Ruprecht-Karls-Universitat Heidelberg
Fakultat fur Mathematik und Informatik

Statistical Postprocessing of Ensemble Forecasts for Temperature: The Importance of Spatial Modeling

Diplomarbeit von
Kira Feldmann

Juni 2012

Betreuer: Dr. Michael Scheuerer
Dr. Thordis Thorarinsdottir
Prof. Dr. Tilmann Gneiting

Abstract

In the recent past, ensemble prediction systems have become state of the art in the meteorological community. However, ensembles often underestimate the uncertainty in numerical weather prediction, resulting in underdispersive and thus uncalibrated forecasts. In order to employ the full potential of the ensemble, statistical postprocessing is needed. However, many of the current approaches, such as Bayesian model averaging (BMA) or ensemble model output statistics (EMOS), focus on forecasts at single locations, not taking spatial correlation between different observation sites into account. In this thesis, we discuss the existing method of spatial BMA, which combines the geostatistical output perturbation method (GOP) for modeling the spatial structure of the observation field, with BMA, in order to produce calibrated and sharp forecasts for whole weather fields. We propose a similar approach that employs EMOS instead of BMA. In a case study, we apply the methods to 21-hour ahead forecasts of surface temperature over Germany, issued by COSMO-DE-EPS. The multivariate forecasts were capable of capturing the spatial structure of the weather field and turn out to be calibrated and sharp, while showing an improvement over the raw ensemble as well as the reference forecasts.

Zusammenfassung

In der nahen Vergangenheit sind ensemblebasierte Wettervorhersagesysteme immer populärer geworden. Dennoch unterschätzen die Ensembles häufig die Unsicherheit numerischer Wettervorhersagen, wodurch die Vorhersagen unterdispersiv werden und folglich nicht kalibriert sind. Um die volle Leistungsfähigkeit des Ensembles zu entfalten, werden statistische Nachbearbeitungsverfahren benötigt. Jedoch konzentriert sich die Mehrheit dieser Methoden, wie unter anderem Bayesian model averaging (BMA) oder ensemble model output statistics (EMOS), auf Vorhersagen an einzelnen Orten und berücksichtigt dabei nicht die räumliche Korrelation zwischen den verschiedenen Beobachtungsstationen. In dieser Arbeit diskutieren wir die vorhandene Methode spatial BMA, welche die Kombination von zwei Nachbearbeitungsverfahren darstellt. Geostatistcal output perturbation method (GOP) modelliert die räumliche Struktur des Wetterfeldes und BMA inkludiert die Ensembleinformation, so dass kalibrierte und scharfe Wettervorhersagen für ganze Wetterfelder produziert werden. Wir schlagen zusätzlich einen ähnlichen Ansatz vor, der BMA durch EMOS ersetzt. In einer Fallstudie wenden wir die Methoden auf 21-stündige Temperaturvorhersagen von COSMO-DE-EPS für Deutschland an. Die multivariaten Vorhersagen schaffen es, die räumliche Struktur des Wetterfeldes wiederzugeben und erzeugen somit scharfe und kalibrierte Vorhersagen, die eine Verbesserung gegenüber dem unbearbeiteten Ensemble sowie den Referenzvorhersagemethoden darstellen.

Contents

Chapter 1

Introduction

”My interest is in the future because I am going to spend the rest of my life there.”
C.F. Kettering (1876-1958), inventor, scientist, engineer, businessman, philosopher

Predictions of the future have always been of great interest for mankind. Especially today, weather forecasts are a matter of high economical and social value, as they find applications in many different areas. Accurate predictions are essential for the growing field of renewable energies, management of air traffic, and natural disaster control, just to name a few.

In the past century, huge developments have been made in the field of forecasting weather quantities. With the advent of computer simulation, the rise of deterministic numerical weather prediction began in the early 1950s. At the same time, concerns over rigorous determinism, based on the principle that the future state of a system can be entirely described by its present state, started to grow (Lewis, 2005). Following the path of determinism, sources of uncertainty in numerical weather forecasts, such as imperfections in model formulation or insufficiency in the description of initial and boundary conditions, are not addressed (Leutbecher and Palmer, 2008).

In order to resolve these shortcomings, the first ensemble prediction systems were developed in the early 1990s (Lewis, 2005). An ensemble consists of multiple runs of the numerical weather prediction model, with variations in mathematical representations of the development of the atmosphere, initial conditions or lateral boundaries, and thus seeks to quantify the sources of uncertainty in deterministic forecasts. However, ensemble prediction systems are often underdispersed and tend to be biased (Hamill and Colucci, 1997).

To address these issues, a variety of statistical postprocessing methods for ensembles have been proposed (Wilks and Hamill, 2007). These models yield probabilistic forecasts, meaning that they deliver a predictive probability distribution for the weather variable of interest, which generally outperforms the raw ensemble in terms of satisfying the underlying goal of “maximizing sharpness subject to calibration” (Gneiting et al., 2007).

However, the majority of these methods is only applicable to univariate weather quantities at a single location and does not model spatial dependencies between different observation sites, which are of great importance when considering composite quantities, such as minima, maxima, totals or averages. These aggregated quantities are crucial e.g. for highway maintenance operations or flood management. Based on well-established univariate postprocessing techniques, probabilistic forecasts of any composite quantity of interest are straightforward to calculate. However, there is a high possibility that the predictive uncertainty is estimated poorly, as the model does not capture that the site-specific predictive uncertainties are correlated which has a great impact on the overall uncertainty (Thorarinsdottir et al., 2012).

In the case of deterministic temperature forecasts, Gel et al. (2004) propose the geostatistical output perturbation (GOP) method, that uses a geostatistical model in order to simulate spatially consistent temperature forecasts fields. In this thesis, we discuss the approach by Berrocal et al. (2007), that combines the univariate postprocessing method Baysian model averaging (BMA) with GOP. Based on ensemble prediction systems, BMA yields predictive probability density functions, which are weighted averages of densities centered at the bias-corrected forecasts of the individual ensemble members (Raftery et al., 2005). By uniting BMA and GOP, calibrated probabilistic forecasts of entire weather fields are produced based on ensemble prediction systems. In a similar way, we propose combining the univariate postprocessing method based on ensemble model output statistics (EMOS), which produces normal predictive density functions for temperature (Gneiting et al., 2005), with GOP in order to obtain the same goal.

