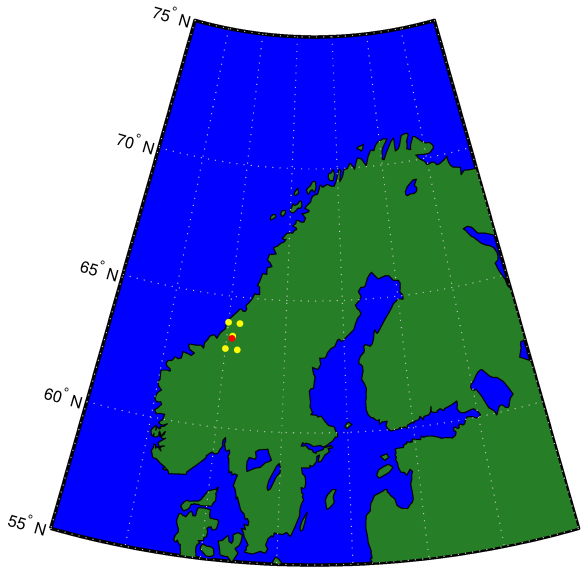


Figure 2.1: *Map over northern Europe. The yellow dots indicate the grid of downloaded ensemble forecast data. The red dot indicates the location of Trondheim – Voll in Norway.*

prediction models a number of times with slightly different starting conditions. The model is based on deterministic simulation models that represent the physics of the atmosphere. One ensemble forecast consists of $M = 50$ separate forecasts made by the same computer mode. The spread between ensemble members indicates the uncertainty of the ensemble forecast.

The forecasts start at 00:00 UTC and are available at 6-hourly intervals up to +384 hours. They are calculated on a grid of resolutions 0.5×0.5. Each ensemble forecast is downloaded on a 3×3 grid from ECMWF with coordinates (63°N-64°N, 10°E-11°E). For further calculations, the forecasts for the coordinate (63.5°N, 10.5°E) are used. We choose the forecasts from this coordinate because this point only is 10.6 kilometers North-West from the observation station at Voll.

In this study we look at 6-hourly forecasts predicted up to 42-hours ahead of surface temperature, initialized at 00:00 UTC. For each 6-hourly interval we step one lead time forward, and in total 8 different lead times are considered. A list of lead times and the corresponding predicted hours ahead is given in Table 2.1. Also the observational time is listed.

Lead time (l)	0	1	2	3	4	5	6	7
Hours predicted ahead	0	6	12	18	24	30	36	42
Observational time (UTC) (t)	00	06	12	18	00	06	12	18

Table 2.1: *Lead time, corresponding hours predicted ahead and corresponding observational time.*

We have n = 1,...,N number of days. Four times a day, at time t = 00, 06, 12, 18 UTC, the temperature is observed, and new forecasts are issued for 8 different lead times. For simplicity, we denote the observation time during one day for $\lambda \in \{1, 2, 3, 4\}$. Hence, the observations observed day n at time λ are denoted as $y_{n,\lambda}$.

2.2. Explanatory analysis

The climate in Trondheim – Voll is characterized by a seasonal variation in temperature. This can be observed in Figure 2.2, where daily mean observed temperature, $\bar{y}_n = \frac{1}{4}\sum_{\lambda=1}^{4} y_{n,\lambda}$, between January 1st 2007 and December 1st is plotted. In order to give a better overview of these trends, we only plot the daily mean observed temperature, $\bar{y}_n$, for year 2011, see Figure 2.3. $\bar{y}_n$ seems to be seasonally stationary, in that the mean temperature seems to be the same every year. However, 2010 was a colder year with $\bar{y}_n$ up to almost 3 degrees lower than the other years, see Table 2.2.

In order to see if there is variation in temperature during the day, we subtract $\bar{y}_n$ from $y_{n,\lambda}$, $\delta_n = y_{n,\lambda} - \bar{y}_n$, see Figur 2.4. We observe that there is less variation in temperature at time 06:00 UTC and 12:00 UTC (λ = 2, 3) than at time 00:00 UTC and 18:00 UTC (λ = 1, 4). Furthermore, we note an outlier at almost -15°C at time 00:00 UTC. This indicates that there was large variation in temperature that day.

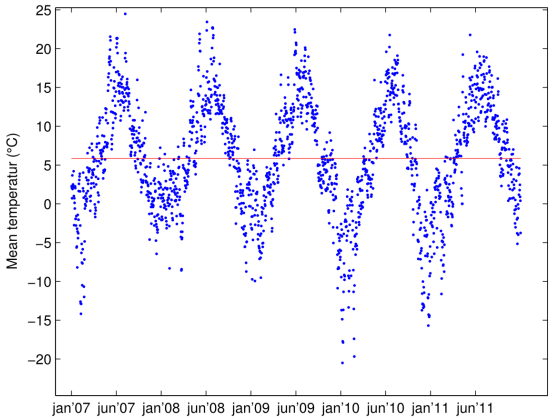


Figure 2.2: *Daily mean observed temperature, $\bar{y}_n = \frac{1}{4}\sum_{\lambda=1}^{4} y_{n,\lambda}$, between year 2007 and 2011. The blue dots represent the daily mean observed temperature. The red line represent the mean temperature between 2007 and 2011 which is 5.86° C*

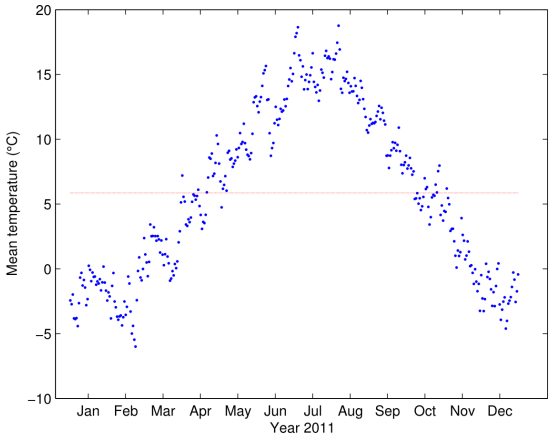


Figure 2.3: *Daily mean observed temperature, $\bar{y}_n = \frac{1}{4}\sum_{\lambda=1}^{4} y_{n,\lambda}$, year 2011. The blue dots represent the daily mean observed temperature. The red line represent the mean temperature for year 2011 and is 6.80° C.*

Mean temp. (°C)	UCT			
Year	**00:00**	**06:00**	**12:00**	**18:00**
2007	4.83	4.87	7.51	6.72
2008	5.37	5.35	8.15	7.51
2009	5.04	4.98	7.67	6.89
2010	2.66	2.56	5.24	4.53
2011	5.68	5.76	8.34	7.43

Table 2.2: *Observed mean temperature at time 00:00, 06:00, 12:00 and 18:00 UTC for each year.*

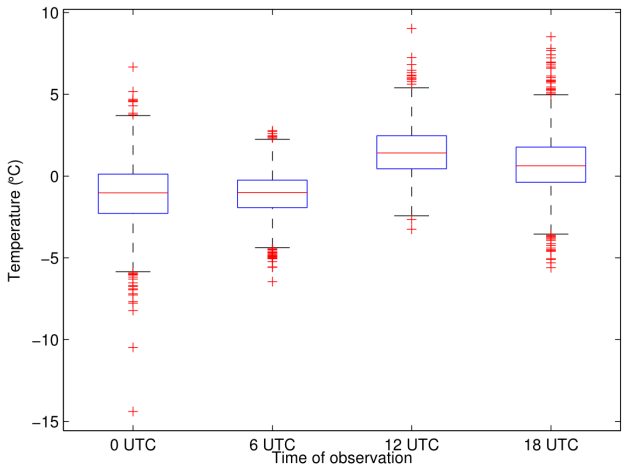


Figure 2.4: *$\delta_n = y_{n,\lambda} - \bar{y}_n$, showing the variation of observed temperature during the day for year 2007 to 2011. The blue box shows the 25-75% quartile, the red line the median and the red crosses the outliers.*

An illustration of the ensemble forecasts, $x_{t,2}$, for lead time 2 January 2011 is shown in Figure 2.5. Each ensemble member, $x_{m,t,2}$, is plotted as a blue circle. The red circle denotes the corresponding observation. We observe that in this example, the ensemble length, $x_{t,2}^{length} = max(x_{m,t,2}) - min(x_{m,t,2})$, is between 1°C and 6°C. Further, we note that the observations lie mostly above the ensemble range.

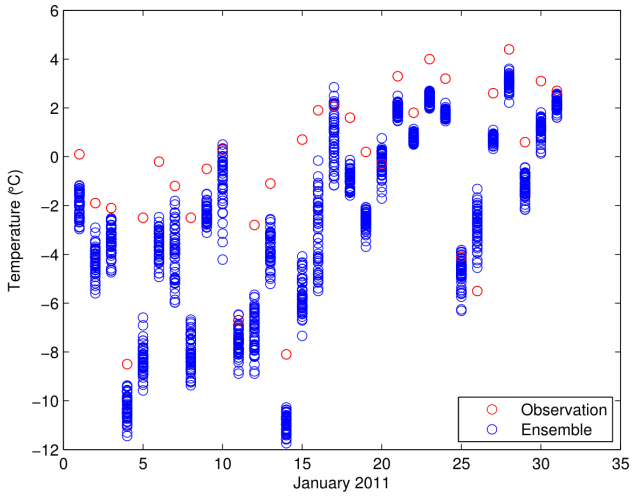


Figure 2.5: *Plot of ensemble forecasts at $l = 2$ and corresponding observations at 12:00 UCT for January 2011.*

For $l = 0$ the ensemble forecasts are issued at 00:00 UTC. Since the ensemble forecasts are made at the same time as they are observed, small ensemble intervals are expected. We plot the length of all ensemble forecast intervals, $x_{t,l}^{length}$, between year 2007 and 2011 to verify that this is true for our data, see Figure 2.6. However, we observe that the lengths make a jump around June 2010 where $x_{t,l}^{length}$ becomes larger. The reason is that there has been done changes in the initial perturbations [21]. On the 22nd of June a new configuration was implemented which has a more reliable spread in the short range.

Table 2.3 shows the error between the observational mean and the ensemble mean, $e = \bar{y}_n - \bar{x}_{t,l}$. We observe a positive error. This indicates that the observational mean almost is consistently larger than the ensemble mean. The bold numbers in the table denote the largest error for lead time l. We note that the error is larger in 6 of 8 lead times in year 2010.

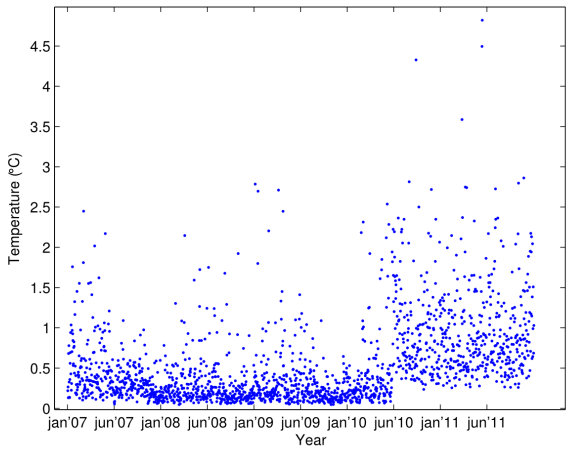


Figure 2.6: *Length of ensemble intervals,* $x_{t,0}^{length} = max(x_{m,t,0}) - min(x_{m,t,0})$ *between year 2007 and 2011 for* $l = 0$.

Year / Error (°C)	l**=0**	l**=1**	l**=2**	l**=3**
2007	1.04	1.30	1.06	1.45
2008	1.37	1.83	1.40	2.14
2009	**1.46**	1.97	**1.53**	2.25
2010	0.81	**2.24**	1.50	**2.92**
2011	0.55	1.41	0.65	1.74
	l**=4**	l**=5**	l**=6**	l**=7**
2007	1.52	1.44	1.18	1.61
2008	2.40	2.10	1.40	2.18
2009	2.66	2.23	**1.66**	2.39
2010	**3.60**	**2.83**	**1.66**	**3.00**
2011	2.30	2.05	1.08	2.03

Table 2.3: *Error between the observational mean and the ensemble mean,* $e = \bar{y}_n - \bar{x}_{t,l}$, *for each year and each lead time. The bold numbers in the table denote the largest error for lead time* l.

In order to see if there is variation in temperature forecasts during the day, we divide the lead times into two days, Day a and Day b. Let $l = 0$, 1, 2, 3 correspond to time 00:00, 06:00, 12:00 and 18:00 UTC Day a, and $l = 4$, 5, 6, 7 correspond to time 00:00, 06:00, 12:00 and 18:00 UTC Day b. The mean of each ensemble forecast at lead time l is given as $\bar{x}_{t,l} = \frac{1}{M}\sum_{m=1}^{M} x_{m,t,l}$. Let then the mean of $\bar{x}_{t,l}$ for Day a be defined as $\bar{\bar{x}}_n^a = \frac{1}{4}\sum_{l=0}^{3} \bar{x}_l$. For Day b the mean of $\bar{x}_{t,l}$ is defined as $\bar{\bar{x}}_a = \frac{1}{4}\sum_{l=4}^{7} \bar{x}_{t,l}$. Subtracting $\bar{\bar{x}}_n^a$ from $\bar{x}_{t,l}$ for $l = 0$, 1, 2 ,3, and $\bar{\bar{x}}_n^b$ from $\bar{x}_{t,l}$ for $l = 4$, 5, 6, 7 shows the variation in temperature ensemble forecasts during the day, see Figure 2.7. We observe that the variation in ensemble forecasts during the day is quite similar to the variation in observed temperature during the day, see Figure 2.4. Furthermore, we note that there is less variation at 06:00 UTC and 12:00 UTC than at 00:00 UTC and 18:00 UTC, as we also observed for the observed temperature.

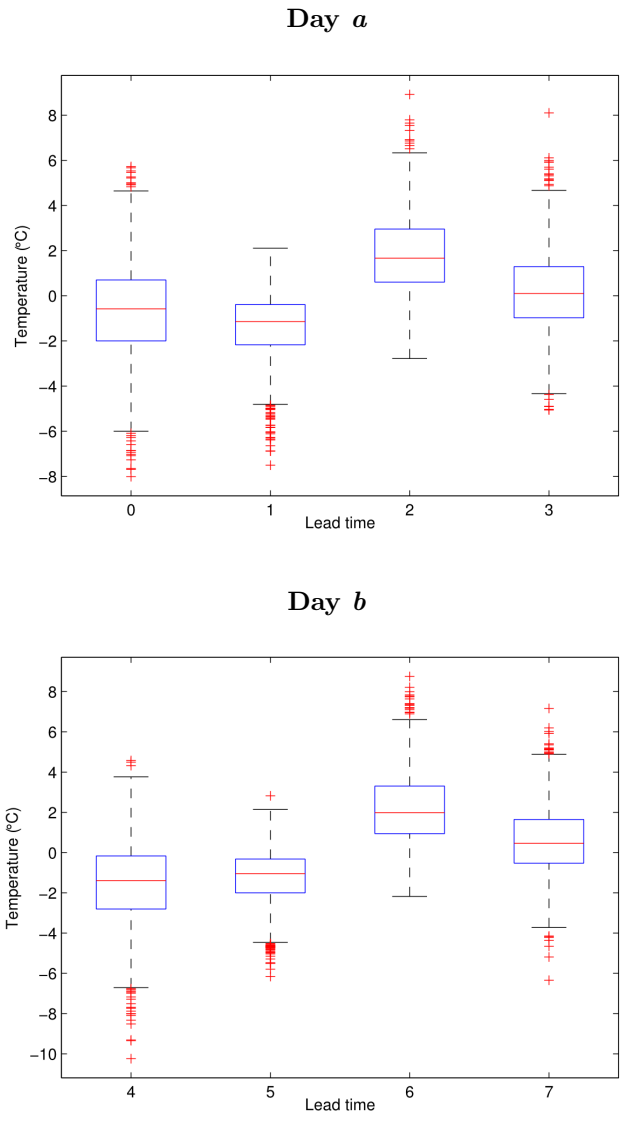


Figure 2.7: *Spread for temperature ensemble forecast during the day for year 2007 to 2011. The blue box shows the 25-75% quartile, the red line the median and the red crosses the outliers. Day a contains data from lead time 0, 1, 2 and 3, and Day b contains data from lead time 4, 5, 6 and 7.*

3. Background

This chapter gives a brief review of two different post-processing techniques for producing calibrated ensemble forecasts for univariate models. How to evaluate the models is explained in section 3, and in section 4 we describe two different parameter estimation methods.

As introduced in the previous chapter, we consider ensemble forecasts for 8 different lead times for several days. Each ensemble forecast consists of M ensemble member forecasts. Let t be the issue time, l the lead time and m the number of an ensemble member in an ensemble forecast. An ensemble member forecast can then be denoted as $x_{m,t,l}$. However, in this chapter we only consider one lead time at one issue time, which simplifies the notation to x_m.

3.1. Bayesian Model Averaging

Raftery et al. [26] introduced Bayesian model averaging (BMA) as a method for generating calibrated predictive probability density functions (pdf) from ensemble forecasts. It allows for combination of different dynamical models, such as numerical weather prediction models, and treats each ensemble member forecast as a statistical model. For non-exchangeable ensemble forecasts, where the ordering of the ensemble members is set, each ensemble member forecast is associated with a pdf, $p_m(y|x_m, \theta_m^{bma})$. Here y is the quantity of interest and θ_m^{bma} are the parameters of the m'th component pdf. The ensemble BMA predictive model is then given by a mixture of the pdf's,

$$p(y|x_1, ..., x_M; \theta_1^{bma}, ..., \theta_M^{bma}) = \sum_{m=1}^{M} w_m p_m(y|x_m, \theta_m^{bma}), \tag{3.1}$$

where M is the total number of forecasts in an ensemble forecast. The weights w_m can be interpreted as posterior probabilities based on ensemble member m's relative performance in the training period. These are assumed to be non-negative with $\sum_{i=1}^{M} w_m = 1$ [9, 26]. The choice of pdf, p_m depends on the weather variable of interest. Raftery et al. [26] consider normal distribution to be appropriate for temperature. This seems to be reasonable because temperature often has normal distributed errors [9, 26]. The distribution is centred so that the conditional pdf, $p_m(y|x_m, \theta_m^{bma})$ is a normal pdf with mean $\eta_m = \alpha_m + \beta_m x_m$ and standard deviation τ_m. The mean can be viewed as a simple linear bias-correction of the ensemble member forecasts.

In this study all the ensemble members are exchangeable. Hence, the BMA weights w_m and model parameters θ_m^{bma} can be considered to be equal for all ensemble members, $x_1, ..., x_M$. The BMA predictive model in Eq. 3.1 can therefore be rewritten as

$$p(y|x_1, ..., x_M, \theta^{bma}) = \sum_{m=1}^{M} w p_m(y|x_m, \theta^{bma}), \tag{3.2}$$

where $\theta^{bma} = \{\alpha, \beta, \tau\}$ and the weights, $w = 1/M$, are constant. The distribution is centred so that the conditional pdf, $p_m(y|x_m, \theta^{bma})$ is a normal pdf with mean $\eta_m = \alpha + \beta x_m$ and standard deviation τ, where α and β are the bias-correction parameters. This can be denoted as

$$p_m(y|x_m, \theta^{bma}) \sim \mathcal{N}(\alpha + \beta x_m, \tau^2). \tag{3.3}$$

An illustration of the BMA predictive pdf is given in Figure 3.1. The predictive pdf is a weighted sum of five normal distributed pdf's (the thin blue lines). Also the observation (black vertical line), the values of the raw ensemble members (black circles) and the bias-corrected ensemble member forecasts (black crosses) are shown. In this example the ensemble members are exchangeable. Hence, the parameters $\alpha = 1.71$, $\beta = 0.92$, $\tau = 0.98$ and the weights $w = 1/5$ are the same for all ensemble member forecasts. We observe that the raw ensemble member forecasts mostly lie outside the predictive pdf, and that the corrected ensemble members are much more calibrated in that the observation falls within the ensemble

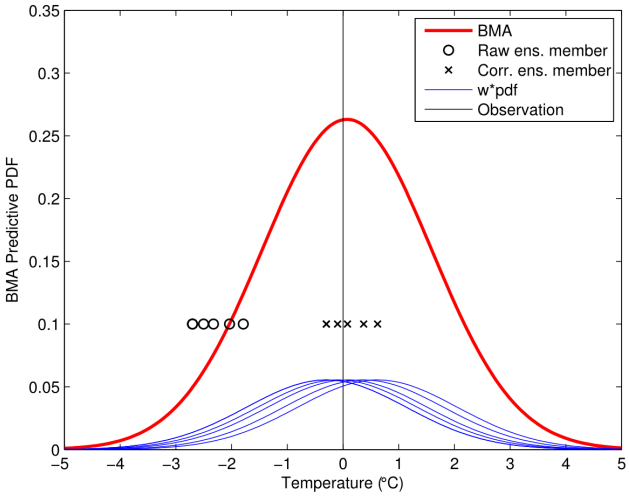


Figure 3.1: *BMA predictive pdf (thick curve) and its five components (thin curves) for January 31st 2008. Parameter estimates are* $\alpha = 1.71$, $\beta = 0.92$, $\tau = 0.98$ *and* $w = 1/5$. *Each circle is one of the five raw ensemble members, the crosses are the bias-corrected ensemble members and the solid vertical line is the observation.*

range. The BMA predictive pdf is a result of the blue single pdf's. We observe that the observation almost is in the center of the predictive pdf.

The BMA predictive mean, $\mu^{bma} = \sum_{m=1}^{M} w(\alpha + \beta x_m)$ can also be used as a deterministic forecast and can be compared with the mean of the raw ensemble forecasts. The BMA predicted variance can be written as [25]

$$\sigma^{bma^2} = \sum_{m=1}^{M} w \left((\alpha + \beta x_m) - \sum_{i=1}^{M} w(\alpha + \beta x_i) \right)^2 + \tau^2. \qquad (3.4)$$

3.2. Ensemble Model Output Statistics

Ensemble model output statistics (EMOS) is a post-processing technique based on multiple linear regression. It is easy to implement and corrects the forecast bias and underdispersion. For temperature it fits a normal distribution to the ensemble member forecasts. EMOS was first proposed by Gneiting et al. [16] and is an extension to the Model Output Statistics (MOS) technique developed by Glahn and Lowry [10].

The EMOS predictive pdf's are normal distributed with predictive mean $\mu = a + b_1x_1 + ... + b_Mx_M$ and predictive variance $\sigma^2 = c + dS^2$. In this approach, the predictive mean is a bias-corrected weighted average of the ensemble forecasts, where a is a bias-correction and $b_1, ..., b_M$ are regression coefficients. The predictive variance is modeled as a linear function of the ensemble forecast variance $S^2 = \frac{1}{M}\sum_{m=1}^{M}(x_m - \bar{x})^2$, where $\bar{x} = \frac{1}{M}\sum_{m=1}^{M} x_m$ is the mean of the ensemble forecast. Hence, the normal predictive distribution for the variable of interest, y can then be written as

$$p(y|x_1, ..., x_M, \theta^{emos}) \sim \mathcal{N}(a + b_1x_1 + ... + b_Mx_M, c + dS^2). \tag{3.5}$$

However, in this study we have exchangeable ensemble member forecasts. Thus, the EMOS technique is simplified so that the predictive mean of the normal distribution becomes a linear function, $\mu = a + b\bar{x}$. We can therefore rewrite the normal predictive distribution given in Eq. 3.5 as

$$p(y|\bar{x}, \theta^{emos}) \sim \mathcal{N}(a + b\bar{x}, c + dS^2), \tag{3.6}$$

where $\theta^{emos} = \{a, b, c, d\}$. We note that the EMOS approach only conditions on a single model considered to be the best one. Thus, it yields one predictive pdf [16]. In contrast, the BMA approach makes use of multiple models and fits a mixture density distribution as predictive pdf [26].

3.3. Assessment methods

In this section we present assessment methods to evaluate the performance of univariate and bivariate BMA and EMOS models. Evaluation of predictions is an important step in any forecast process. For deterministic forecasts, score functions, such as mean absolute error (MAE) or root mean square error (RMSE) are sufficient methods for evaluating the quality of the forecasts. However, according to Gneiting et al. [12], the goal is to make the ensemble forecasts as sharp as possible subject to its calibration. Calibration is a measure of statistical consistency between observations and ensemble forecasts. For example, if an event is predicted to occur with probability 60%, on average it should happen in about 60% in the long run. Sharpness measures the concentration of the predictive distribution, meaning that the sharper a calibrated predictive distribution is, the less uncertainty and the better its performance. The continuous ranked probability score (CRPS) is often used to evaluate the calibration and sharpness of probabilistic forecasts [15]. In order to assess the calibration of a forecast, we will also consider probability integral transform (PIT) histograms [12] for predictive distribution.

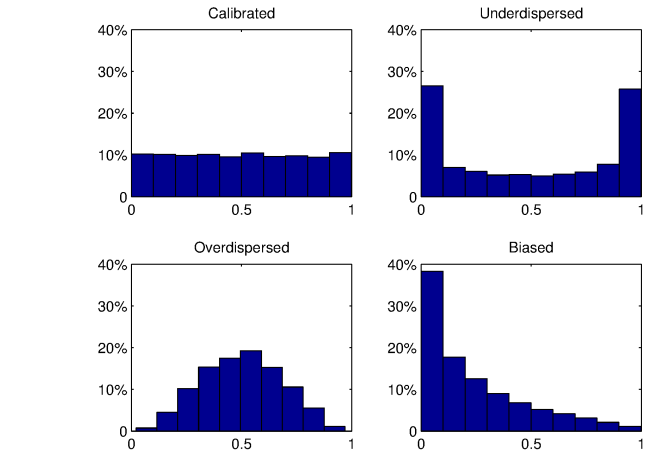
3.3.1. Assessing calibration

There are several ways to assess calibration. For ensemble forecasts the verification rank histogram (VRH) [1, 7, 19] can be used to evaluate the raw ensemble forecasts. In order to compute VRH, one arranges the ensemble forecasts and the corresponding observation y in increasing order. Hence, it is possible to check which index the verifying observation becomes in the range from 1 to $M+1$. In a calibrated ensemble, the verifying observation is equally likely to get any of the indexes. In the long run the VRH should therefore be uniform if the predictive pdf is calibrated.

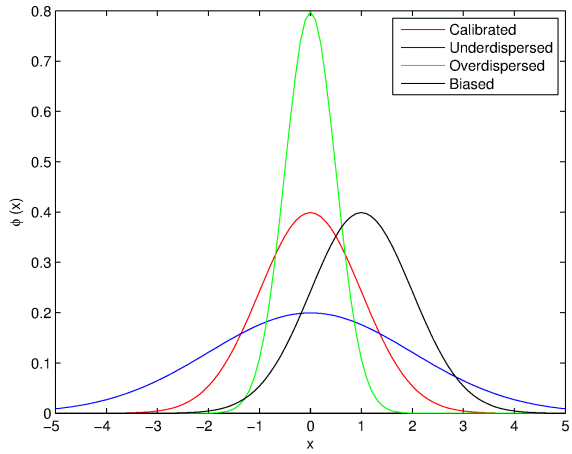
A continuous analog of the VRH is the probability integral transform (PIT) histogram. It is a common tool used to evaluate the calibration of a univariate predictive forecast distribution [8, 12]. Let F be the predictive cdf of the observation y, then the probability integral transform is defined by

$$\text{PIT} = F(y) \sim \mathcal{U}[0, 1]. \tag{3.7}$$

Here, PIT is a number between 0 and 1. PIT's interpretation is the same as for the VRH, meaning that the PIT histogram should be closely to uniform when it is calibrated.



a)



b)

Figure 3.2: *a) Hypothetical verification rank histograms for a well calibrated* ($\mathcal{N}(0,1)$)*, an underdispersive* ($\mathcal{N}(0,0.5)$)*, an overdispersive* ($\mathcal{N}(0,2)$)*, and a biased* ($\mathcal{N}(1,1)$) *predictive distribution are plotted. b) The corresponding predictive distribution.* ϕ *denotes the pdf.*

An illustration of a PIT-histogram for different hypothetical distributions is shown in Figure 3.2. In this example, all data points are random samples from a standard normal, $\mathcal{N}(0, 1)$ distribution. A U-shaped histogram indicates that the ensemble forecasts are underdispersed in that the observations too often lie outside of the ensemble range. A hump-shaped diagram indicates that the variance is too wide, meaning that the ensemble forecasts are overdispersed. A triangle-shaped histogram indicates that the ensemble forecasts are biased, while a nearly flat histogram suggests calibration.

3.3.2. Assessing sharpness

Although the PIT-histogram is a good tool for evaluating calibration of probabilistic forecasts, it is not sufficient to tell if a post-processing technique is useful or not. As mentioned in the introduction, we want the ensemble forecasts to be as sharp as possible subject to its calibration [12]. We therefore need to combine evaluation of calibration with assessing sharpness to identify if the BMA and EMOS approaches yield both well calibrated and sharp ensemble forecasts. In this study we use proper scoring rules to assess both sharpness and calibration.

Proper scoring rules

A scoring rule is said to be a strictly proper scoring rule when the forecaster gets the best score by forecasting his or her true beliefs, although it may be possible to get the same score by using a different forecast [15]. A widely used proper scoring rule in the assessment of the quality of probabilistic forecasts is the Continuous rank probability score (CRPS) [7, 15, 20]. The CRPS gives a joint assessment of calibration and sharpness and is negatively orientated, such that a smaller CRPS means a better forecast. A perfect forecast gives CRPS score of 0. The CRPS is sensitive to distance in that it is penalizing predictions that are far away from the actual observation. Let F be the cumulative distribution function (cdf) and y_{t+l} the observed quantity, which in our case is the 2-m temperature at issue time t and lead time l. The CRPS measures the difference between the predicted and the occurred cdf's. For a univariate forecast the standard form is defined as

$$\mathrm{CRPS}_t(F, y_t) = \int_{-\infty}^{\infty} \left[F(\xi) - H(\xi - y)\right]^2 d\xi, \tag{3.8}$$

where t is the issue time and H is the Heaviside function,

$$H(\xi - y) = \begin{cases} 0 & \text{for} \quad \xi < y \\ 1 & \text{for} \quad \xi \geq y. \end{cases} \tag{3.9}$$

Gneiting et al.[15] show that the CRPS for one lead time also can be written as

$$\mathrm{CRPS}_t(F, y_t) = \mathbb{E}|x_{m,t} - y_t| - \frac{1}{2}\mathbb{E}|x_{m,t} - x^*_{m,t}|, \tag{3.10}$$

where $x_{m,t}$ and $x^*_{m,t}$ are independent random variables with cdf F, and $\mathbb{E}$ denotes the expectation. Hence, it is possible to calculate CRPS from the predictive cdf from an ensemble forecast of size M:

$$\mathrm{CRPS}_t(F_{ens}, y_t) = \frac{1}{M}\sum_{m=1}^{M}|x_{m,t} - y_t| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{n=1}^{M}|x_{m,t} - x_{n,t}|, \tag{3.11}$$

where y_t is the observation and $x_{m,t}$ the ensemble members. F_{ens} is a discrete predictive distribution from a forecast ensemble of size M. The CRPS can also be estimated by sampling. To evaluate a forecast procedure we average the CRPS_t over T forecast-observation pairs,

$$\mathrm{CRPS} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{CRPS}_t. \tag{3.12}$$

For a deterministic forecast the CRPS reduces to Mean absolute error (MAE). MAE is a scoring function used for evaluating deterministic forecasts. It measures how the values of the forecasts differ from the values of the observations and is defined as

$$\mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T}|y_{t+l} - \bar{x}_{t,l}|. \tag{3.13}$$

Here y_{t+l} is the observation and $\bar{x}_{t,l}$ is the mean of the ensemble forecast at issue time t and lead time l. T is the total number of days evaluated.

The integral in Eq. 3.10 is not on a closed form, meaning that it can cause problems to some distributions. Gneiting et al. [16] derived an analytic expression for the CRPS when the cdf is normal distributed with mean μ and variance σ^2. By repeated partial integration in Eq. 3.10 the CRPS for a normal distribution becomes

$$\begin{aligned}&\mathrm{CRPS}_t(\mathcal{N}(\mu_{m,t}, \sigma_t^2), y_t) = \\ &\quad \sigma\left(\frac{y_t - \mu_{m,t}}{\sigma_t}\left[2\Phi\left(\frac{y_t - \mu_{m,t}}{\sigma_t}\right) - 1\right] + 2\phi\left(\frac{y_t - \mu_{m,t}}{\sigma_t}\right) - \frac{1}{\sqrt{\pi}}\right),\end{aligned} \tag{3.14}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and the cdf of a standard normal random variable at the normalized prediction error, $(y_t - \mu_t)/\sigma_t$ [13, 18]. Eq. 3.14 is used to evaluate the EMOS model. However, for BMA, F is the cdf of a Gaussian mixture distribution, meaning that the distribution is composed of two or more normal distributed pdf's. Hence, given a set of n cdf's, $F_1, ..., F_n$, the mixture distribution for non-exchangeable ensemble forecasts, introduced by Raftery et al. [26] can be written as

$$F = \sum_{i=1}^{n} w_i F_i, \tag{3.15}$$

where w_i are non-negative weights and $\sum w_i = 1$. For exchangeable ensemble forecasts a closed form solution for this is

$$\begin{aligned}&\mathrm{CRPS}_t\left(\sum_{m=1}^{M} w\mathcal{N}(\mu_{m,t}, \sigma_t^2), y_t\right) = \\ &\quad \sum_{m=1}^{M} wA(y_t - \mu_{m,t}, \sigma_t^2) - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M} w^2 A(\mu_{m,t} - \mu_{n,t}, \sigma_t^2 + \sigma_t^2),\end{aligned} \tag{3.16}$$

where A is

$$A(\mu_{m,t}, \sigma_t^2) = 2\sigma_t \phi\left(\frac{\mu_{m,t}}{\sigma_t}\right) + \mu_{m,t}\left(2\Phi\left(\frac{\mu_{m,t}}{\sigma_t}\right) - 1\right). \tag{3.17}$$

Eq. 3.16 is used to evaluate the BMA model.

An illustration of a predictive cdf is given in Figure 3.3. It is plotted together with the corresponding observation. The grey area between the observation and the predicted cdf is the value returned by the CRPS.