The remainder of this thesis is organized as follows. Chapter 2 gives an introduction to COSMO-DE-EPS, the 20-member ensemble prediction system developed by the German Meteorological Service (DWD). We describe its construction and evaluate the predictive performance of 21-hour ahead forecasts of surface temperature over Germany in 2011. In Chapter 3, we provide details of the univariate postprocessing methods EMOS and BMA as well as a rather simple approach, based solely on least squares regression. In a case study, we apply these models to COSMO-DE-EPS and compare their predictive performance with reference forecasts. At the end of this chapter, we investigate an EMOS approach, where we replace the commonly used normal distribution with a Student's t-distribution, in order to cope with some of the challenges in statistical postprocessing of COSMO-DE-EPS. Chapter 4 describes the multivariate GOP and different ways of estimating the corresponding parameters. Then, we combine GOP with EMOS$^+$ and BMA respectively in order to produce calibrated and sharp forecasts for entire weather fields, followed by a case study based on COSMO-DE-EPS. The thesis closes with a discussion in Chapter 5, where we summarize the results, describe possible improvements of the methods, address some unresolved issues and hint at subjects of further research. The Appendix describes verification methods, compares the empirical variogram calculation of two different R packages and provides additional results for the models presented in Chapter 4.

Chapter 2

COSMO-DE-EPS

This chapter serves as an introduction to the ensemble forecasting system COSMO-DE-EPS (COnsortium for Small-scale MOdeling - DE - Ensemble Prediction System), which we will use as a basis for the competing postprocessing methods in the following chapters. Starting with some properties of the ensemble, we then describe its construction and end with an evaluation of the forecasting performance.

2.1 Properties and Construction

COSMO-DE-EPS is a 20-member ensemble prediction system, developed by the German Meteorological Service, with a planned extension to 40 members. Its pre-operational phase started on 9 December 2010 and the operational phase was launched on 22 May 2012 (Theis and Gebhardt, 2012). The forecasts are made for lead times from 0 up to 21 hours on a grid covering Germany. The horizontal and vertical spacing between grid-points is 2.8 km, resulting in meso-γ-scale predictions (Peralta and Buchhold, 2011; Theis et al., 2011).

The EPS is based on a convection-permitting configuration, COSMO-DE, of the numerical weather prediction model COSMO (Steppeler et al., 2003; Baldauf et al., 2011). Usually, prediction ensembles are created by perturbing the initial conditions and model physics of the corresponding numerical forecast model, based on an idea by Leith (1974). In case of the COSMO-DE-EPS, variations in lateral boundary conditions are also included, in order to generate multiple deterministic predictions for one location (Gebhardt et al., 2011). In particular, five different configurations of the COSMO-DE model yield perturbations in the model physics. On the other hand, the ensemble member forecasts rely on diverse lateral

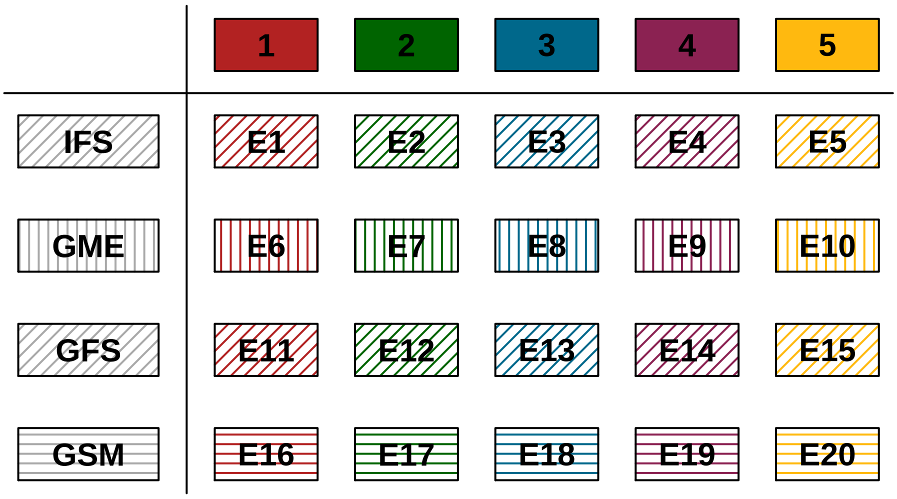

Figure 2.1: Construction of COSMO-DE-EPS: IFS, GME, GFS, GSM represent the four global models, which provide the initial and lateral boundary conditions, and 1-5 correspond to the different configurations in the model physics of COSMO-DE. By running the four sets of restricting conditions through the five distinct model formulations, 20 ensemble members E1-E20 are generated.

boundary and initial conditions, provided by four dissimilar global models (Peralta et al., 2012): Integrated Forecast System (IFS), run by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Janssen and Bidlot, 2002), DWD's global model (GME) (Majewski et al., 2002), Global Forecast System (GFS) (Environmental Modeling Center, 2003), operated by the American National Center for Environmental Prediction (NCEP) and Global Spectral Model (GSM), generated by the Meteorological Agency of Japan (Zhan et al., 1995). When plugging these four distinct sets of conditions in the five configurations of COSMO-DE, 20 members are generated. This is visualized in Figure 2.1.

2.2 Forecasting Performance

We consider 21-hour ahead forecasts of surface temperature, initialized at 00:00 UTC, within the time frame from 10 December 2010 until 30 November 2011. Using a training period of 25 days for our postprocessing methods in Chapters 3 and 4, we start making forecasts on 5 January 2011 and thus evaluate the raw ensemble over the same period of time. If at least one member is missing at every location on a specific day, we omit this day completely,

Table 2.1: Scores of the raw ensemble, as well as its average width and coverage of the nominal 90.5% prediction interval, aggregated over 5 January 2011 until 30 November 2011.