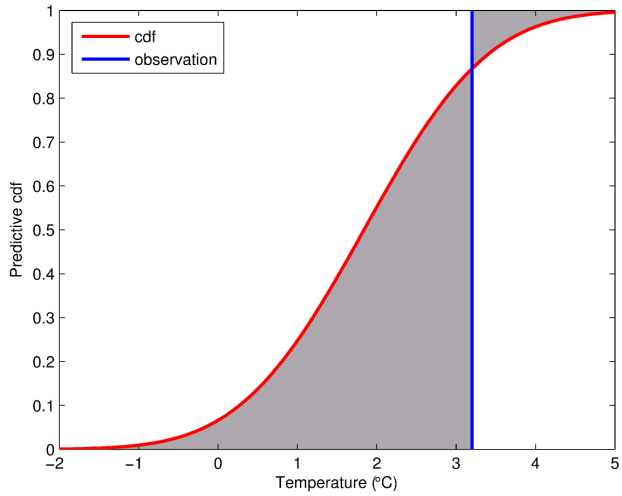


Figure 3.3: *Illustration of CRPS for January 31st 2011. The forecasts are made twelve hours ahead and the corresponding observation is observed at 12:00 UCT. The figure shows the computation of the CRPS, which is the grey area between the cdf of the forecast and the observation, $y = 32$.*

Energy score (ES) is a generalization of the CRPS to evaluate multivariate probabilistic forecasts [15]. For one forecast-observation pair at one lead time, it is defined by

$$\mathrm{ES}_t(F_t, \mathbf{Y}_t) = \mathbb{E}\,\|\mathbf{X}_t - \mathbf{Y}_t\| - \frac{1}{2}\mathbb{E}\,\|\mathbf{X}_t - \mathbf{X}_t^*\| \tag{3.18}$$

where $\|\cdot\|$ denotes the Euclidean norm, and $\mathbf{X}_t$ and $\mathbf{X}_t^*$ are independent

random vectors from the multivariate probabilistic forecast with cdf F_t. $\mathbf{Y}_t$ is a vector of observations. It can be calculated using samples of the multivariate probabilistic forecast

$$\mathrm{ES}_t(F_t, \vec{y}_t) = \frac{1}{M}\sum_{m=1}^{M} \|\mathbf{X}_{m,t} - \mathbf{Y}_t\| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{n=1}^{M} \|\mathbf{X}_{m,t} - \mathbf{X}_{n,t}\|\,, \quad (3.19)$$

where $\mathbf{X}_{1,t}, ..., \mathbf{X}_{m,t}$ are m independent vectors sampled from the multivariate probabilistic forecast. The temporal average of all values of ES_t is denoted as ES, where T is the number of forecast-observation pairs,

$$\mathrm{ES} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{ES}_t. \quad (3.20)$$

3.4. Parameter estimation

We use two different methods when estimating the parameters: Maximum likelihood estimation (MLE) and minimum CRPS estimation (minCRPS).

3.4.1. Maximum likelihood estimation

Maximum likelihood estimation (MLE) was proposed by R. A. Fisher and is a common method to estimate the parameters of a statistical model. Generally, for a given data set and a basic statistical model, the maximum likelihood method selects values of the parameters that maximize the likelihood function.

BMA

At one lead time l, we have the training data, $y_{t-1}, y_{t-2}..., y_{t-K}$ of K observations with conditional pdf's, $p_m(y_{t-k}|x_{m,t-k}, \theta^{bma})$. For BMA with exchangeable ensemble forecasts, the parameters α and β are estimated by simple linear regression, and only the standard deviation, τ, is estimated by MLE. The likelihood function $\mathcal{L}$ for mixture normal distributed functions is defined by

$$\mathcal{L}(x_1, ..., x_M, \tau | y_{t-1}, ...y_{t-K}) = \prod_{k=t-K}^{t-1} \left[\sum_{m=1}^{M} w p_m(y_k | x_{m,k}, \tau) \right]. \quad (3.21)$$

In practice, it is easier to find the MLE by maximizing the log-likelihood function. This is because the logarithm is a monotonically increasing function. Equation 3.21 can then be rewritten as

$$\log \mathcal{L}(x_1, ..., x_M, \tau | y_{t-1}, ..., y_{t-K}) = \sum_{k=t-K}^{t-1} \log \left[\sum_{m=1}^{M} w p_m(y_k | x_{m,k}, \tau) \right], \quad (3.22)$$

where we consider equal weights and parameters. Because of the logarithm of a sum in the function, this problem is not possible to solve analytically. Therefore it is necessary to find a numerical method to compute the MLE.

EMOS

Given the training data $y_{t-1}, y_{t-2}..., y_{t-K}$ of K observations, the EMOS parameters, a, b, c and d are all estimated at once with MLE. Equally as for the BMA, it is more convenient to maximize the logarithm of the likelihood function. The log-likelihood function for the EMOS model Eq. 3.6 is

$$\log \mathcal{L}(x_1, ..., x_M, a, b, c, d | y_{t-1}, ..., y_{t-K}) = -\frac{1}{2} \left\{ K \log(2\pi) + \sum_{k=t-K}^{t-1} \frac{[y_k - (a + b\bar{x}_k)]^2}{c + dS_k^2} + \sum_{k=t-K}^{t-1} \log(c + dS_k^2) \right\}. \quad (3.23)$$

To make sure that the density is a valid probability distribution, we set c and d as non-negative parameters [16]. To include these constraints in the EMOS model we write

$$c = \gamma^2$$
$$d = \delta^2.$$

3.4.2. Minimum CRPS estimation

An other estimation method is the Minimum CRPS (minCRPS) estimation, suggested by Gneiting et al. [16].

BMA

Similarly as for MLE, we use the training data $y_{t-1}, y_{t-2}..., y_{t-K}$ of K observations. By taking into account that we have equal weights and exchangeable ensemble member forecasts, we can rewrite Equation 3.16 as

$$\text{minCRPS}\left(\sum_{m=1}^{M} w\mathcal{N}(\eta_m, \tau^2), y\right) = \frac{1}{K}\sum_{k=t-K}^{t-1}\left(\sum_{m=1}^{M} wA(y_k - \eta_{m,k}, \tau^2) - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M} w^2 A(\eta_{m,k} - \eta_{n,k}, 2\tau^2)\right), \tag{3.24}$$

where $w = 1/M$ where $\sum_{m=1}^{M} w = 1$, and A is defined as Eq. 3.17. It is minimized numerically.

EMOS

Also for EMOS, we estimate the model parameters by minimizing the CRPS value for the training data. We calculate the minimum CRPS estimation by minimizing Equation 3.25. Hence, minimizing equation can be written as

$$\text{minCRPS}[\mathcal{N}(\mu, \sigma^2), y] = \frac{1}{K}\sum_{k=t-K}^{t-1}\left[\sigma\left(\frac{y_k - \mu_k}{\sigma}\left(2\Phi\left(\frac{y_k - \mu_k}{\sigma}\right) - 1\right) + 2\phi\left(\frac{y_k - \mu_k}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right)\right], \tag{3.25}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and the cdf of a standard normal random variable at the normalized prediction error, $(y-\mu)/\sigma$ [13, 18]. As for BMA, it is minimized numerically.

3.5. Software

In this thesis the software MATLAB is used for the analysis. Ensemble forecasts were loaded from GRIB-files. Gridded Information in Binary (GRIB) files are outputs directly from Numerical Weather Prediction programs. The toolbox *NCTOOLBOX* provides methods for data access.

4. Bivariate forecasts

In order to account for lead time dependencies between errors at a single location and single quantity, we need bivariate statistical post-processing techniques. In this chapter we introduce an extended version of the BMA and EMOS models for bivariate lead times. Let t be the issue time, l the lead time and m the number of an ensemble member in an ensemble forecast. We then denote $\mathbf{x}_{m,t} = \{x_{m,t,l} : l \in \mathbf{L}\}$ as a vector with ensemble member forecasts at lead time l of a set of lead times $\mathbf{L}$, and $\mathbf{y}_t = \{y_{t+l} : l \in \mathbf{L}\}$ the corresponding observation vector. For simplicity we suppress issue time from the notation and denote the ensemble member forecast and observation vector as $\mathbf{x}_m$ and $\mathbf{y}$. Additionally, we let $\mathbf{L} = \{1, 2\}$ label the elements in the vectors. This is only in order to simplify the notation and does not mean that we have set the two lead times to be $l = 1$ and $l = 2$.

After introducing the extended version of the BMA and EMOS models in section 4.1 and 4.2, we present how to estimate the bivariate parameters with the estimation methods by MLE in section 4.3. In section 4.4 we derive how to estimate the parameters with minimum ES (minES).

4.1. Bivariate Normal distributed BMA

Bivariate normal distributed BMA is an extended version of the univariate BMA model introduced in section 3.1. It takes two lead times into account at a time in order to account for lead time dependencies between errors. Let $\Theta^{bma} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, r\}$, where $\boldsymbol{\alpha} = [\alpha_1 \ \alpha_2]$', $\boldsymbol{\beta} = [\beta_1 \ \beta_2]$', $\boldsymbol{\tau}^2 = [\tau_1^2 \ \tau_2^2]$' and r the correlation coefficient. For exchangeable ensemble members, each pair of ensemble member forecasts is associated with a pdf, $p(\mathbf{y}|\mathbf{x}_1, ..., \mathbf{x}_M, \Theta^{bma})$. The bivariate normal BMA is a mixture of M bi-

variate Gaussian pdf's and equals

$$p(\mathbf{y}|\mathbf{x}_1, ..., \mathbf{x}_M, \Theta^{bma}) = \sum_{m=1}^{M} w p_m(\mathbf{y}|\mathbf{x}_m, \Theta^{bma}), \tag{4.1}$$

where M is the total number of forecasts in the ensemble, and the weights are $w = \frac{1}{M}$. Normal distribution is considered for temperature ensemble forecasts. The distribution is centred so that the conditional pdf, $p_m(\mathbf{y}|\mathbf{x}_m, \Theta^{bma})$ has a mean, $\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{x}_m$ and covariance matrix $\boldsymbol{\Sigma}^{bma}$. This can be denoted as

$$p_m(\mathbf{y}|\mathbf{x}_m, \Theta^{bma}) \sim \mathcal{MVN}(\boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{x}_m, \boldsymbol{\Sigma}^{bma}). \tag{4.2}$$

were $\mathbf{y} = [y_1\ y_2]$', $\mathbf{x}_m = [x_{1m}\ x_{2m}]$' and

$$\boldsymbol{\Sigma}^{bma} = \begin{bmatrix} \tau_1^2 & r\tau_1\tau_2 \\ r\tau_1\tau_2 & \tau_2^2 \end{bmatrix}. \tag{4.3}$$

Hence, the distribution $p_m(\mathbf{y}|\mathbf{x}_m, \Theta^{bma})$ can be written as

$$p_m\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \middle| \begin{bmatrix} x_{1m} \\ x_{2m} \end{bmatrix}, \Theta^{bma}\right) \sim \mathcal{MVN}\left(\begin{bmatrix} \alpha_1 + \beta_1 x_{1m} \\ \alpha_2 + \beta_2 x_{2m} \end{bmatrix}, \begin{bmatrix} \tau_1^2 & r\tau_1\tau_2 \\ r\tau_1\tau_2 & \tau_2^2 \end{bmatrix}\right). \tag{4.4}$$

In order to illustrate bivariate BMA predictive pdf, we let the parameters be $\boldsymbol{\alpha} = [0\ 0]$', $\boldsymbol{\beta} = [1\ 1]$', $\tau_1 = 0.2$, $\tau_2 = 0.2$, $\rho = 0.8$ and apply these on an ensemble forecasts with $M = 5$. Hence, the weights are $w = 1/5$. In Figure 4.1 the estimated probability density contours for the bivariate normal mixture distribution are plotted. We observe that the ensemble members are divided into two clusters. Values of the bias-corrected ensemble members (black dots) are also shown.

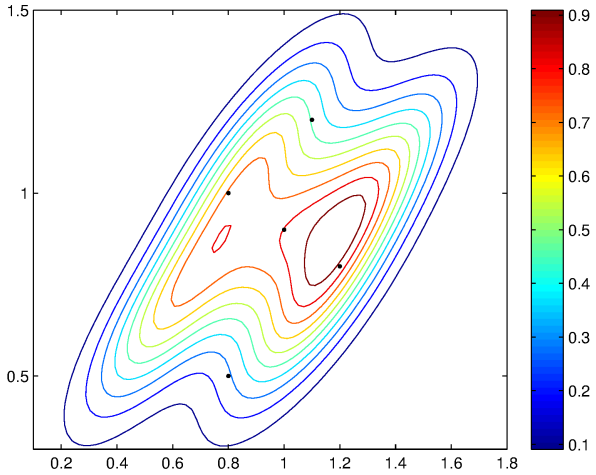


Figure 4.1: *Extended predictive BMA pdf. The parameters are* $\boldsymbol{\alpha}$ *= [0 0]', ***β*** = [1 1]',* $\tau_1 = 0.2$, $\tau_2 = 0.2$, $\rho = 0.8$ *and the weights are* $w = 1/5$.

4.2. Bivariate Normal distributed EMOS

The bivariate normal EMOS model is an extended version of the univariate EMOS model introduced in section 3.2. Let $\Theta^{emos} = \{\mathbf{a}, \mathbf{b}, \boldsymbol{\sigma}, \rho\}$, where $\boldsymbol{a} = [a_1 \ a_2]'$, $\boldsymbol{b} = [b_1 \ b_2]'$, $\boldsymbol{\sigma} = [\sigma_1^2 \ \sigma_2^2]'$ and ρ the correlation coefficient. The EMOS predictive pdf is normal distributed with predictive mean, $\boldsymbol{\mu} = \mathbf{a} + \mathbf{b}\bar{\mathbf{x}}$ and covariance $\boldsymbol{\Sigma}^{emos}$ and is denoted as

$$p(\mathbf{y}|\bar{\mathbf{x}}, \Theta^{emos}) \sim \mathcal{MVN}(\mathbf{a} + \mathbf{b}\bar{\mathbf{x}}, \boldsymbol{\Sigma}^{emos}), \tag{4.5}$$

where $\mathbf{y} = [y_1 \ y_2]'$, $\bar{\mathbf{x}} = [\bar{x}_1 \ \bar{x}_2]'$ and

$$\boldsymbol{\Sigma}^{emos} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{4.6}$$

Here the predictive variance is $\sigma^2_{\mathbf{L}} = c + dS^2_{\mathbf{L}}$, where $S^2_{\mathbf{L}} = \frac{1}{M}\sum_{m=1}^{M}(x_m - \bar{x})^2$. Remembering the notation, $\mathbf{L} = \{1,2\}$, the bivariate normal predictive EMOS pdf, $p(\mathbf{y}|\bar{\mathbf{x}}, \Theta^{emos})$, equals

$$p\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \middle| \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}, \Theta^{emos}\right) \sim \mathcal{MVN}\left(\begin{bmatrix} a_1 + b_1\bar{x}_1 \\ a_2 + b_2\bar{x}_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right). \tag{4.7}$$

In order to illustrate bivariate EMOS predictive pdf, we use the same parameters and ensemble forecast as in the example of bivariate predictive BMA pdf, see section 4.1. Figure 4.2 shows the estimated probability density contours for the bivariate normal distribution. Values of the bias-corrected ensemble members (black dots) are also shown.

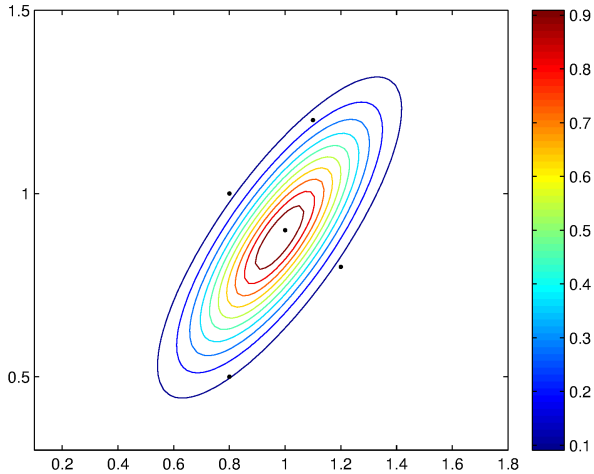


Figure 4.2: *Extended predictive EMOS pdf. The same parameters as for the extended predictive BMA pdf are used (see figure text 4.1).*

4.3. ML estimation for bivariate model

Given the training data $\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_{t-K}$ of K observations with conditional pdf $p(\mathbf{y}_{t-k}|\mathbf{x}_{m,t-k}, \Theta)$. The log-likelihood function for the bivariate BMA is defined by

$$\log \mathcal{L}(\mathbf{x}_1, ..., \mathbf{x}_M, \mathbf{\Sigma}^{bma}|\mathbf{y}_{t-1}, ..., \mathbf{y}_{t-K}) = \sum_{k=t-K}^{t-1} \log \left[\sum_{m=1}^{M} w p_m(\mathbf{y}_k|\mathbf{x}_{m,k}, \mathbf{\Sigma}^{bma}) \right], \tag{4.8}$$

where $\mathbf{\Sigma}^{bma}$ is the covariance matrix in Eq. 4.3. The log-likelihood function for the bivariate EMOS is defined by

$$\log \mathcal{L}(\bar{\mathbf{x}}, \mathbf{\Sigma}^{emos}|\mathbf{y}_{t-1}, ..., \mathbf{y}_{t-K}) = \sum_{k=t-K}^{t-1} \log \left[p(\mathbf{y}_k|\bar{\mathbf{x}}_k, \mathbf{\Sigma}^{emos}) \right], \tag{4.9}$$

where $\mathbf{\Sigma}^{emos}$ is the covariance matrix in Eq. 4.6. In practice we estimate the parameters in two steps. For the normal distributed BMA model, we first estimate the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. This is done in the same way as we did for univariate models, see section 3.4. In the next step, we then estimate the correlation coefficient, r. Similarly for EMOS, we first estimate $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{d}$ before we estimate the correlation coefficient, ρ.

4.4. Minimum ES estimation for bivariate model

A second estimating method, minimum ES (minES), is proposed for estimating the bivariate model correlation parameter. For computational simplicity, we estimate the correlation coefficient with minES by sampling from the predictive BMA and EMOS distribution. For convenience the following trick is used:

Algorithm for estimating ρ

· Given a set of correlation coefficients, $\rho \in$ {-0.95:0.01:0.95};
for *each* ρ **do**
| · calculate the ES;
end
· find ρ giving the lowest ES;

Note that we here used the parameter notation for EMOS model. The same procedure is done for the BMA model. In more details, for one simulation, s, one training day, k, and one given ρ, we sample $\mathbf{y}_s^{emos}$ from a predictive EMOS distribution, $p(\mathbf{y}|\bar{\mathbf{x}}, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \rho)$, by the following algorithm:

Algorithm for sampling $\mathbf{y}_s^{emos}$

Input: $\bar{\mathbf{x}}, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \rho$;
· generate $\boldsymbol{\varepsilon}_s^{emos} \overset{\text{iid}}{\sim} \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}(\rho)^{emos})$;
· set $\mathbf{y}_s^{emos} = \mathbf{a} + \mathbf{b}\bar{\mathbf{x}} + \boldsymbol{\varepsilon}_s^{emos}$;

We note that $\mathbf{y}_s^{emos}$ varies with an independent normal distributed error term. For the BMA model we have to take into account the ensemble members. For one simulation, s, one training day, k, and one given r, we sample $\mathbf{y}_s^{bma}$ from a predictive BMA distribution, $p_m(\mathbf{y}|\mathbf{x}_m, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, r)$, by the following algorithm:

Algorithm for sampling $\mathbf{y}_s^{bma}$

Input: $\mathbf{x}_1, ..., \mathbf{x}_M, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, r$
· draw $m \overset{\text{iid}}{\sim} Unif[1, 50]$;
· generate $\boldsymbol{\varepsilon}_s^{bma} \overset{\text{iid}}{\sim} \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}(r)^{bma})$;
· $\mathbf{y}_s^{bma} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{x}_m + \boldsymbol{\varepsilon}_s^{bma}$;

In the algorithm above, an index m is drawn in order to select a random member from the ensemble forecasts. This ensemble member forecast is further used when sampling from the predictive BMA distribution. In order to generate normal random variables, $\boldsymbol{\varepsilon}_s^{emos} \overset{\text{iid}}{\sim} \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}(\rho)^{emos})$, the Cholesky decomposition is used. This decomposition commonly used in e.g. the Monte Carlo method [2]. For one training day, k, one given ρ and $s = 1, ..., N_s$ simulations, we generate $\boldsymbol{\varepsilon}_s^{emos}$ as follows:

Algorithm for generating ε_s

Input: $\rho, \boldsymbol{\Sigma}^{emos}$
· set $\boldsymbol{\Sigma}(\rho)^{emos} = \mathbf{Q}^{-1}(\rho)^{emos}$;
· find $\mathcal{L}$ such that $\mathcal{L}^T\mathcal{L} = \mathbf{Q}(\rho)^{emos}$;
for *1:N_s* **do**
  · generate $\mathbf{z} \overset{\text{iid}}{\sim} \mathcal{MVN}(\mathbf{0}, \mathbf{I})$;
  · set $\boldsymbol{\varepsilon}_s^{emos} = \mathcal{L}^T\mathbf{z}$
end

We compute the lower-triangular, $\mathcal{L}$, by decomposing the correlation matrix, $\boldsymbol{\Sigma}^{emos}$. Applying this to a standard normal distributed vector, $\mathbf{z}$, we obtain normal random variables, $\boldsymbol{\varepsilon}_s^{emos}$.

The same procedure is used for the BMA model. However, in order to get smoother Gaussian curves, we use the same generated $\mathbf{z}$ for the simulations for each given correlation parameter. How to find the number of simulations, N_s, is discussed in the next chapter. A more detailed algorithm for the estimation of the correlation coefficient for BMA model is given in the following algorithm:

Detailed algorithm for BMA

Input: $\mathbf{x}_1, ..., \mathbf{x}_M, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, r$
· draw $\vec{m} \overset{\text{iid}}{\sim} Unif[1, M]$ of vector size $[1\ N_s]$;
· generate $\mathbf{z} \overset{\text{iid}}{\sim} \mathcal{MVN}(\mathbf{0}, \mathbf{I})$ of matrix size $[2\ N_s]$;
for $r = -0.95 : 0.01 : 0.95$ **do**
  · set $\boldsymbol{\Sigma}(r)^{bma} = \mathbf{Q}^{-1}(r)^{bma}$;
  · find $\mathcal{L}$ such that $\mathcal{L}^T\mathcal{L} = \mathbf{Q}(r)^{bma}$;
  for k = *1:K* **do**
    for s = *1:*N_s **do**
      · set $\boldsymbol{\varepsilon}_s^{emos} = \mathcal{L}^T\mathbf{z}$;
      · set $\mathbf{y}_s^{bma} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{x}_m + \boldsymbol{\varepsilon}_s^{bma}$;
    end
    · calculate ES_k;
  end
  · calculate $\text{ES} = \frac{1}{K}\sum_{k=1}^{K} \text{ES}_k$;
end
· find r giving the lowest ES;

A similar algorithm is used for the EMOS model. However, in this case it is not necessary to draw a random index, m, since the EMOS model uses the mean of the ensemble members as input.

5. Simulation study

In this study the parameters of BMA and EMOS are estimated with two different methods: by MLE and minCRPS estimation for univariate models, and with MLE and minES estimation for bivariate models. A simulation study is performed in order to explore the properties of the estimates, τ_{MLE}, $\tau_{minCRPS}$, σ_{MLE}, $\sigma_{minCRPS}$, r_{MLE} and ρ_{MLE}. The subscripts MLE, minCRPS and minES denote the estimation method used to estimate the parameters. We do a different study for r_{minES} and ρ_{minES}, where we find out how many simulations which are needed to estimate the parameters with minES estimation. This is because a simulation optimization routine is used for minES estimation.

Univariate model

A training period of $K = 30$ days and ensemble forecasts for $l = 2$ are considered for univariate models. How the training period length is found is described in the next chapter.

The normal predictive pdf of the univariate BMA model is defined as $p_m(y|x_m, \theta^{bma}) \sim \mathcal{N}(\alpha + \beta x_m, \tau^2)$. In order to test whether the estimates, τ, are consistent, we sample normal mixture distributed observations, Y^{bma}, from a known pdf for each of the training days. First, we randomly extract an index i from a uniform distribution with length equal to the number of ensemble members in an ensemble forecast, $M = 50$. The ensemble member of the daily ensemble with index i, x_i, is then used as the mean when sampling the observation, Y^{bma}. Hence, given a standard deviation, τ_g, we sample the observations from the pdf $p(Y^{bma}|x_i, \tau_g) \sim \mathcal{N}(x_i, \tau_g)$. Furthermore, we estimate the standard deviation, τ, with both the MLE and minCRPS estimation using Y^{bma} as

observational input. If τ is similar to the given parameter, τ_g, we conclude that our estimates are consistent.

We sample the observation Y^{bma} for BMA for the given standard deviations $\tau_g = 0.01$, 0.1, 0.5, 1, 1.5 and 2. In order to compare the estimation methods, we use the same Y^{bma} for both the MLE and minCRPS estimation. We sample Y^{bma} 100 times for the same given τ_g in order to calculate a 95% confidence interval (CI).

Table 5.1 a) lists the result for the MLE estimates. We observe that τ_{MLE} is close to τ_g and that the confidence intervals are relatively small. We also note that τ_{MLE} is three out of six times greater than τ_g.

a) **BMA, MLE**

τ_g	τ_{MLE}	95% CI	Length of CI
0.01	0.0102	[0.0098, 0.0106]	0.0008
0.1	0.1048	[0.0937, 0.1159]	0.0222
0.5	0.4748	[0.4553, 0.4944]	0.0391
1.0	0.9734	[0.9457, 1.0010]	0.0553
1.5	1.4669	[1.4329, 1.5010]	0.0681
2.0	2.0236	[1.9726, 2.0746]	0.1020

b) **BMA, minCRPS**

τ_g	$\tau_{minCRPS}$	95% CI	Length of CI
0.01	0.0194	[0.0780, 0.0109]	0.0671
0.1	0.1335	[0.1181, 0.1489]	0.0308
0.5	0.4578	[0.4360, 0.4796]	0.0436
1.0	0.9417	[0.9108, 0.9727]	0.0619
1.5	1.4275	[1.3897, 1.4654]	0.0757
2.0	1.9847	[1.9254, 2.0440]	0.1186

Table 5.1: *Simulation study for τ_{MLE} and $\tau_{minCRSP}$. In order to calculate a 95% confidence interval (CI), the observation Y^{bma} is sampled 100 times for the same given standard deviation, τ_g. A training length of $K = 30$ days and data from $l = 2$ is used.*

For minCRPS estimates, we observe that $\tau_{minCRPS}$ is not as close to τ_g as when we use the MLE method, see Table 5.1 b). Additionally, we note that the confidence intervals are slightly larger for $\tau_{minCRPS}$ than for τ_{MLE}. We also observe that $\tau_{minCRPS}$ mostly lies below τ_g. Hence, the minCRPS estimation method seems to underestimate the parameter.

a) **EMOS, MLE**

σ_g	σ_{MLE}	95% CI	Length of CI
0.01	0.0101	[0.0098, 0.0105]	0.0007
0.1	0.1055	[0.1011, 0.1099]	0.0088
0.5	0.5326	[0.5123, 0.5529]	0.0406
1.0	1.0548	[1.0096, 1.0999]	0.0903
1.5	1.6091	[1.5405, 1.6776]	0.1371
2.0	2.0937	[1.9990, 2.1884]	0.1894

b) **EMOS, minCRPS**

σ_g	$\sigma_{minCRPS}$	95% CI	Length of CI
0.01	0.0095	[0.0092, 0.0097]	0.0005
0.1	0.0970	[0.0939, 0.1001]	0.0062
0.5	0.4959	[0.4800, 0.5119]	0.0319
1.0	0.9694	[0.9365, 1.0024]	0.0659
1.5	1.4656	[1.4154, 1.5157]	0.1003
2.0	1.9288	[1.8635, 1.9942]	0.1307

Table 5.2: *Simulation study for σ_{MLE} and $\sigma_{minCRSP}$. In order to calculate a 95% confidence interval (CI), the observation Y^{emos} is sampled 100 times for the same given standard deviation, σ_g. A training length of $K = 30$ days and data from $l = 2$ is used.*

The normal distributed predictive pdf for the univariate EMOS model is defined as $p(y|\bar{x}, \theta^{emos}) \sim \mathcal{N}(a + b\bar{x}, c + dS^2)$. We test whether the EMOS predictive standard deviations, $\sigma = c + dS^2$, are consistent in a similar way as for the univariate BMA model. However, for EMOS the mean of the ensemble forecasts, $\bar{x}$, is used as the mean when sampling the observations Y^{emos}. There is therefore no need for extracting an index i in this case. Hence, given a predictive standard deviation, σ_g, we sample the observations from the pdf $p(y_s^{emos}|\bar{x}, \sigma_g) \sim \mathcal{N}(\bar{x}, \sigma_g^2)$. Furthermore, we estimate

the predictive standard deviation, σ, with both the MLE and minCRPS estimation using Y^{emos} as observational input.

The observation Y^{emos} for EMOS is sampled with the given standard deviations, $\sigma_g = 0.01$, 0.1, 0.5, 1, 1.5 and 2. In a similar way as for BMA, Y^{emos} is sampled 100 times for each given predictive standard deviation, σ_g. In order to compare the estimation methods, we use the same Y^{emos} for both the MLE and minCRPS estimation. We observe that σ_{MLE} is close to σ_g most of the times, and that the confidence intervals are relatively small, see Table 5.2 a). However, σ_{MLE} is greater than σ_g for all cases. Hence, it seems like MLE overestimates the EMOS estimates.

Table 5.2 b) lists the results for minCRPS estimates. We observe that $\sigma_{minCRPS}$ is smaller than σ_g for all cases. Hence, it seems like minCRPS underestimates the EMOS parameters. Furthermore, we note that the confidence intervals are smaller for $\sigma_{minCRPS}$ than for σ_{MLE}.

Bivariate model

A training period of $K = 80$ days and ensemble forecasts for $l = 5$ and $l = 6$ are used for bivariate models with MLE estimates. How the training period is found is discussed in the next chapter.

We remember that for bivariate models, the ensemble members are vectors, $\mathbf{x}_m = \{x, l : l \in \mathbf{L}\}$. Here we suppress the issue time. The normal bivariate predictive pdf of the BMA model is defined by Eq. 4.1. The process is similar as for univariate BMA, where we first extract an index $i \overset{\text{iid}}{\sim} Unif[1, M]$. The ensemble member with this index, $\mathbf{x}_i$, is then used as input for the mean when sampling the observations, $\mathbf{Y}^{bma}$. Hence, given a correlation coefficient, r_g, we sample bivariate normal mixture distributed observations, $\mathbf{Y}^{bma}$, from a known pdf

$$p_m(\mathbf{y}|\mathbf{x}_i, \mathbf{\Sigma}_g^{bma}) \sim \mathcal{MVN}(\mathbf{x}_i, \mathbf{\Sigma}_g^{bma}), \tag{5.1}$$

where

$$\mathbf{\Sigma}_g^{bma} = \begin{bmatrix} \tau_1^2 & r_g\tau_1\tau_2 \\ r_g\tau_1\tau_2 & \tau_2^2 \end{bmatrix}. \tag{5.2}$$

We sample $\mathbf{Y}^{bma}$ for the given correlation coefficients $r_g = 0.01$, 0.1, 0.5 and 0.8. As for the univariate models, $\mathbf{Y}^{bma}$ is sampled 100 times for each r_g in order to calculate a 95% confidence interval. However, for bivariate models, the estimation process is done in two steps. First, the standard deviations are estimated. Next, we estimate the correlation coefficient. Table 5.3 a) shows the result for BMA with MLE estimates. We observe that r_{MLE} is relatively close to r_g, except for when $r_g = 0.5$. Furthermore, we note that three out of four times r_{MLE} is greater than r_g. This indicates that MLE overestimates the correlation coefficient for the BMA model.

a) **BMA, MLE**

r_g	r_{MLE}	95% CI	Length of CI
0.01	0.0207	[-0.0058, 0.0472]	0.0530
0.1	0.1112	[0.0864, 0.1359]	0.0495
0.5	0.4311	[0.4135, 0.4486]	0.0351
0.8	0.8123	[0.7987, 0.8569]	0.0582

b) **EMOS, MLE**

σ_g	ρ_{MLE}	95% CI	Length of CI
0.01	0.0101	[-0.0177,0.0379]	0.0556
0.1	0.0995	[0.0720, 0.1269]	0.0549
0.5	0.4920	[0.4734, 0.5106]	0.0372
0.8	0.7810	[0.7711, 0.7908]	0.0197

Table 5.3: *a) Simulation study for r_{MLE}. b) Simulation study for $\rho_M LE$. In order to calculate a 95% confidence interval (CI), the observation Y^{bma} and Y^{emos} is sampled 100 times for the same given standard deviation, r_g and ρ_g, respectively. A training length of $K = 80$ days and data from $l = 5$ and $l = 6$ is used.*

The normal bivariate predictive pdf of EMOS model is defined by Eq. 4.5. When sampling the observations, $\mathbf{y}_s^{emos}$, the mean of the daily ensemble, $\bar{\mathbf{x}}$, is used as input for the mean. Given a correlation coefficient ρ_g, we sample $\mathbf{y}_s^{emos}$ from a known pdf

$$p_m(\mathbf{y}|\bar{\mathbf{x}}, \mathbf{\Sigma}_g^{emos}) \sim \mathcal{MVN}(\bar{\mathbf{x}}, \mathbf{\Sigma}_g^{emos}), \tag{5.3}$$

where

$$\mathbf{\Sigma}_g^{emos} = \begin{bmatrix} \sigma_1^2 & \rho_g \sigma_1 \sigma_2 \\ \rho_g \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{5.4}$$

Table 5.3 b) shows the estimates, ρ_{MLE}, for EMOS. We observe that ρ_{MLE} is close to ρ_g. Furthermore, we note that three out of four times ρ_{MLE} is smaller than ρ_g. This indicates that MLE underestimates the correlation coefficient for the EMOS model. We also note that the width of the confidence intervals for ρ_{MLE} are similar to the with of the confidence interval for r_{MLE}.

Furthermore, we do a simulation study for minES estimates, where we find out how many simulations, s, which are needed to optimize a consistent correlation coefficient. For both the BMA and EMOS models, we use data from year 2007 as training period and test the estimates on year 2008. We test the different simulation lengths $s = 50, 100, 150, ..., 300$. The same optimization routine as explained in subsection 4.4 is used. Figure 5.1 shows the result for each simulation length. We observe that BMA converges to ES = 1.53 after slightly more than 150 simulations. EMOS converges to 1.47 after around 150 simulations. Hence, we choose to use 200 simulations for BMA and 150 simulations for EMOS when estimating the parameters with minES.

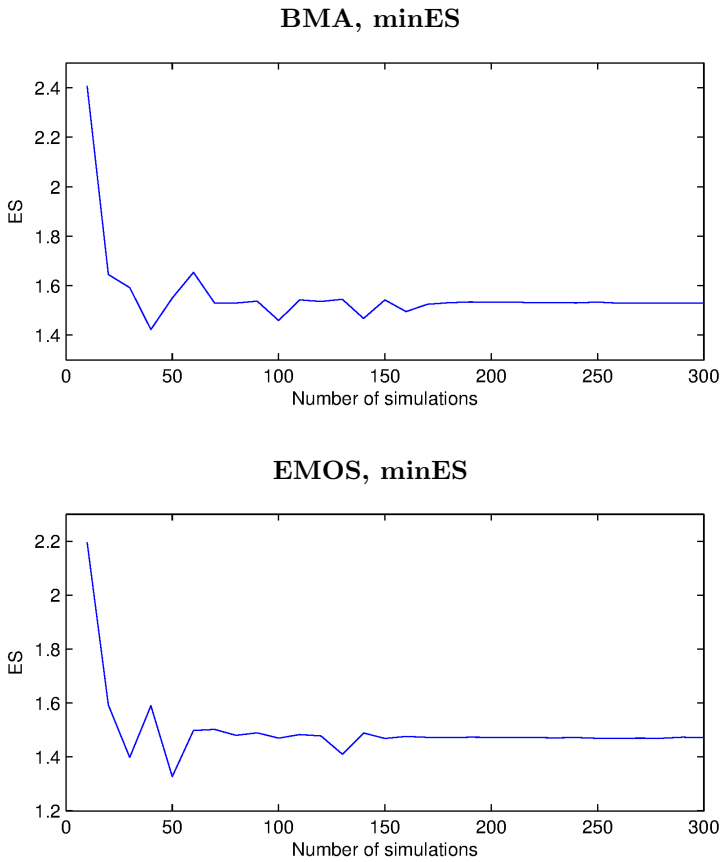


Figure 5.1: *ES for $l = 5$ and $l = 6$ using different lengths of simulations. Data from year 2007 is used as training period. The estimates are tested on year 2008.*

6. Case study: Temperature forecast of Trondheim - Voll

In this chapter, we first introduce the raw ensemble data before we describe how to choose the length of the training period. Furthermore, we present the results of applying the BMA and EMOS techniques for both univariate and bivariate models on ensemble forecasts of 2-m temperature in Trondheim – Voll. Ensemble forecasts from ECMWF between January 1st 2007 and December 31st 2011 were used. During this period, there were $T = 1826$ corresponding observations for each lead time, $l = 0$, 1, 2, 3, and $T = 1825$ corresponding observations for each $l = 4$, 5, 6, 7. The number of ensemble members is $M = 50$. Furthermore, we compare two different estimation methods, MLE and minCRPS for univariate models, and MLE and minES for bivariate models. For univariate models we use the approaches and estimation methods on all 8 different lead times. The extended bivariate BMA and EMOS techniques are only applied on $l = 5$ and 6.