CRPS	MAE	RMSE	Width	Coverage
1.77°C	1.57°C	2.27°C	1.50°C	26.97%

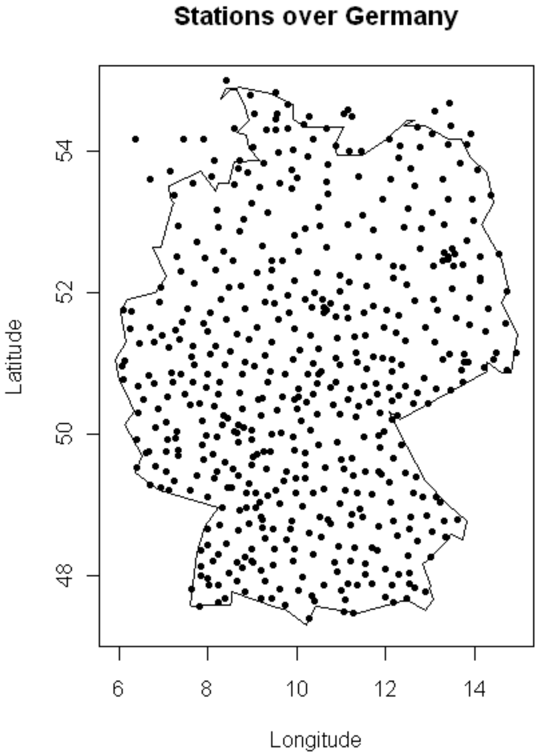

Figure 2.2: Map of observational locations in Germany

pretending it never occurred. By following this approach, ten days are eliminated with 346 days remaining.

Spread over Germany, there are 515 SYNOP stations, as seen in Figure 2.2. However, the forecasts' grid points do usually not match the observation locations and so the ensemble output is bilinearly interpolated, in order to produce predictions for the observation sites. When applying this procedure, forecasts and observations are provided for the 515 locations. However, bilinear interpolation does not account for variation of temperature due to changes in altitude. Especially, if the surrounding grid points are situated substantially beneath the observational site, the corresponding prediction shows a significant negative bias. This occurs at Germany's highest mountain Zugspitze at 2690 m above sea level and consequently we choose to eliminate this station, as the corresponding forecasts prove to be unreliable

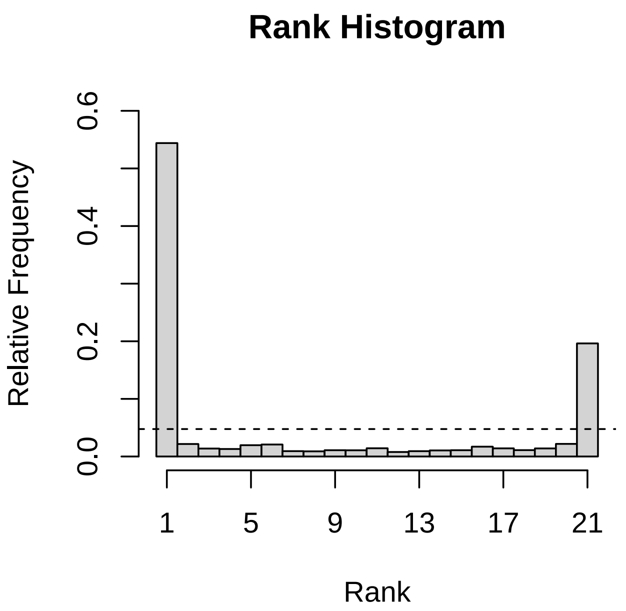

Figure 2.3: Rank histogram of the raw ensemble aggregated over 514 observation sites and the time period from 5 January 2011 until 30 November 2011.

and distort the overall performance results. In total, we evaluate forecasts for 117,879 verifying observations at 514 sites over 346 days.

As has been discussed in the literature, ensemble forecasts tend to be underdispersive and often show a positive spread-error correlation (see e.g. Hamill and Colucci, 1997; Buizza, 1997; Eckel and Clifford, 2005). Calculating the latter for the COSMO-DE-EPS via the absolute error and the ensemble range, the outcome equals 0.11, which motivates the application of the postprocessing methods EMOS and BMA in Chapter 3, as both of them incorporate the ensemble variance information.

When assessing the performance, we apply a selection of the methods presented in Appendix A. The rank histogram in Figure 2.3 clearly shows underdispersion of the ensemble, as the majority of the observations fall beneath the ensemble's predicted minimum temperature or above its maximum which leads to the conclusion that the ensemble width is too narrow. By looking at Table 2.1, this idea is supported, as the average coverage of the nominal 90.5% prediction interval yields only 26.97%. Consequently, its average width of 1.50°C is highly underestimated and the spread needs to be larger in order to generate calibrated forecasts.

A contributing factor to the small spread could be the fact that the forecast errors of the distinct ensemble members are highly correlated, which can be seen in Figure 2.4, showing the monthly and overall correlation coefficients between the members. The correlation coefficients of the error terms range from 0.82 up to 1, implying a great dependency between the errors. In these image plots, there are some noticeable patterns, which can be traced back to the construction of COSMO-DE-EPS. Obviously, the members E1-E5, E6-E-10, E11-E15 and E16-E20 always show a particular high correlation, due to the fact that they are based on initial and lateral boundary conditions provided by the same global model. In addition, within every block of five succeeding members, there is a similar pattern of coefficients visible, which can be related to the same configurations of the model physics. Seasonal changes have a small impact on the correlation coefficient, as the correlation slightly reduces during the summer months, but increases again in autumn.

While evaluating the forecasting performance, another useful tool to compare competing forecasting methods are the scores in Table 2.1. For the mean absolute error (MAE) and root mean square error (RMSE), the COSMO-DE-EPS yields good results, which indicates that the ensemble's predictive median and mean are close to the verifying observations. However, as mentioned before, the overall spread of the ensemble is too small, which leaves room for improvement, especially in the continuous rank probability score (CRPS). We will investigate this in the following two chapters.

This chapter has introduced the basics of the COSMO-DE-EPS by describing its characteristics and modulation. In its current state, the predictions are very accurate, but the ensemble spread is too small and its members are highly correlated. In consequence, the need for postprocessing arises, in order to address these deficiencies.

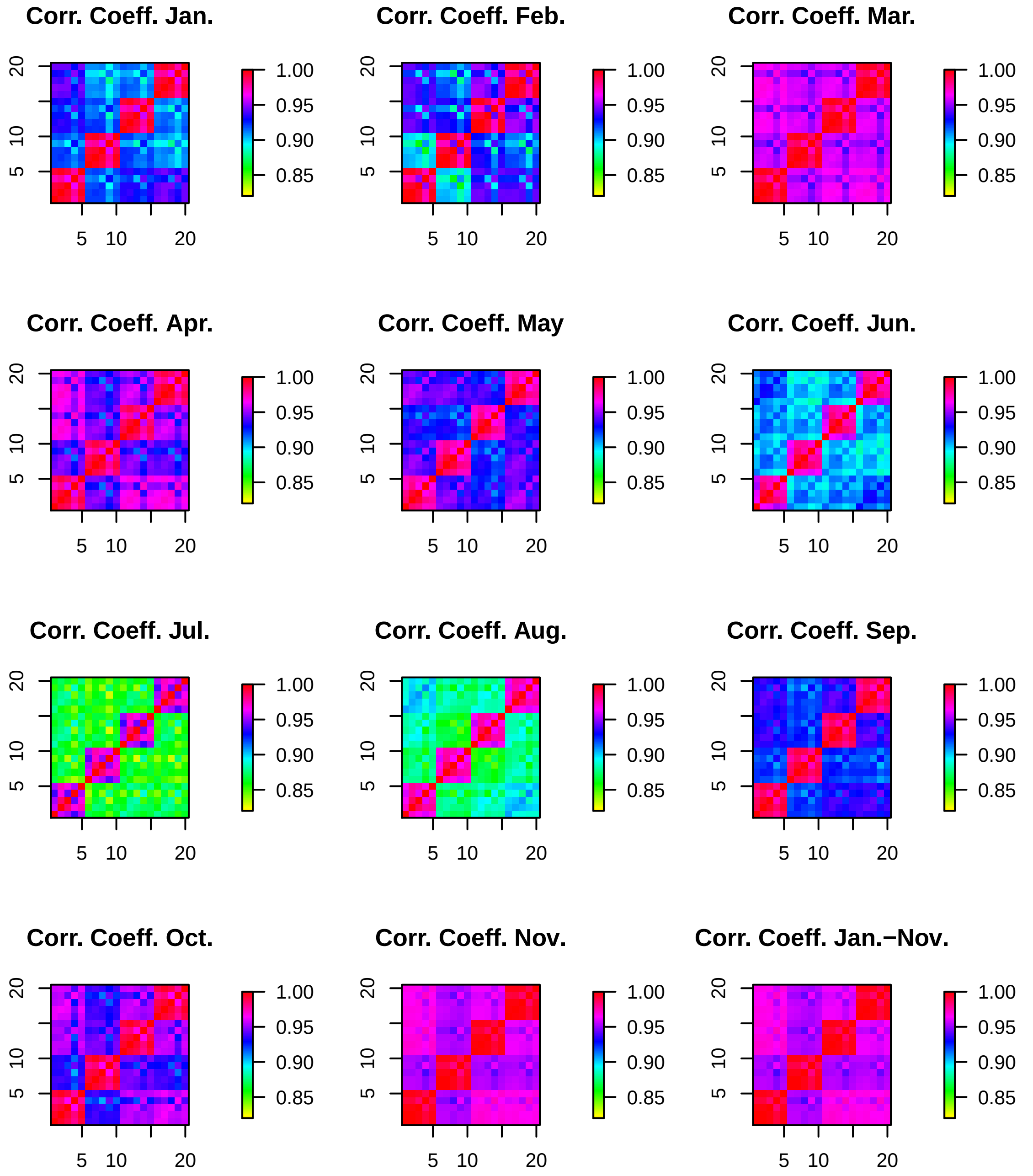

Figure 2.4: Image plots of monthly correlation coefficients between the members of the ensemble, where the horizontal and vertical axis represent the members of the COSMO-DE-EPS. The plot in the last panel covers the entire time period.

Chapter 3

Univariate Postprocessing

Most forecast ensembles, including the COSMO-DE-EPS (introduced in Chapter 2), show a positive spread-error correlation, while at the same time being uncalibrated. In order to address theses issues, a variety of statistical postprocessing techniques have been developed. In this chapter, we present different methods for univariate postprocessing of ensemble forecasts. All the procedures have in common that they yield full predictive probability distributions, which strive to satisfy the underlying goal of "maximizing sharpness subject to calibration" (Gneiting et al., 2007). In this thesis, we focus on techniques for the postprocessing of temperature.

At first, we briefly demonstrate the principles of the well known techniques BMA and EMOS, followed by a method which is solely based on least squares regression. We name this method linear model forecasts (LMF). Subsequently, we explain procedures to obtain reference forecasts in order to compare the overall predictive performance. After discussing the application of all techniques to the COSMO-DE-EPS, we finalize with an EMOS approach, where we replace the commonly used normal distribution with a Student's t-distribution, in an attempt to resolve some of the issues that arise in the statistical postprocessing of COSMO-DE-EPS.

3.1 Bayesian Model Averaging (BMA)

BMA is a standard statistical approach for combining competing statistical models and has a broad application in e.g. social and health sciences (Hoeting et al., 1999). Its advantage over other techniques, such as conventional regression analysis, is based on the fact that

BMA makes use of multiple models, in contrast to methods which soley use a single model that is deemed to be the best. Only using a single model often leads to an underestimation of the uncertainty in the process of model selection. In this thesis, we follow the extension of BMA from statistical models to dynamical models by Raftery et al. (2005), for the purpose of producing calibrated and sharp predictive distributions.