6.1. Raw ensemble data

As mentioned in the introduction, ensemble forecasts tend to be underdispersive [19]. In order to see that this also is true in our case, verification rank histograms for all lead times are plotted, see Figure 6.1. We observe that the histograms are slightly U-shaped. However, we note that the corresponding observations more often fell above than below the ensemble range for all lead times. Thus, our raw ensemble forecasts are biased and underdispersed. Hence, post-processing techniques are needed in order to calibrate them.

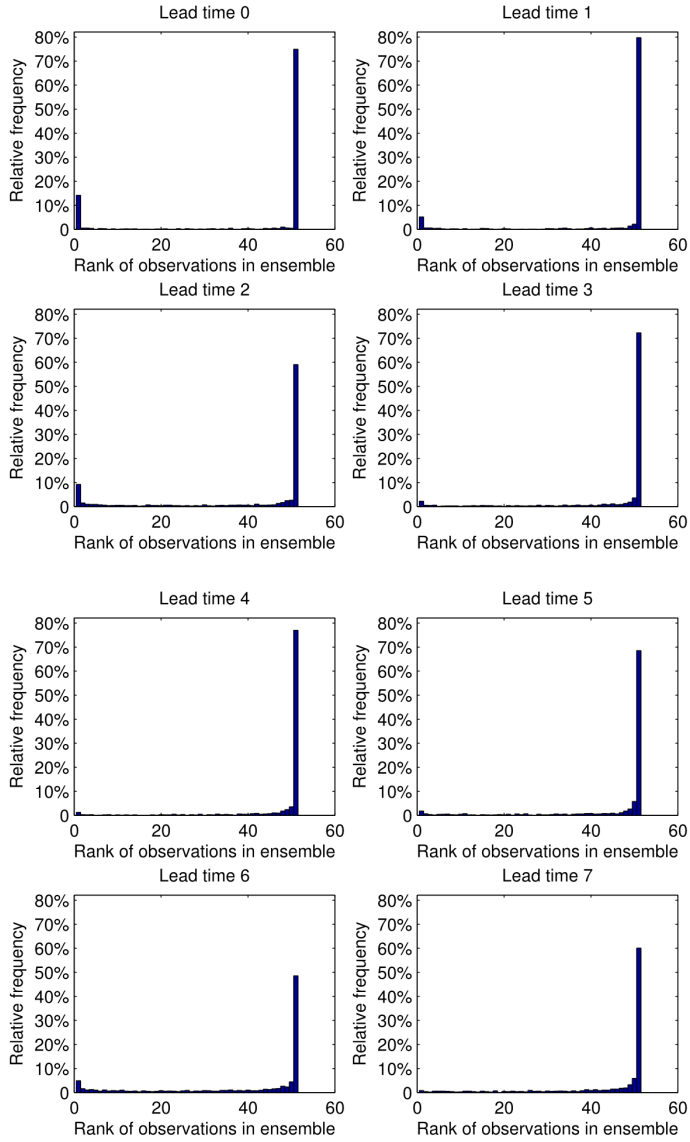


Figure 6.1: *Verification rank histogram for the raw ensemble forecasts for all lead times. The entire verification period between January 1st 2007 and December 31st 2011 was used. The height of the bars in each histogram indicates the percentage of cases for which the observation fell in each of the 51 bins.*

6.2. Length of training period

The model parameters are estimated using data from a "sliding window" training period consisting of the previous K days. A training period should be a balance between rapidly adapting of seasonal variation and give precise estimates of parameters. Hence, it is a trade-off: we want the training period to be as short as possible, but on the other hand, the longer the training period, the better estimates of the parameters. According to Gneiting et al. [12] one should maximize the sharpness subject to calibration. Therefore proper scoring rules, as e.g. CRPS are useful tools when choosing the training length. Raftery et al. [26] used a training period of 25 days to estimate temperature parameters to the BMA model. Gneiting et al. [16] chose a sliding 40-day training period for EMOS temperature forecasts. In order to find out how many training days we should use for the univariate models for our data, we minimise MAE and CRPS. For the bivariate models we minimise ES.

Univariate model

The minimisation of MAE and CRPS is done for both the univariate BMA and EMOS models, and for both the MLE and minCRPS estimation methods. Hence, we find the length of the training period for four different combinations. We consider the training period lengths, $K = 10, 15, 20, ..., 50$ for $l = 2$ to see which of these gives the lowest score. The results are plotted in Figure 6.2, showing MAE and CRPS as a function of the training period lengths. We observe that all four combinations decrease up to $K = 30$ for both MLE and CRPS. After that, they increase slightly at $K = 35$ before they decreases again. There seems to be little difference in MAE and CRPS between $K = 30$ and $K = 40$, 45, 50. Remembering that there is an advantage of using a short training period, we choose a training period of $K = 30$.

Bivariate model

We find the length of the training period for the bivariate BMA and EMOS models by minimising ES. This is only done for the MLE method, since the parameters with minES are found by sampling. For simplicity, we therefore choose to use the same training period length, found for MLE estimation, for minES estimation. Hence, for bivariate models, we only

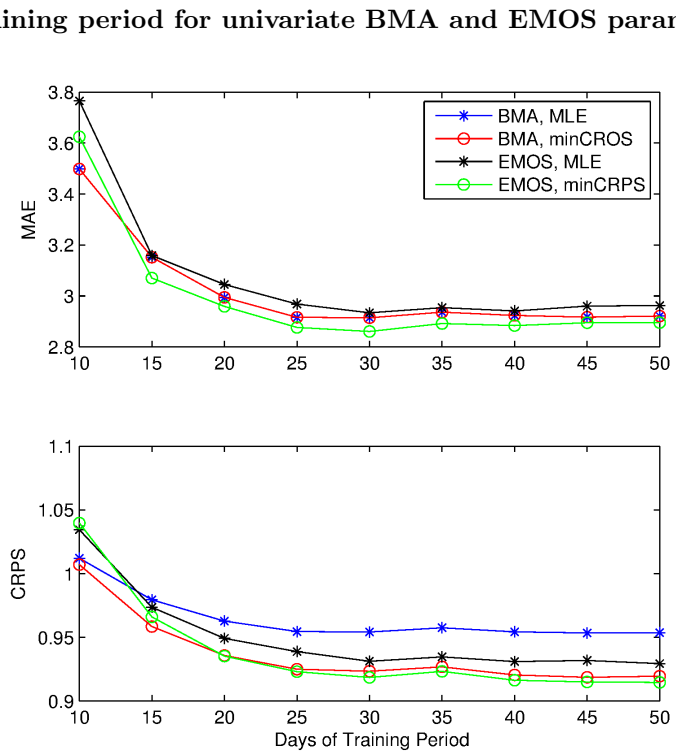


Figure 6.2: *Comparison of training period lengths for temperature at lead time 2. Both a plot for BMA parameters estimated with MLE and with minCRPS is shown.*

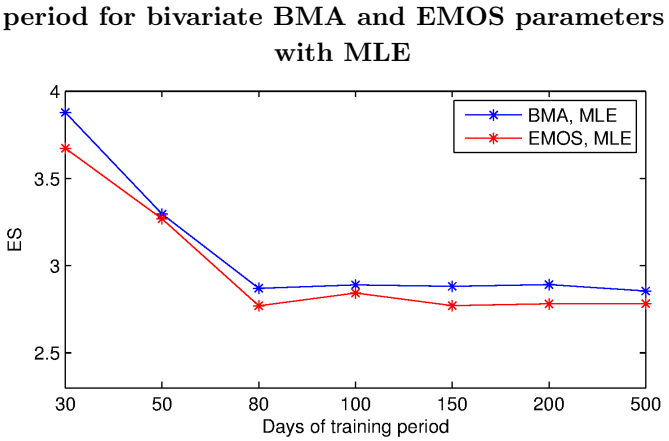


Figure 6.3: *Comparison of training period lengths for temperature at lead time 5 and 6. The plot shows ES of multivariate BMA and EMOS model.*

find the length of training period for two different settings.

We consider the training periods lengths, $K = 30$, 50,..., 500 for $l = 5$ and 6. The results are plotted in Figure 6.3, showing ES as a function of the training period lengths. We observe that the bivariate EMOS model has the lowest ES with length of training period, $K = 80$. The bivariate BMA model has the lowest ES with $K = 500$ training days. However, the ES is not very different between $K = 80$ and $K = 500$ for the BMA model. Hence, we choose to use a training period of $K = 80$ for the bivariate models. We note that bivariate BMA gives higher ES than the bivariate EMOS model.

6.3. Univariate model

In this section, we present results for the univariate BMA and EMOS models, which are presented in section 3.1 and 3.2, respectively. A training period of $K = 30$ days is applied to the data presented in Chapter 2.

In the next subsections, assessment methods, explained in section 3.3, are used to evaluate the univariate models. First calibration is assessed by PIT-histograms and prediction intervals. Furthermore, sharpness is eval-

uated by proper scoring rules, which were explained in subsection 3.3.2. We also compare the predictive standard deviations before the CRPS is calculated.

6.3.1. Assessment of univariate predictive performance

The PIT-histogram is the most common method for identifying calibration of probabilistic forecast [8]. A verification rank histogram (VRH) of the bias-corrected ensemble forecasts, shows the importance of post-processing, see Figure 6.4. The U-shape indicates that the bias-corrected ensemble forecasts are very underdispersed for all lead times. This means that the observations fell as much below as above the ensemble range. We also note that the observations much more often fell outside the ensemble range for $l = 0$ than for e.g. $l = 7$.

Figures 6.5 – 6.8 shows the PIT-histograms for the BMA and EMOS models. We observe that these are close to uniform. Hence, they are very well calibrated. Figures 6.5 and 6.6 show the PIT-histograms for the BMA model with parameters estimated with MLE and minCRPS, respectively. Comparing the estimation methods, we observe that the PIT-histograms are almost identical. The same result is observed for the EMOS model, see Figure 6.7 and 6.8. This indicates that there is little difference between the estimation methods. However, we note that BMA give slightly different PIT-histograms than the EMOS. For lead time, $l = 0$ to $l = 6$ the PIT-histograms for the EMOS models seem to be slightly U-shaped. This indicates underdispersion. For lead time 6 and 7 they seem to be more hump-shaped, which indicates overdispersion. It seems like the BMA model is slightly better calibrated than the EMOS model.

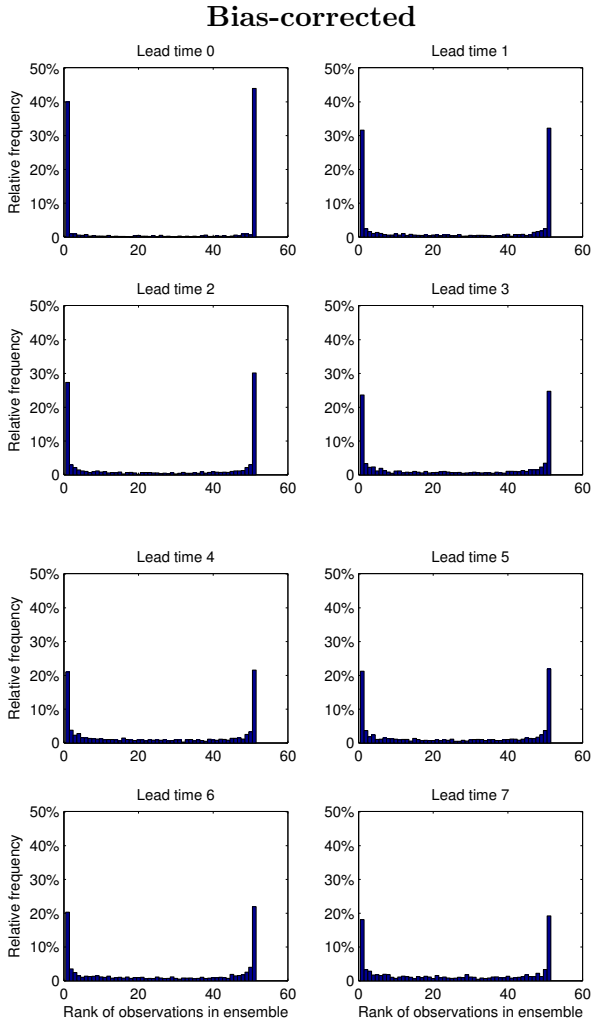


Figure 6.4: *Verification rank histogram for the bias-corrected ensemble forecast for all lead times from 2007 to 2011. The height of the bars in each histogram indicates the percentage of cases for which the observation fell in each of the 51 bins.*

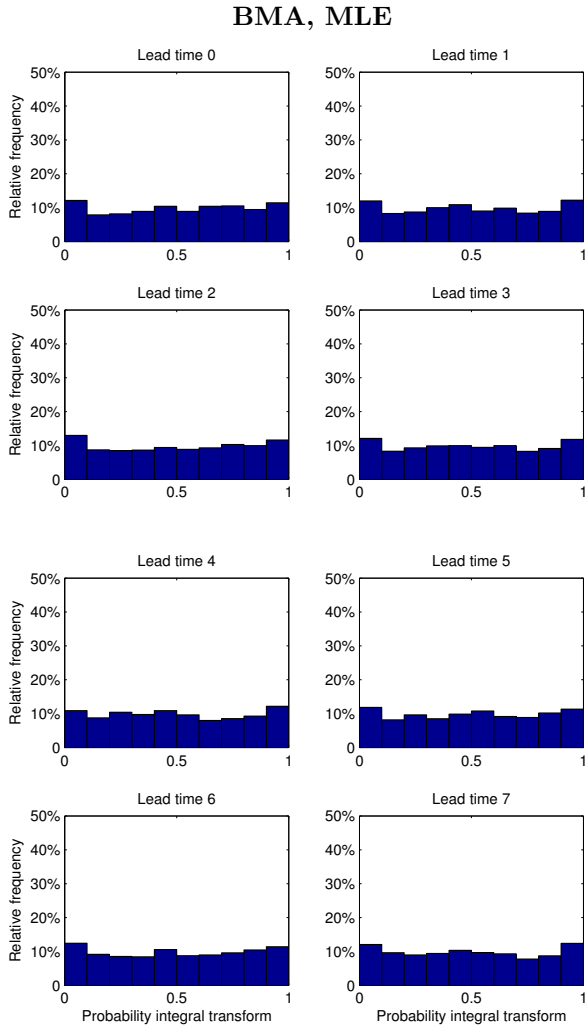


Figure 6.5: *PIT histogram for BMA ensemble forecast for all lead times from 2007 to 2011. The height of the bars in each histogram indicates the percentage of cases for which the observation fell in each of the 51 bins.*

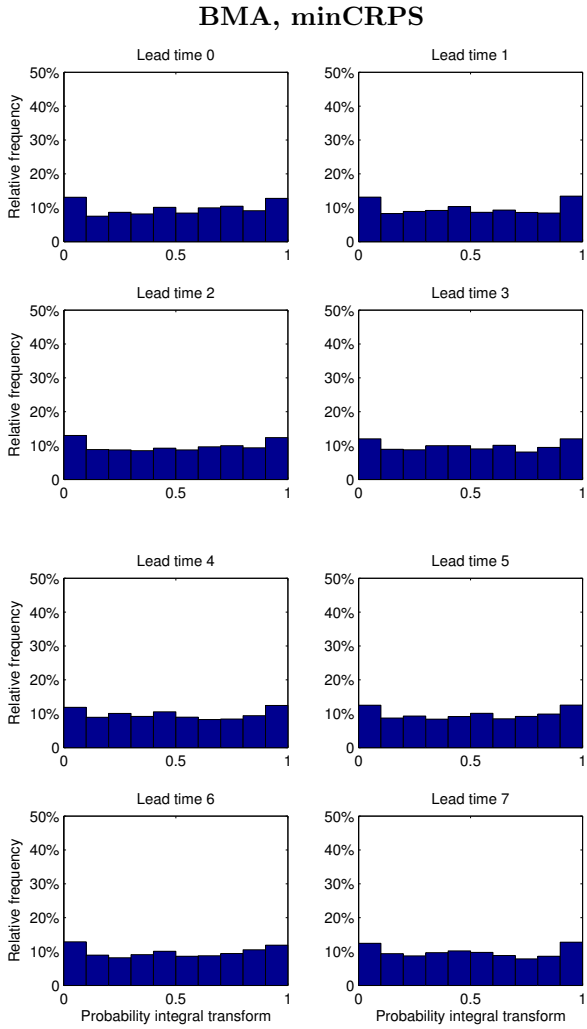


Figure 6.6: *PIT histogram for BMA ensemble forecast for all lead times from 2007 to 2011. The height of the bars in each histogram indicates the percentage of cases for which the observation fell in each of the 51 bins.*