Let y_s denote the weather variable of interest at location $s \in \mathcal{S}$, and $f_{1s}, ..., f_{Ms}$ the corresponding forecasts of the M-member ensemble. In the framework of BMA, we assign each member a conditional probability density function $p_m\left(y_s | f_{ms}\right)$, or kernel density, which we can interpret as the conditional density of y_s, given member $m \in M$ being the most skillful within the ensemble. Then, the predictive density for y_s equals

$$p\left(y_s | f_{1s}, ..., f_{Ms}\right) = \sum_{m=1}^{M} w_m p_m\left(y_s | f_{ms}\right),$$

where $w_1, ..., w_M$ are non-negative weights that add up to 1. The weights are determined by the member's skill in the training period; a higher weight reflects a more reliable member, whereas a lower weight is associated with a weak performance.

In the setting of BMA, temperature is modeled with a normal distribution. Thus, the kernels are univariate normal densities, centered at each member's bias-corrected forecast $a_{ms} + b_{ms} f_{ms}$,

$$y_s | f_{ms} \sim \mathcal{N}(a_{ms} + b_{ms} f_{ms}, \sigma^2)$$

with a common variance σ^2.

If it is necessary to produce deterministic forecasts, the BMA predictive mean, which is a weighted average over the bias-corrected forecast,

$$\mathrm{E}\left(y | f_1, ..., f_M\right) = \sum_{m=1}^{M} w_m\left(a_m + b_m f_m\right)$$

can be useful. The overall variance of y_s in the BMA setting is

$$\mathrm{Var}(y_s | f_{1s}, .., f_{ms}) = \sum_{m=1}^{M} w_m \left(\left(a_m + b_m f_{ms}\right) - \sum_{m=1}^{M} w_m\left(a_m + b_m f_{ms}\right) \right)^2 + \sigma^2,$$

where the first part of the sum on the right-hand side is the between-forecast variance while the second part represents the within-forecast variance.

Given a set of verifying observations y_{st} and associated ensemble forecasts $f_{1st}, ..., f_{Mst}$ for day t within a training period $\mathcal{T}$, the BMA parameters are obtained in three steps. First, the member-specific parameters a_m and b_m, for $m = 1, ..., M$, are individually determined via simple linear regression. In the second step, the weights w_m, $m = 1, ..., M$, and the variance σ^2 are estimated by maximizing the log-likelihood function

$$l(w_1, ..., w_M, \sigma^2) = \sum_{s,t} \log \left(\sum_{m=1}^{M} w_m p\left(y_t | f_{kst}\right) \right),$$

which, for simplicity, is based on the assumption that the forecast errors are independent in space and time. However, the maximum cannot be found analytically and Raftery et al. (2005) thus employ the expectation–maximization (EM) algorithm. In the final and voluntary step, the parameter estimate for σ^2 may be refined by minimizing the CRPS, a proper scoring rule described in Appendix A, over the training period. When implementing the BMA approach, we utilize the `R` package `ensembleBMA` by Fraley et al. (2011), which yields the desired predictive distributions.

Apart from generating forecasts for normally distributed variables, such as temperature or sea level pressure, BMA approaches for other weather quantities have been studied. Sloughter et al. (2007) propose a gamma distribution with a point mass in zero for modeling precipitation, and Sloughter et al. (2010) present a BMA method to predict wind speed using a gamma distribution. In addition, Bao et al. (2010) describe future wind directions via von Mises distributions.

3.2 Ensemble Model Output Statistics (EMOS)

EMOS, also called non-homogeneous Gaussian regression, is a form of multiple linear regression and an extension to the model output statistics technique. In contrast to BMA, EMOS is based on a single predictive density, whose parameters depend on the ensemble forecasts.

For temperature, a normal distribution is again employed to describe the future state of the variable (Gneiting et al., 2005). At location $s \in \mathcal{S}$, the predictive distribution is

$$y_s | f_{1s}, ..., f_{Ms} \sim \mathcal{N}(a_1 + b_1 f_{1s} + ... + b_M f_{Ms}, c + d S_s^2),$$

which forms a linear model with the forecasts as predictors and the temperature y_s as predictand. The variance is modeled as a linear function of the ensemble variance S_s^2. In this approach, the variance term counteracts the underdispersion of the ensemble, while additionally accounting for the positive spread-error correlation by incorporating the ensemble information.

In contrast to BMA, the coefficients $b_1, ..., b_M$ can take any value in $\mathbb{R}$. However, negative values are difficult to interpret. Hence, the authors suggest restricting the coefficients to be non-negative and name this approach EMOS$^+$. In this extended framework, the coefficients reflect the relative performance of the ensemble members over the training period. We further investigate a simpler variant, EMOS mean, where the ensemble mean is used as a predictor rather than the individual members.

For the calculation of the parameters, Gneiting et al. (2005) use minimum CRPS and maximum likelihood estimation. According to their results, the minimum CRPS estimation outperforms the latter and hence we employ this method as well. Over a set of training data which contains the past forecasts and corresponding observations, the parameters optimizing the score are chosen. In the case of a normal distribution, the CRPS can be written in a closed form (Gneiting et al., 2005) and the computational cost is therefore substantially reduced.

Further developments of EMOS for wind speed, wind gust and wind vectors can be found in Thorarinsdottir and Gneiting (2010), Thorarinsdottir and Johnson (2012) and Schuhen et al. (2012), respectively.

3.3 Linear Model Forecast (LMF)

Originally, we developed the LMF in the context of modeling EMOS with a Student's t-distribution. In this framework, the technique performs very well and we hence include it here in a more general setting. LMF is based on two subsequent applications of linear least squares regression. In the first step, we calculate the bias correction parameters a and b for the ensemble mean $\bar{f}$ over a training period. We could include all members in the model formulation, however, due to the high correlation of the forecast errors within COSMO-DE-EPS, there is not much lost when following this simpler approach. Then, we use linear regression equations with the squared values of the residuals as predictands and the local ensemble variances as predictors, in order to obtain the variance parameters c and d. Finally, the predictive density results in

$$y_s|f_{1s}, ..., f_{Ms} \sim \mathcal{N}(a + b\bar{f}, c + dS^2),$$

where S represents the standard deviation of the ensemble.