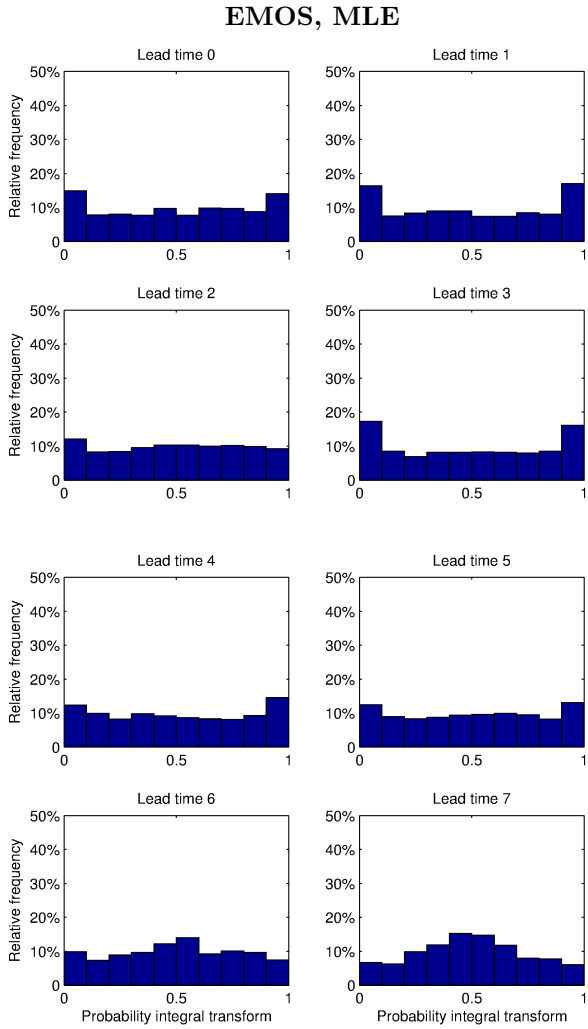


Figure 6.7: *PIT histogram for EMOS forecast for all lead times from 2007 to 2011. The height of the bars in each histogram indicates the percentage of cases for which the observation fell in each of the 51 bins.*

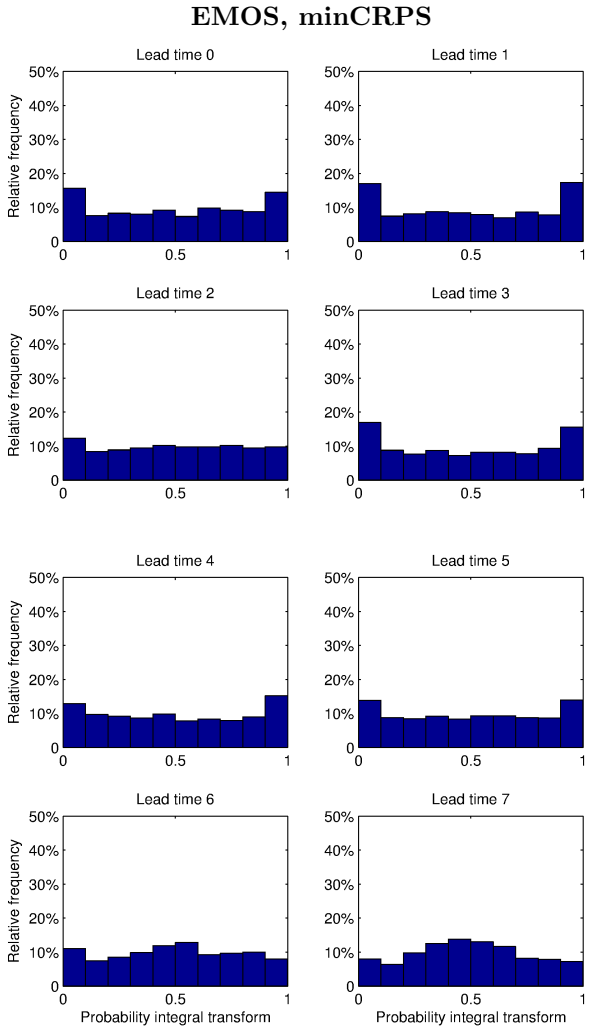


Figure 6.8: *PIT histogram for BMA ensemble forecast for all lead times from 2007 to 2011. The height of the bars in each histogram indicates the percentage of cases for which the observation fell in each of the 51 bins.*

Another method for measuring calibration is to find the coverage of prediction intervals. Figures 6.9 and 6.10 show the percentage of observations within a certain prediction interval for raw ensemble forecasts and bias-corrected ensemble forecasts, respectively. The dashed line shows where the values would be if the ensemble members were perfectly calibrated. We observe that the forecasts are calibrated for non of the lead times, neither for raw or bias-corrected ensemble forecasts. However, we note that the bias-corrected ensemble forecasts are slightly closer to be calibrated than the raw ensemble forecasts.

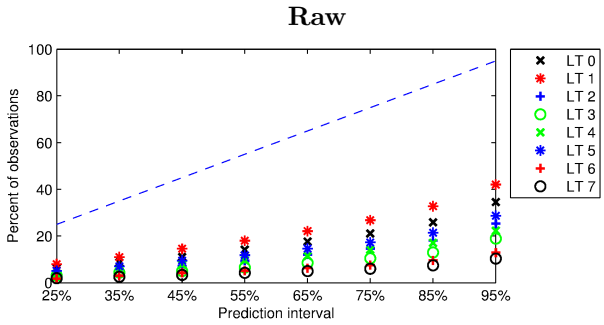


Figure 6.9: *Percentage of observations within a certain prediction interval for raw ensemble forecasts. The dashed line shows where the values would have been if the ensemble members were perfectly calibrated. Each lead time (LT) is plotted.*

Figure 6.11 shows the BMA and EMOS predictive pdf's with MLE estimates. We observe that these ensemble forecasts are almost perfectly calibrated up to a 55% prediction interval. For the reminding prediction intervals, the percentage of observations lie slightly below.

Figure 6.12 shows the BMA and EMOS predictive pdf's with minCRPS estimates. We observe that BMA and EMOS are almost perfectly calibrated up to a 45% prediction interval. For the reminding prediction intervals, the percentage of observations lie slightly below.

Comparing the CRPS estimates with the MLE estimates, we observe that the minCRPS estimates give slightly less calibrated ensemble forecasts than the MLE estimates. Comparing the models, they seem to give similar

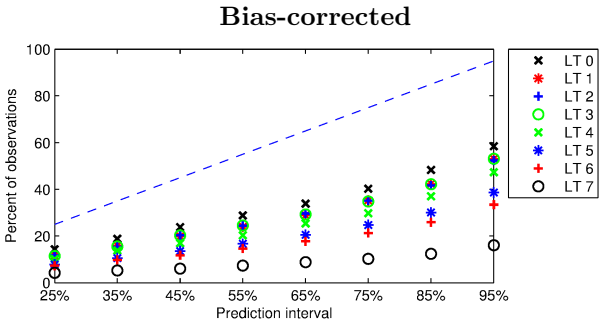


Figure 6.10: *Percentage of observations within a certain prediction interval for bias-corrected ensemble forecasts. The dashed line shows where the values would have been if the ensemble members were perfectly calibrated. Each lead time (LT) is plotted.*

results. This is the same result as we observed when evaluating calibration with PIT-histograms.

However, although the PIT-histogram and prediction intervals are good tools for evaluating calibration of probabilistic forecasts, it is not sufficient to tell if a post-processing technique is useful or not. We therefore assess the sharpness of BMA and EMOS by evaluating the width of 95% prediction intervals.

Figure 6.13 shows how sharp the BMA and EMOS predictive pdf's are for MLE and minCRPS estimates. We observe that all four cases are quite similar, and that the prediction intervals become wider for each lead time. The width of the prediction intervals should ideally be shorter for more accurate predictions. It is hard to tell which of the models and estimation methods make the sharpest ensemble forecasts. We therefore examine the BMA and EMOS predictive standard deviations closer. The predictive BMA standard deviations are denoted as σ^{bma}_{MLE}, $\sigma^{bma}_{minCRPS}$ and are calculated from Eq. 3.4. Similarly, the predictive EMOS standard deviations from Eq. 3.6 are denoted as σ^{emos}_{MLE}, $\sigma^{emos}_{minCRPS}$. The subscript MLE and minCRPS denote the estimation method used to estimate the parameters.

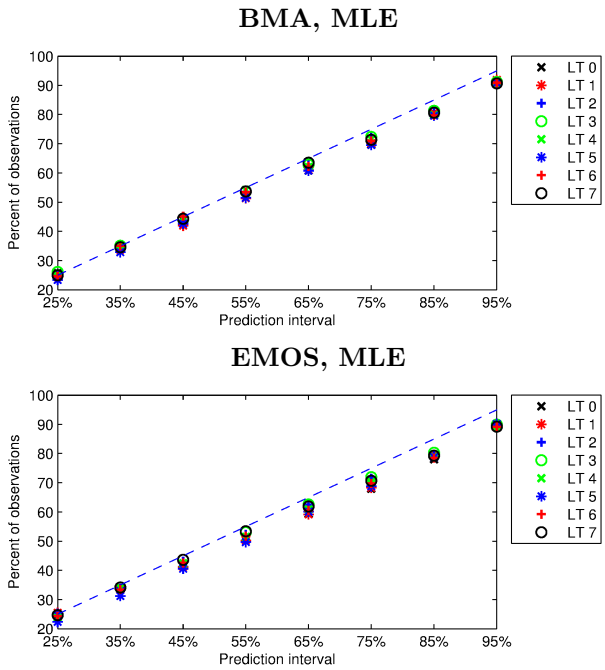


Figure 6.11: *Percentage of observations within a certain prediction interval for BMA and EMOS forecasts with MLE estimates. The dashed line shows where the values would have been if the ensemble members were perfectly calibrated. Each lead time (LT) is plotted.*

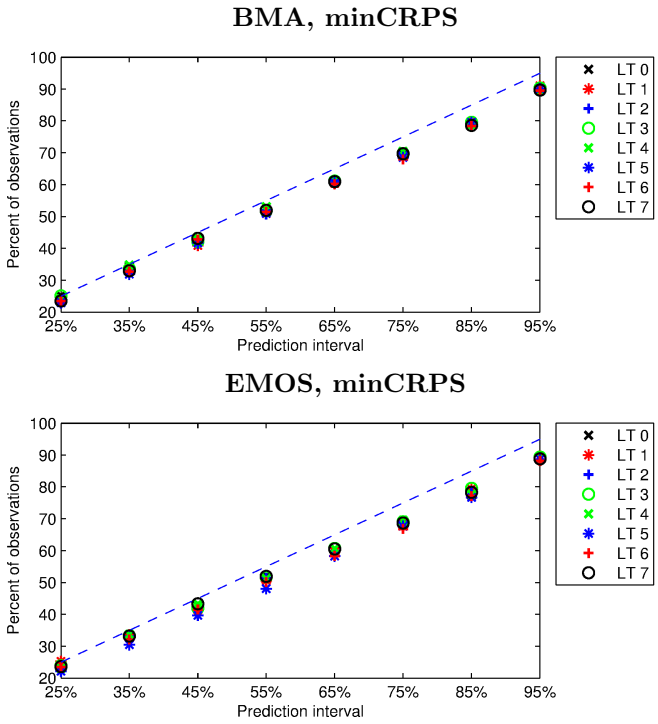


Figure 6.12: *Percentage of observations within a certain prediction interval for BMA and EMOS forecasts with minCRPS estimates. The dashed line shows where the values would have been if the ensemble members were perfectly calibrated. Each lead time (LT) is plotted.*

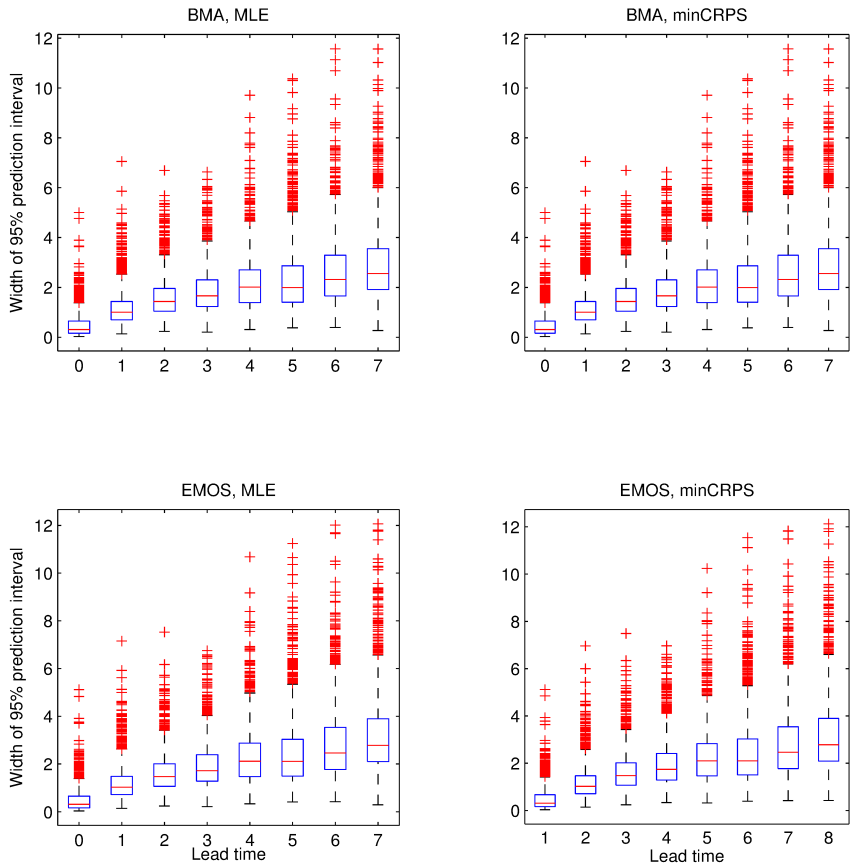


Figure 6.13: *Width of 95% prediction interval for each lead time between year 2007 and 2011. The blue box shows the 25-75% quantiles, the red line the median and the red crosses the outliers.*

The predictive standard deviations were estimated for the entire verification period for each lead times. σ^{bma}_{MLE} and $\sigma^{bma}_{minCRPS}$ are very similar for both estimation methods, MLE and minCRPS (see Figure 6.14). We also note that σ^{bma} grows with higher lead times. However, we do observe small differences between the BMA predictive standard deviations. This is better observed in Figure 6.15, where the difference between the BMA predictive standard deviations, $\sigma^{bma}_{MLE} - \sigma^{bma}_{minCRPS}$ are shown. There are indications of σ^{bma}_{MLE} lying slightly above $\sigma^{bma}_{minCRPS}$ most of the time for all lead times. This can especially be observed in extreme situations. For example, around April 2011 there was a storm caused by low air pressure meeting high pressure from Great Britain [23]. This made it hard to predict the temperature, and we note that MLE estimates were clearly higher than minCRPS estimates around that time for all lead times. We also note that the differences between σ^{bma}_{MLE} and $\sigma^{bma}_{minCRPS}$ are quite similar for all lead times. The intercept in Figure 6.14 are σ^{bma}_{MLE} and $\sigma^{bma}_{minCRPS}$ when fitting models using all $K = 1825$ days. We note that σ^{bma}_{MLE} lies above $\sigma^{bma}_{minCRPS}$ for all lead times. However, the values of σ^{bma} vary little from lead time to lead time.

Furthermore, we observe that also σ^{emos}_{MLE} and $\sigma^{emos}_{minCRPS}$ are very similar for both estimation methods, MLE and minCRPS, but only up to $l = 5$, see Figure 6.16. For $l = 6$ and 7 we note that σ^{emos} grows significantly, which also leads to bigger differences in the estimates. This is better observed in Figure 6.17, where the differences between the EMOS predictive standard deviations, $\sigma^{emos}_{MLE} - \sigma^{emos}_{minCRPS}$ are plotted. As for BMA predictive standard deviations, it also seems like σ^{emos}_{MLE} lies slightly above $\sigma^{emos}_{minCRPS}$ most of the time for all lead times. Furthermore, we again note that MLE estimates are higher than minCRPS estimates in extreme situations. However, the error between σ_{MLE} and $\sigma_{minCRPS}$ is more stable for BMA than for EMOS. Additionally, we note that the intercept in Figure 6.16 varies from lead time to lead time. However, again σ^{emos}_{MLE} lies above $\sigma^{emos}_{minCRPS}$ for all lead times.