3.4 Reference Forecasts

In order to compare the postprocessing methods discussed above with standard approaches, we also include some simpler techniques. Error dressing and the bias-corrected ensemble are both based on ensemble predictions systems, whereas the climatology ensemble is generated via past observations.

Error Dressing

Error dressing, proposed by Gneiting et al. (2008), is an effortless way of postprocessing ensemble outputs, which accounts for bias and dispersion errors. We employ a global variant of the method, where at each station $s \in \mathcal{S}$ we calculate the empirical errors $e_s = y_s - \bar{f}_s$ of the ensemble mean $\bar{f}_s$ over the training period. Let $\bar{e}$ denote the mean of the sample of empirical errors over all stations and σ^2 the empirical variance, then the predictive density equals

$$y_s|f_{1s}, ..., f_{Ms} \sim \mathcal{N}(\bar{f}_s + \bar{e}, \sigma^2).$$

Bias-Corrected Ensemble

The bias-corrected ensemble is a slight variation of the original ensemble, which often outperforms the former. Considering each member individually, we use linear least squares regression, with the forecasts f_{mst} as predictors and the verifying observations y_{st} as predictands for day t in the training period $\mathcal{T}$. When following this approach, we obtain the member-specific bias-correcting parameters $a_{m,t+1}, b_{m,t+1}$ for the succeeding day $t+1$. Then, the values $a_{m,t+1} + b_{m,t+1}f_{m,s,t+1}$, $m = 1, ..., M$, form the bias-corrected ensemble.

Climatology

In contrast to the aforementioned methods, the climatology ensemble is not based on an EPS, instead past observations from the training period for the statistical postprocessing methods are used to produce forecasts. When fitting a normal distribution to the observations

Table 3.1: Univariate scores as well as width and coverage of the nominal 90.5% prediction interval. The results cover the time period from 5 January 2011 until 30 November 2011 and are averaged over all observation sites and days.

				Prediction Intervals	
	CRPS (°C)	MAE (°C)	RMSE (°C)	Width (°C)	Coverage (%)
Raw Ensemble	1.77	1.57	2.27	1.50	26.97
Bias-Corrected Ensemble	1.30	1.48	1.89	1.20	28.51
BMA	1.04	1.46	1.86	5.91	88.82
EMOS	1.02	1.43	1.83	5.61	88.70
EMOS$^+$	1.04	1.46	1.87	5.76	87.99
EMOS Mean	1.05	1.48	1.89	5.89	88.20
Error Dressing	1.07	1.50	1.91	4.73	80.12
LMF	1.07	1.48	1.89	6.13	89.53
Climatology	2.25	3.16	4.00	11.79	85.25

contained in the training set, we use the empirical mean and variance as the distribution parameters. Unlike the error dressing ensemble, this procedure is performed locally at every single station $s \in \mathcal{S}$. In this way, for every observation site $s \in \mathcal{S}$, a different distribution function is obtained.

3.5 Results

All presented methods - except for the climatology ensemble - share the mutual characteristic that they use the structural patterns of past forecast errors, with the purpose of improving the current prediction. Hence, past data, containing the ensemble forecasts as well as the verifying observations, is employed for the estimation of the model parameters. In this thesis, we use the sliding window approach, which means, when estimating the model parameters for a specific day, the data of a certain number of consecutive, previous days is utilized in this process.

For the determination of the length of the training period, there is a trade-off (Gneiting et al., 2005). A short span has the ability to adjust more quickly to seasonal variation, whereas parameter estimation based on a longer time frame is more stable and less prone to variability. However, as our given data hardly covers one year, we do not have the opportunity for out-of-sample comparison of different lengths for the training period and consequently use 25 days, as proposed by Berrocal et al. (2007).

When implementing the introduced procedures, we use the output of COSMO-DE-EPS, presented in Chapter 2. Forecast of 21h-ahead surface temperature are available from 10 December 2010 until 30 November 2011. With a 25-day training period, we begin forecasting on 5 January 2011.

For the assessment of the forecasting performance, we evaluate the predictive densities of the competing models with the techniques presented in Appendix A. In the case of the raw and the bias-corrected ensemble, we replace the output with a normal distribution, whose parameters are the empirical mean and variance of the respective ensemble.

Figure 3.1 shows the rank histograms, which are based on samples of the ensemble size 20 for all postprocessing models. The raw as well as the bias-corrected ensemble show a U-shape and thus underestimate the forecasting uncertainty. The remaining postprocessing techniques correct this flaw, demonstrated by nearly uniformly appearing histograms. The scores in Table 3.1 confirm the improvement achieved by postprocessing. Considering the CRPS, EMOS performs best, closely followed by BMA and EMOS$^+$. Due to the high correlation of the forecast errors between the members of the COSMO-DE-EPS (see Chapter 2), the performance of EMOS mean is very good in comparison to the more sophisticated models. The subsequent methods, the LMF and the error dressing ensemble, although being simple, yield very good results. When viewing the scores, the climatology performs worst, as it is not based on predictions, but only past observations.

For the MAE and the RMSE, all postprocessed models yield comparable results, which is to be expected as the respective predictive means are based on estimations via least squares regression. In terms of deterministic forecasts, the raw and climatology ensemble produce poor results reflected in a rather high MAE and RMSE.

Again, Table 3.1 confirms the underestimated spread of the raw and bias-corrected ensemble, not even covering 30% for a nominal $19/21 \approx 90.5\%$ prediction interval. Most of the other techniques fulfill the requirement nearly or entirely, while still creating much sharper forecasts than climatology. The average width of the prediction interval by the climatology ensemble is rather large, since the variation in temperature observed in the previous 25 days is prone to high variability.