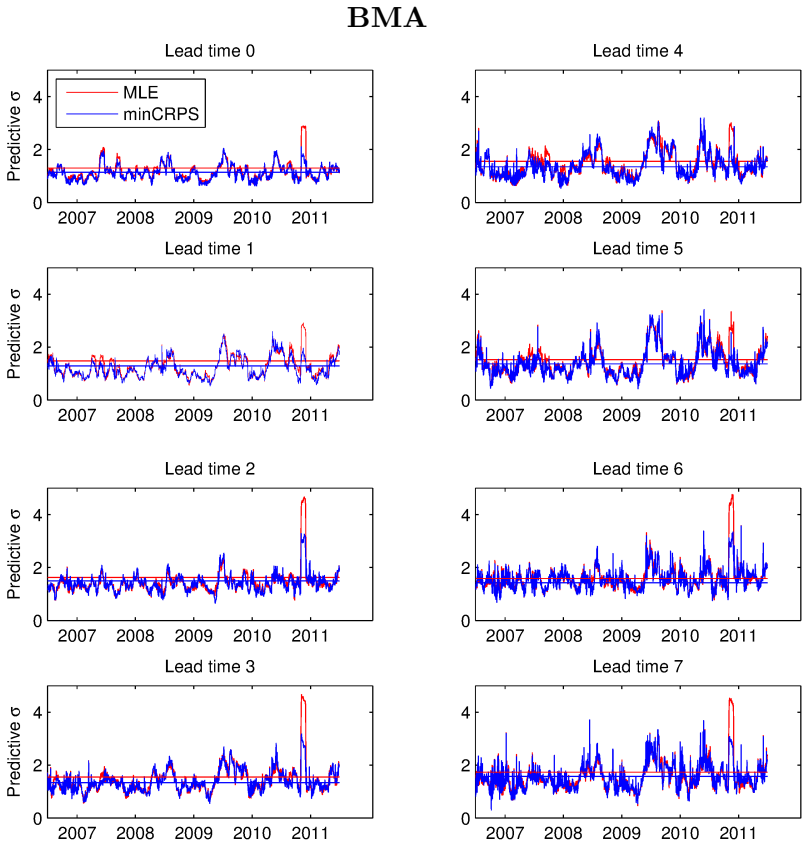


Figure 6.14: *BMA predictive standard deviation estimated over the entire verification period for all lead times with both MLE and minCRPS. The straight lines are the predictive σ using all years as training period.*

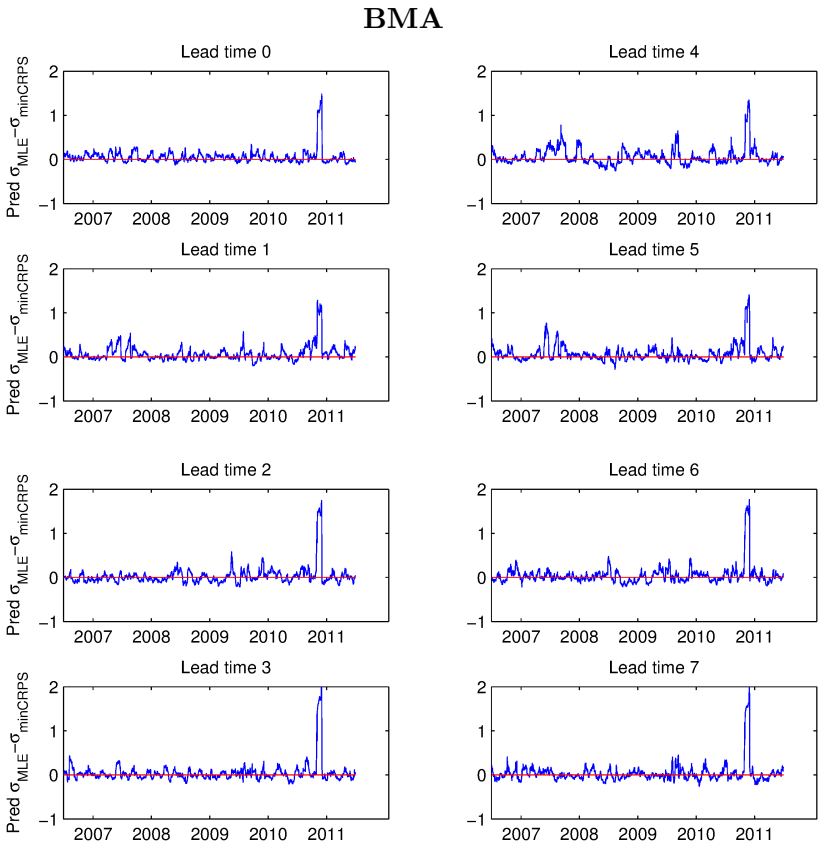


Figure 6.15: *Difference between the MLE and minCRPS predictive standard deviations.*

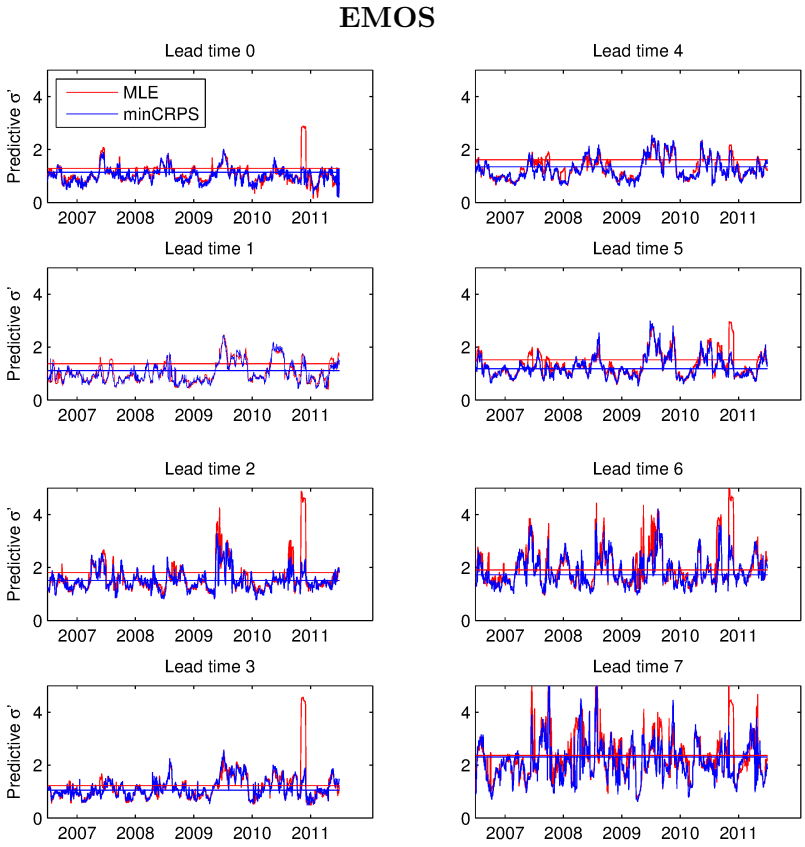


Figure 6.16: *EMOS predictive standard deviation estimated over the entire verification period for all lead times with both MLE and minCRPS. The straight lines are the predictive σ' using all years as training period.*

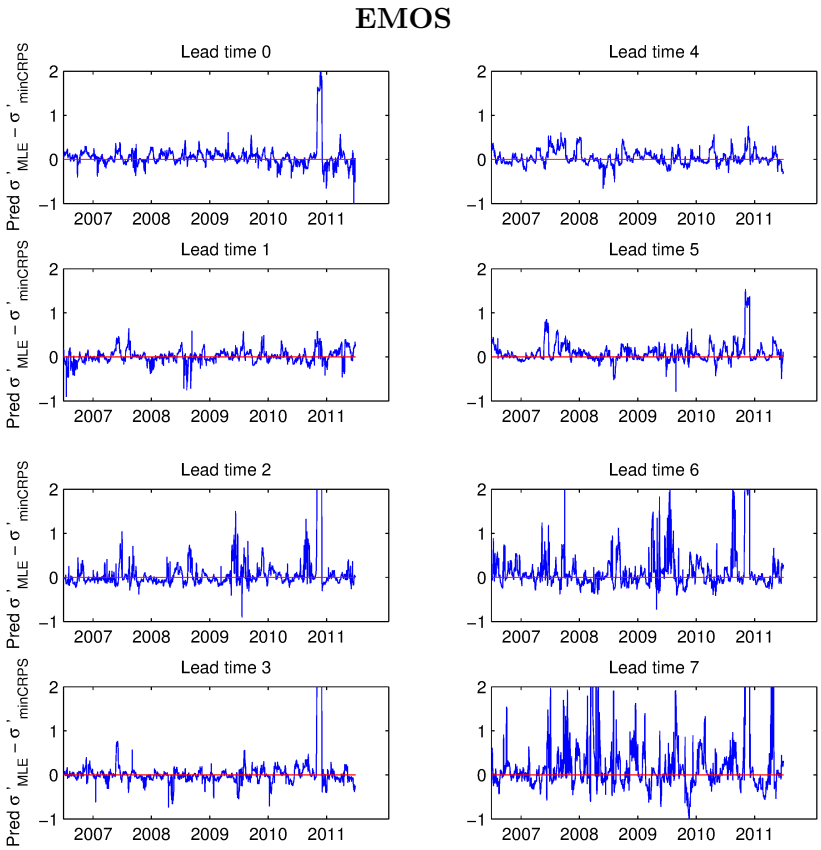


Figure 6.17: *Difference between MLE and minCRPS predictive standard deviations.*

The CRPS gives a joint assessment of calibration and sharpness. We recall that for a deterministic forecast the CRPS reduces to the MAE, hence we can compare these two scores. Table 6.1 lists the resulting CRPS for the deterministic and probabilistic forecasts using $K = 30$ days as training period. The scores are better for the BMA and EMOS techniques than for raw and bias-corrected ensemble forecasts. This is also observed in Figure 6.18, where the scores from Table 6.1 are plotted. The mean of raw ensemble forecasts give by far the poorest score. Comparing only the predictive BMA and EMOS performance, we note that $\tau_{minCRPS}$ gives overall the lowest CRPS. Hence, the BMA model with minCRPS estimates give the most calibrated and sharpest ensemble forecasts.

CRPS $K = 30$	l**=0**	l**=1**	l**=2**	l**=3**
Raw ensemble forecast	1.37	1.80	1.57	2.07
Mean of raw ensemble forecast	1.44	1.96	1.78	2.32
Mean of bias-corr. ensemble forecast	1.03	1.12	1.28	1.24
Bias-corr. ensemble forecast	0.96	0.98	1.10	1.04
BMA τ_{MLE}	**0.75**	**0.82**	**0.92**	**0.91**
BMA $\tau_{minCRPS}$	**0.75**	**0.82**	**0.92**	**0.91**
EMOS σ_{MLE}	0.76	0.84	0.93	0.93
EMOS $\sigma_{minCRPS}$	**0.75**	**0.82**	**0.92**	0.92
CRPS $K = 30$	l**=4**	l**=5**	l**=6**	l**=7**
Raw ensemble forecast	2.32	2.03	1.67	2.13
Mean of raw ensemble forecast	2.62	2.34	2.00	2.52
Mean of bias-corr. ensemble forecast	1.27	1.26	1.40	1.37
Bias-corr. ensemble forecast	1.04	1.03	1.14	1.11
BMA τ_{MLE}	0.92	**0.92**	**1.02**	1.01
BMA $\tau_{minCRPS}$	**0.91**	**0.92**	**1.02**	**1.00**
EMOS σ_{MLE}	0.92	0.93	1.04	1.05
EMOS $\sigma_{minCRPS}$	0.92	**0.92**	**1.02**	1.03

Table 6.1: *CRPS for temperature forecasts between year 2007 and 2011. The bias-corrected ensemble forecast, BMA τ_{MLE}, $\tau_{minCRPS}$ and EMOS σ_{MLE}, $\sigma_{minCRPS}$ were trained on a sliding 30-day period.*

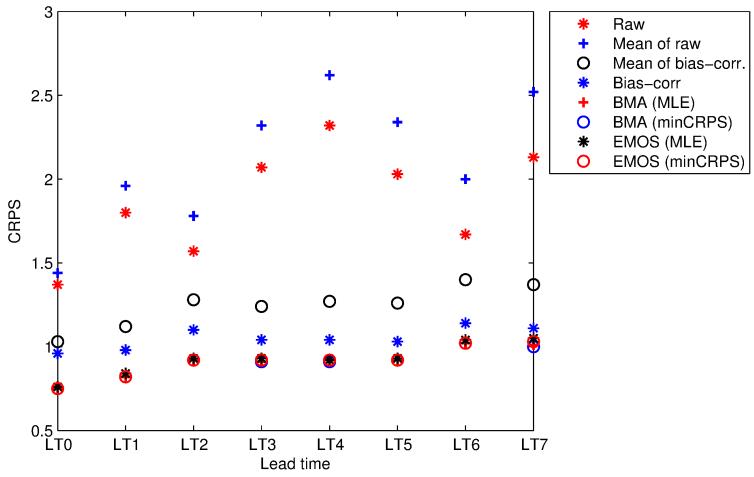


Figure 6.18: *CRPS for temperature forecasts using a $K = 30$ day training period. The accurate values are listed in Table 6.1*

CRPS $K = 1825$	l**=0**	l**=1**	l**=2**	l**=3**
BMA τ_{MLE}	0.76	0.81	**0.91**	**0.92**
BMA $\tau_{minCRPS}$	0.76	0.81	**0.91**	**0.92**
EMOS σ_{MLE}	0.76	0.81	**0.91**	**0.92**
EMOS $\sigma_{minCRPS}$	**0.75**	**0.80**	**0.91**	**0.92**
CRPS $K = 1825$	l**=4**	l**=5**	l**=6**	l**=7**
BMA τ_{MLE}	**0.95**	**0.92**	1.02	1.01
BMA $\tau_{minCRPS}$	**0.95**	**0.92**	**1.00**	1.01
EMOS σ_{MLE}	**0.95**	**0.92**	**1.00**	1.01
EMOS $\sigma_{minCRPS}$	**0.95**	**0.92**	**1.00**	**1.00**

Table 6.2: *CRPS for temperature forecasts where the model was fitted to all data, $K = 1825$ is used.*

The temperature in Trondheim can be hard to predict due to a lot of changes in the weather during the day. It might therefore be hard to adapt these varieties with a training period of only 30 days. Hence, we fitted a model using all $K = 1825$ available days. However, we observe that the scores are very similar to the scores when using a training period of $K = 30$ days, see Table 6.2. Thus, this time the EMOS method with minCRPS estimates gives the best score.

Taking into account the PIT-histogram, CRPS and prediction interval using a $K = 30$ day training period, the BMA model with minCRPS estimates gives the most calibrated and sharpest ensemble forecasts. The BMA model might be better than EMOS because single ensemble members give more information than using only the mean of the ensemble members, $\bar{x}$. Moreover, the minCRPS estimation method might be better than MLE because the minCRPS is more robust. MLE performs best when the model is perfect. However, when fitting a model using all $K = 1825$ days, the EMOS model with minCRPS gives the best score.

6.4. Bivariate model

In this section we present the results for the bivariate BMA and EMOS models. These are extended versions of the univariate BMA and EMOS models presented by Raftery et al. [26] and Gneiting et al. [16], respectively. We use two different estimation methods, MLE and minES to estimate the model parameters. In section 6.2 we found that a training length of $K = 80$ days is sufficient for bivariate models. However, instead of using a sliding training period, we choose to fit the models using one year, $S = 365$ days, of data. The parameters are evaluated on the next year. We also fitted a seasonal model using one season, $S = 92$ days, of data and evaluated the parameters on the same season the next year. The seasons are divided into: Winter (Dec. – Feb.), Spring (Mar. – May), Summer (Jan. – Aug.) and Autumn (Sept. – Nov.).

Correlation

Before we start applying the extended approaches to ensemble forecasts, we are interested in finding out if there is lead time correlation between errors. In the previous section we saw that the estimation methods, MLE

and minCRPS, were quite similar. Hence, we only consider the difference between the observations and the median of the BMA and EMOS probabilistic forecasts with minCRPS estimates and suppress the MLE estimates. The errors are denoted as:

- $e_{t,l} = y_{t+l} - median[x_{m,t,l}]$
- $e_{t,l}^{bma} = y_{t+l} - median[p(y|x_{m,t,l}, \theta^{mle})]$
- $e_{t,l}^{emos} = y_{t+l} - median[p(y|\bar{x}_{t,l}, \theta^{emos})]$

For each of the errors, we calculate the empirical correlations, shown in Table 6.3, 6.4 and 6.5, respectively. In all three matrices we observe that there is small lead time dependencies between errors. The bold numbers in the matrices indicate the highest correlation in each row. We note that the highest empirical lead time correlation between errors is when l and $l+1$ for l = 1,...,6. This is observed in all three matrices. Furthermore, there is little correlation between $l = 0$ and the reminding lead times. This makes sense in that $l = 0$ was issued at the same time as it was observed. Thus, there is little uncertainty in the ensemble forecasts for $l = 0$ compared to the other lead times.

				$e_{t,l}$				
l	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	1.0	**0.30**	0.25	0.12	0.09	0.03	0.05	0.01
1		1.0	**0.53**	0.50	0.43	0.36	0.29	0.32
2			1.0	**0.65**	0.40	0.37	0.42	0.37
3				1.0	**0.64**	0.54	0.46	0.50
4					1.0	**0.67**	0.43	0.45
5						1.0	**0.60**	0.56
6							1.0	**0.70**
7								1.0

Table 6.3: *Correlation of error between observation and the median of the raw ensemble member forecasts; $e_{t,l} = y_{t+l} - median[x_{m,t,l}]$. The bold numbers indicate the highest correlation in each row.*

$e^{bma}_{t,l}$

l	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	1.0	**0.30**	0.27	0.16	0.09	0.02	0.05	0.03
1		1.0	**0.46**	0.40	0.33	0.20	0.15	0.14
2			1.0	**0.58**	0.35	0.23	0.25	0.16
3				1.0	**0.58**	0.40	0.31	0.30
4					1.0	**0.60**	0.38	0.33
5						1.0	**0.53**	0.44
6							1.0	**0.63**
7								1.0

Table 6.4: *Correlation of error between observation and the median of the BMA ensemble member forecasts;* $e^{bma}_{t,l} = y_{t+l} - median[p(y|x_{m,t,l}, \theta^{mle})]$.

$e^{emos}_{t,l}$

l	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	1.0	**0.29**	0.26	0.15	0.08	0.02	0.04	0.03
1		1.0	**0.43**	0.38	0.32	0.20	0.13	0.12
2			1.0	**0.57**	0.32	0.22	0.23	0.16
3				1.0	**0.56**	0.40	0.30	0.31
4					1.0	**0.58**	0.34	0.30
5						1.0	**0.51**	0.43
6							1.0	**0.62**
7								1.0

Table 6.5: *Correlation of error between observation and the EMOS ensemble member forecasts;* $e^{emos}_{t,l} = y_{t+l} - median[p(y|\bar{x}_{t,l}, \theta^{emos})]$.

Furthermore, we note that the raw ensemble forecasts show overall higher correlation between the errors than the BMA and EMOS forecasts. Lower correlation between the errors for BMA and EMOS might have been caused by the calibration process. Furthermore we observe that $e^{bma}_{t,l}$ shows slightly higher correlation than $e^{emos}_{t,l}$ between l and *l+1* for l = 0,...,6. However, we note that the correlation between l and *l+1* for l = 2,...,6 are stable, in that the correlation is almost the same. We therefore choose to focus on $l = 5$ and $l = 6$ in the parameter estimation and evaluation part.

Correlation coefficient

In this subsection, we apply the estimation methods, explained in section 4.3 and 4.4, on data for $l = 5$ and $l = 6$. Table 6.6 shows the estimates when fitting models using one year, $K = 365$ days, of data. Also the estimates when fitting a models using all data, $K = 1825$ days, are shown. We observe that there is correlation between $l = 5$ and 6. We note that the tables list the correlation parameter, r, for BMA, and the *predictive* correlation parameter, ρ, for EMOS.

	r		ρ	
Year	**BMA**		**EMOS**	
	MLE	**minES**	**MLE**	**minES**
2007	0.66	0.68	0.58	0.60
2008	0.70	0.67	0.68	0.63
2009	0.57	0.53	0.54	0.55
2010	0.78	0.74	0.64	0.68
All	0.73	0.71	0.59	0.63

Table 6.6: *Correlation coefficient, ρ between $l = 5$ and $l = 6$ using one year as training period. The bold numbers indicate the highest correlation coefficient for each year.*

Furthermore, a model is fitted using one season, $K = 92$ days, of data. Table 6.7 lists r and ρ for the different seasons. Comparing the seasons, we note that there is less correlation between the lead times during Summer than the other seasons for both BMA and EMOS.

Assessment of bivariate predictive performance

The performance of the models is evaluated with ES, which gives a joint assessment of calibration and sharpness. Table 6.8 lists the ES for BMA and EMOS when fitting the model to one year of data, $K = 365$ days, and applying the estimates on the next year. For example, $07 \Rightarrow 08$ in Table 6.8 means that all available data from year 2007 are used to fit the model, and that the estimates are evaluated on year 2008. The bold numbers show the lowest ES for each year.

	r		ρ	
Season	**BMA**		**EMOS**	
Winter	**MLE**	**minES**	**MLE**	**minES**
2007	0.73	0.77	0.63	0.65
2008	0.68	0.65	0.54	0.55
2009	0.91	0.89	0.74	0.76
2010	0.92	0.92	0.86	0.87
All	0.90	0.94	0.80	0.84
Spring				
2007	0.66	0.63	0.62	0.63
2008	0.73	0.67	0.55	0.56
2009	0.81	0.79	0.73	0.77
2010	0.67	0.61	0.52	0.57
All	0.74	0.76	0.62	0.62
Summer				
2007	0.39	0.36	0.35	0.33
2008	0.38	0.31	0.31	0.25
2009	0.25	0.26	0.15	0.22
2010	0.19	0.18	0.12	0.16
All	0.35	0.34	0.20	0.25
Autumn				
2007	0.67	0.67	0.51	0.54
2008	0.77	0.74	0.67	0.71
2009	0.79	0.71	0.67	0.63
2010	0.74	0.71	0.65	0.59
All	0.85	0.83	0.63	0.68

Table 6.7: *Correlation coefficient for $l = 5$ and $l = 6$ using one year as training period. The bold numbers indicate the highest correlation coefficient for each season and year.*

We observe than the EMOS model with minES estimates gives the lowest ES for all years except for one. For the BMA model the estimation methods give almost the same ES. However, they are always higher than for the EMOS model. This indicates that the EMOS model with minCRPS estimates is the most calibrated and sharpest model.

We evaluate the seasonal fitted models by applying the parameters on the next season, see Table 6.9. We again observe that the EMOS model with minES estimates gives the lowest ES.

ES

Year	**BMA**		**EMOS**	
	MLE	**minES**	**MLE**	**minES**
07 ⇒ 08	1.53	1.53	**1.46**	1.47
08 ⇒ 09	1.57	1.58	1.54	**1.52**
09 ⇒ 10	1.84	1.84	1.76	**1.75**
10 ⇒ 11	1.96	1.96	1.81	**1.80**
All ⇒ All	1.79	1.78	1.73	**1.71**

Table 6.8: *ES for $l = 5$ and $l = 6$ using one year as training period. The bold numbers indicate the lowest correlation coefficient for each year.*

Figure 6.19 shows the contour plot for both seasonal BMA and EMOS models with minCRPS estimates, applied on July 1st 2011. Here, the models are fitted on the data of the previous summer season, the summer of 2010. BMA makes use of a mixture distribution, in which each ensemble member corresponds to its own component. We therefore expect contour plots of BMA where more than one clustering is shown. However, we observe that the contour plots for the bivariate BMA models give similar results as for the bivariate EMOS models. This might be due to high standard deviations of the components. The same is observed in Figure 6.20, where the models are fitted to all summer season data, $K = 470$ days and applied on July 1st 2011.

ES

Season	BMA		EMOS	
Winter	**MLE**	**minES**	**MLE**	**minES**
07 ⇒ 08	1.83	1.86	1.71	**1.70**
08 ⇒ 09	1.86	1.86	**1.78**	1.79
09 ⇒ 10	2.65	2.69	2.39	**2.32**
10 ⇒ 11	2.56	2.54	2.37	**2.39**
All ⇒ All	2.09	2.06	**2.04**	**2.04**
Spring				
07 ⇒ 08	1.45	1.46	1.46	**1.43**
08 ⇒ 09	1.58	1.57	1.56	**1.53**
09 ⇒ 10	1.77	1.80	**1.68**	**1.68**
10 ⇒ 11	2.32	2.30	2.21	**2.20**
All ⇒ All	1.71	1.71	1.63	**1.62**
Summer				
07 ⇒ 08	1.45	1.46	1.41	**1.40**
08 ⇒ 09	1.42	1.39	1.28	**1.24**
09 ⇒ 10	1.45	1.44	1.37	**1.35**
10 ⇒ 11	1.78	1.78	1.48	**1.42**
All ⇒ All	1.55	1.55	1.37	**1.35**
Autumn				
07 ⇒ 08	1.71	1.72	1.57	**1.56**
08 ⇒ 09	1.56	1.59	**1.47**	1.49
09 ⇒ 10	2.04	2.03	1.95	**1.89**
10 ⇒ 11	1.99	2.04	1.67	**1.66**
All ⇒ All	1.68	1.68	1.57	**1.56**

Table 6.9: *ES for $l = 5$ and 6 using one season as training period. The bold numbers indicate the lowest correlation coefficient for each season.*

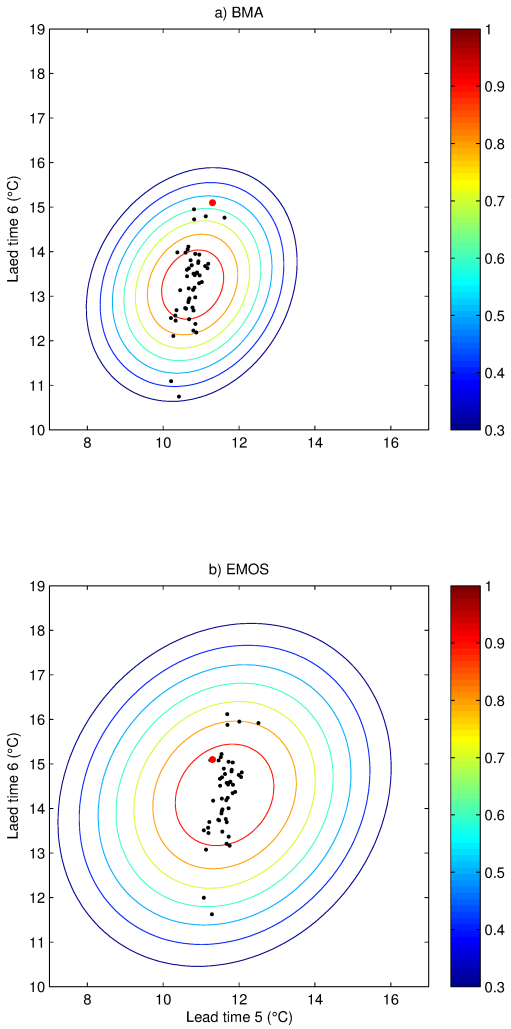


Figure 6.19: *a) Contour plot of BMA with minES estimates, applied on July 1st 2011. b) Contour plot of EMOS with minCRPS estimates applied on July 1st 2011. The model is fitted using all days of summer seasons 2010. The black dots shows the bias-corrected ensemble forecasts. The red dot show the observation that day.*

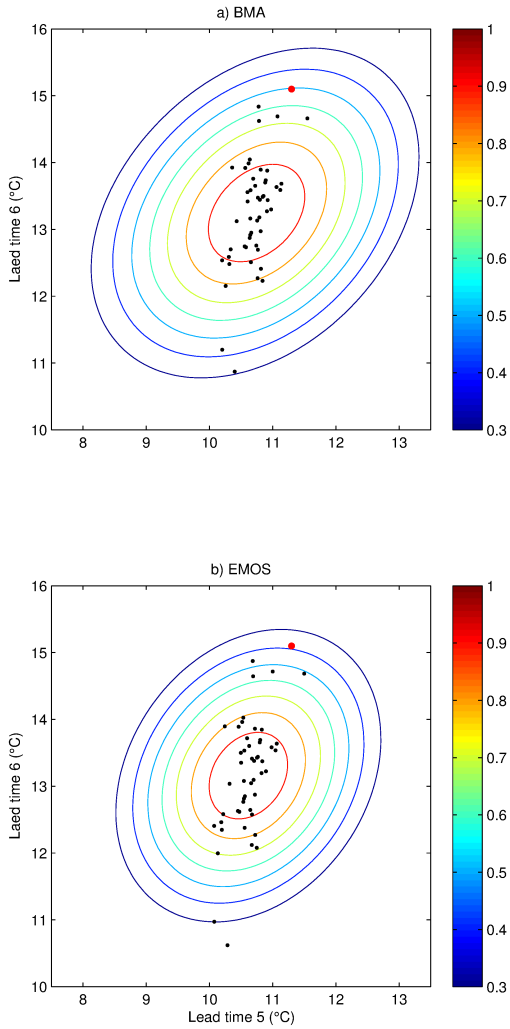


Figure 6.20: *a) Contour plot of BMA with minES estimates, applied on July 1st 2011. b) Contour plot of EMOS with minCRPS estimates applied on July 1st 2011. The model is fitted using all days of summer seasons between year 2007 and 2011. The black dots shows the bias-corrected ensemble forecasts. The red dot show the observation that day.*

7. Discussion and conclusion

In this study, we have used the existing BMA [26] and EMOS [16] post-processing techniques to obtain calibrated and sharp predictive probability distributions for temperature forecasts. These techniques were applied to a single weather quantity, at a single location only.

The most common method for estimation of BMA and EMOS parameters is the Maximum likelihood estimation (MLE) [26, 28]. This method has been used to estimate the parameters of both BMA and EMOS models. For EMOS parameter estimation, we have also seen that minimum CRPS (minCRPS) estimation, proposed by Gneiting et al. [16], has been used [16, 27]. However, to our knowledge, there is only a limited number of studies using the latter estimation method on BMA models. This motivated us to compare both estimation methods for both BMA and EMOS post-processing techniques. Hence, we evaluated four different settings to see if any of these performed differently. These were applied on 8 different lead times.

Furthermore, we proposed an extended bivariate BMA and EMOS model which accounts for correlation in errors. Two different estimation methods were also applied on these models. However, minCRPS estimation is only applicable on univariate models. Hence, we additionally proposed the minimum ES (minES) estimation method for bivariate models. Furthermore, we observed that the highest empirical lead time correlation between error occurs when $l = i$ and $l = i+1$, for $l = 1,...,6$. However, only $l = 5$ and 6 were considered .

The BMA and EMOS models were applied on 2-m surface temperature ensemble forecasts in Trondheim – Voll between January 2007 and December 2011, using European Center for Medium-range Weather Forecasts

(ECMWF) ensembles. A training period of 30 days was used for the univariate models. In order to get better results, longer training periods were necessary for bivariate models. Hence, a model was fitted on one year of data. In order to see if there was seasonal variation, we additionally used one season to fit the model. Following Raftery et al. [26], the distribution of forecast errors is approximated by a normal distribution.

We start by only comparing the models, suppressing that we also have used two different estimation methods. Our results showed well calibrated predictive pdf's for both models. All methods outperformed the raw ensemble forecasts. One might argue that there would be an advantage using BMA over EMOS because BMA models utilizes all ensemble members, while EMOS only makes use of the mean of these. Only using a single model often leads to an underestimation of the uncertainty in the process of the model selection. However, our results showed that both approaches yield nearly the same predictive performance. Still, considering the PIT-histograms, the univariate EMOS models were slightly overdispersed for higher lead times. For lower lead times, there were indications of underdispersiveness. The univariate BMA model seemed to perform slightly better than the EMOS model when only assessing the calibration. Thus, also the BMA model shows hints of underdispersiveness. Furthermore, the EMOS predictive standard deviation showed much more variation for higher lead times than the BMA predictive standard deviation. Yet, CRPS, giving a joint assessment of both calibration and sharpness, indicates that BMA only performs slightly better than the EMOS model for a training period of $K = 30$ days. When we fitted a model using all $K = 1825$ days, the EMOS model yielded an overall lower score than the BMA model. For bivariate models, we used both one year and one season as training period. In both cases the EMOS model outperformed the BMA model. This might be an indications that it is an advantage for EMOS models to use longer training periods.

BMA makes use of a mixture distribution, in which each ensemble member corresponds to its own component. We therefore expected to see contour plots of BMA where more than one clustering is shown. However, the contour plots for the bivariate BMA models give similar results as for the bivariate EMOS models. This might be due to high standard deviations of the components.

Two different training periods for bivariate models were used. First we fitted a model using one year of data. Further, we fitted a model using one season of data, where the parameters became a seasonal index. In both cases we observed that there is high lead time correlation between the errors. Furthermore, we noted that, when fitting a seasonal model, there is lower lead time correlation between the errors during summer than for the other seasons.

We continue with the comparison of the estimation methods. PIT-histograms showed very little difference in calibration between MLE and minCRPS estimation. However, plots of predictive standard deviations, σ, indicate that the MLE method estimates overall higher σ than minCRPS estimation. This was especially observed when fitting the model using all $K = 1825$ days. Furthermore, σ_{MLE} was higher in extreme situations than $\sigma_{minCRPS}$. However, CRPS and ES showed that minCRPS and minES performed better than MLE for univariate and bivariate models, respectively. The reason for this might be that minCRPS is more robust than MLE. The MLE method gives good estimates for perfect models, which we do not have in our case. For univariate models, the EMOS model with MLE estimates yields higher CRPS than both the BMA and EMOS models with minCRPS estimates. The same was observed for both training periods for bivariate models.

Overall, in our case study, the BMA and EMOS models give very similar results for univariate models. However, bivariate EMOS models perform better than bivariate BMA models. Comparing the estimation methods, minCRPS and minES estimation outperformed MLE. Hence, the BMA model with minCRPS estimation performs best for univariate models, when we consider a training period of $K = 30$. For bivariate models, EMOS with minES estimation gives the best results.

There are several directions into which our bivariate BMA and EMOS models could be developed. First of all, only two lead times were considered at a time. Hence, the model could also be extended to account for multivariate correlation. Additionally, our bivariate models are only applicable to temperature forecasts where normal distribution has been approximated. We could therefore extend the models so that one could

apply them to e.g. precipitation, where it seems more reasonable to approximate the conditional pdf by a gamma distribution [29]. Also more explanatory variables, as for example air pressure, could be taken into account in the temperature models.

Bibliography

[1] J. L. Anderson. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9(7):1518–1530, 1996.

[2] C. F. Ansley. An algorithm for the exact likelihood of a mixed autoregressive-moving average process. *Biometrika*, 66(1):59–65, 1979.

[3] V.J. Berrocal, A.E. Raftery, and T. Gneiting. Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, 135(4):1386–1402, 2007.

[4] V.J. Berrocal, A.E. Raftery, T. Gneiting, and R.C. Steed. Probabilistic weather forecasting for winter road maintenance. *Journal of the American Statistical Association*, 105(490):522–537, 2010.

[5] J. Bröcker and H. Kantz. The concept of exchangeability in ensemble forecasting. *Nonlinear Processes in Geophysics*, 18:1–5, 2011.

[6] R. Buizza, P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, 133(5):1076–1097, 2005.

[7] G. Candille and O. Talagrand. Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, 131(609):2131–2150, 2005.

[8] A. P. Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, pages 278–292, 1984.

[9] C. Fraley, A.E. Raftery, and T. Gneiting. Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review*, 138(1):190–202, 2010.

[10] H. R. Glahn and D. A Lowry. The use of model output statistics (MOS) in objective weather forecasting. *Journal of applied meteorology*, 11(8):1203–1211, 1972.

[11] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[12] T. Gneiting, F. Balabdaoui, and A.E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[13] T. Gneiting, K. Larson, K. Westrick, M.G. Genton, and E. Aldrich. Calibrated probabilistic forecasting at the stateline wind energy center. *Journal of the American Statistical Association*, 101(475):968–979, 2006.

[14] T. Gneiting and A. E. Raftery. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, 2005.

[15] T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[16] T. Gneiting, A.E. Raftery, A.H. Westveld III, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.

[17] E. P. Grimit and C. F. Mass. Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Weather and Forecasting*, 17(2):192–205, 2002.

[18] E.P. Grimit, T. Gneiting, VJ Berrocal, and N.A. Johnson. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132(621C):2925–2942, 2006.

[19] T. M. Hamill and S. J. Colucci. Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review*, 125(6):1312–1327, 1997.

[20] H. Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.

[21] L. Isaksen, J. Haseler, R. Buizza, and M. Leutbecher. The new ensemble of data assimilations. *ECMWF Newsletter, Spring*, pages 17–21, 2010.

[22] J. M. Lewis. Roots of ensemble forecasting. *Monthly weather review*, 133(7):1865–1885, 2005.

[23] T. Lygrell. Værboka 2012. ISBN 978-82-516-5517-0, 2012.

[24] T. N. Palmer. The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society*, 128(581):747–774, 2002.

[25] A. E. Raftery. Bayesian model selection in structural equation models. *Sage Focus Editions*, 154:163–163, 1993.

[26] A.E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005.

[27] N. Schuhen, T. L. Thorarinsdottir, and T. Gneiting. Ensemble model output statistics for wind vectors. *arXiv preprint arXiv:1201.2612*, 2012.

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

[29] J.M. Sloughter, A.E. Raftery, T. Gneiting, and C. Fraley. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, 135(9):3209–3220, 2007.