3.6 Extension: EMOS with Student's t-distribution

As the pre-operational phase of COSMO-DE-EPS only started in December 2010, at the beginning of working on this thesis soley small data sets were available. The first one

Table 3.2: Scores in degrees Celsius for different EMOS based forecasts, averaged over the time from 5 January 2011 until 30 March 2011 and all observational sites. EMOS normal is based on a normal distribution. EMOS Student's t employs a Student's distribution with different degrees of freedom ν, but uses the parameters a, b, c, d, estimated by EMOS normal.

Model	CRPS	MAE	RMSE
EMOS normal	1.06	1.48	1.90
EMOS Student's t; $\nu = 3$	1.09	1.48	1.90
EMOS Student's t; $\nu = 5$	1.07	1.48	1.90
EMOS Student's t; $\nu = 100$	1.06	1.48	1.90

comprised observations and forecasts for the time frame from 9 December 2011 until 31 March 2011.

After applying EMOS to the data of this period, the forecasts were not calibrated, as can be seen in Figure 3.2. The middle section of the probability integral transform (PIT) histogram (see Appendix A) appears almost uniform. However, many observations obtain very low and high PIT values, which might mean that the probability mass on the edges of the density is too small, suggesting that the normal distribution is not the best fit to describe the predictive distribution of temperature. This particular pattern with uniformity in the middle, but outliers at the sides might be attributed to the fact that the tails of the normal distribution are too small. In order to address this phenomenon, we study the use of the Student's t-distribution, developed by Gosset (1908), which is an extension of the normal distribution with heavier tails. For simplicity, we fit the Student's t-distribution within an EMOS mean setting.

The density function of the Student's t-distribution equals

$$p(y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} (\nu\pi)^{-\frac{1}{2}} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{3.1}$$

where ν denotes the degrees of freedom and Γ the Gamma function. The parameter ν determines the shape of the tails, as can be seen in Figure 3.3. For $\nu \to \infty$, the Student's t density function converges to the normal density function.

We expand the density in Equation 3.1 by introducing a scale parameter λ and a location parameter μ (see e.g. Bishop (2006)). Then, the density is

$$p(y|\mu,\lambda,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{\lambda}{\nu\pi}\right)^{\frac{1}{2}}\left(1+\frac{\lambda\left(y-\mu\right)^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

with $\mathrm{E}\left(\mathrm{y}\right) = \mu$, for $\nu > 1$, and $\mathrm{Var}(y) = \frac{1}{\lambda}\frac{\nu}{\nu-2}$, for $\nu > 2$.

When estimating parameters, we use the maximum likelihood technique. Since we want to obtain values in terms of linear functions of the ensemble parameters, we substitute the distribution variance by a linear function of the ensemble variance, $\frac{1}{\lambda}\frac{\nu}{\nu-2} = c + dS^2$ (for $\nu > 2$), and perform the same for the mean, $\mu = a + b\bar{f}$. Here, S^2 denotes the ensemble variance, $\bar{f}$ the forecast mean and a, b, c and d are real numbers:

$$p(y|a,b,c,d,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(\left(c+dS^2\right)\left(\nu-2\right)\pi\right)^{-\frac{1}{2}}\left(1+\frac{\left(y-a+b\bar{f}\right)^2}{\left(c+dS^2\right)\left(\nu-2\right)}\right)^{-\frac{\nu+1}{2}}.$$

Then, the likelihood function, given the past observations $y_1, ..., y_n$, equals

$$L(a,b,c,d,\nu|y_1,...,y_n) = \prod_{i=1}^{n} p(y_i|a,b,c,d,\nu).$$

When maximizing this expression, we find the values for the parameters a, b, c, d and ν, which were most likely to have produced the observations $y_1, ...y_n$. For algebraic simplicity and numerical stability, instead of maximizing the likelihood function, we choose to maximize the log-likelihood, which yields the same results. We use the `R` function `optim` with box constraints, based on the algorithm by Bryd et al. (1995), in order to find the maximum of

$$\begin{aligned} l(a,b,c,d,\nu|y_1,...,y_n) &= n\log\left(\Gamma\left(\frac{\nu+1}{2}\right)\right) - n\log\left(\Gamma\left(\frac{\nu}{2}\right)\right) \\ &- \frac{n}{2}\log(\pi\nu) - \frac{1}{2}\sum_{i=1}^{n}\log\left(\left(c+dS_i^2\right)\frac{\nu-2}{2}\right) \\ &- \frac{\nu+1}{2}\sum_{i=1}^{n}\log\left(1+\frac{\left(y_i-\left(a+b\bar{f}_i\right)\right)^2}{\left(\left(c+dS_i^2\right)\frac{\nu-2}{2}\right)\nu}\right), \quad \text{for } \nu > 2. \end{aligned} \tag{3.2}$$

When estimating the parameters, we restrict ν to be greater than two and leave the parameter space of the other variables as large as possible, by setting the lower boundary to $-1,000$ and the upper to $1,000$. However, the parameter estimation turns out to be unstable, as c and d usually take unrealistic values over 20 and the estimates for ν are always as close to two as possible.

In order to stabilize the parameter estimation, we try another approach, which might seem suboptimal. However, if the small tails of the normal density function are the issue, this method nevertheless should detect it. We estimate the parameters a, b, c and d in the regular EMOS setting, by minimizing the CRPS for the normal density function. Then, we plug the results for a, b, c and d into Equation 3.2, so that it depends only on ν. In the final step, we estimate ν separately via maximum likelihood estimation of this equation. For this step, we again use the `R` function `optim` with box constraints, in order to regulate the parameter space for ν. If small tails caused the issue, ν ought to take small values. However, we find that ν always takes values as close to the upper boundary of the constraint as possible. For these large values of ν, the normal and Student's t density function coincide empirically.

In Table 3.2, we follow the aforementioned approach and then force ν to take the values of three, five, and hundred. The results show that there is no improvement when modeling with heavier tails. Instead, the CRPS, which addresses calibration and sharpness simultaneously, is higher for small values of ν and then decreases as ν is assigned greater values.

Hence, the normal density function captures the predictive distribution of temperature better and other aspects must cause the structure of the PIT histogram in Figure 3.2. It might be attributed to the small amount of data, or due to a spatial pattern, which can not be captured by the global parameters of EMOS. However, after receiving larger data sets, we found that the population of the outer bins in the PIT histogram decreased noticeably and we stopped investigating this problem further.

In this chapter, we have presented different approaches for postprocessing of temperature. Most of them yield calibrated and sharp forecasts, by utilizing the structural patterns of past errors for the predictions. However, these procedures do not account for spatial correlation between these errors. This issue we will discuss in the succeeding chapter.

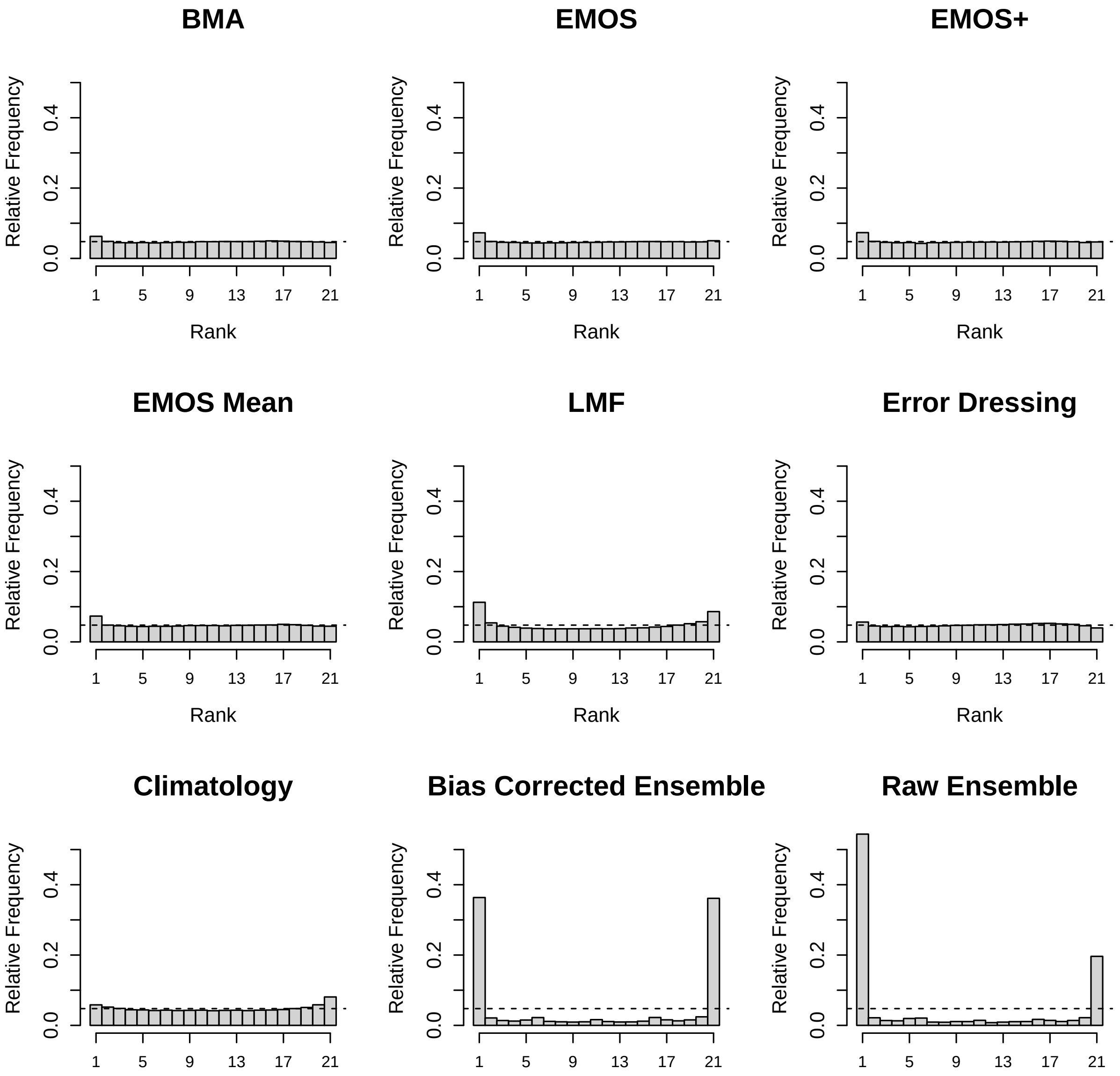

Figure 3.1: Rank histograms of the competing forecasts, aggregated over all stations from 5 January 2011 until 30 November 2011.

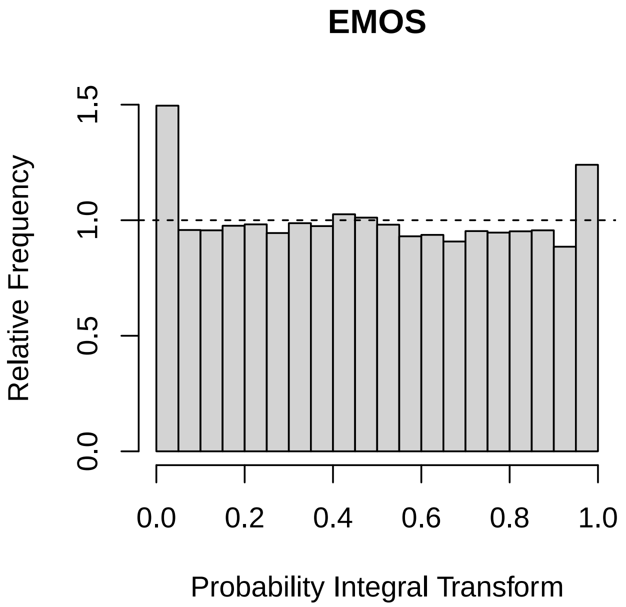

Figure 3.2: PIT histogram for EMOS with values aggregated over 5 January 2011 until 30 March 2011 and all observational sites.

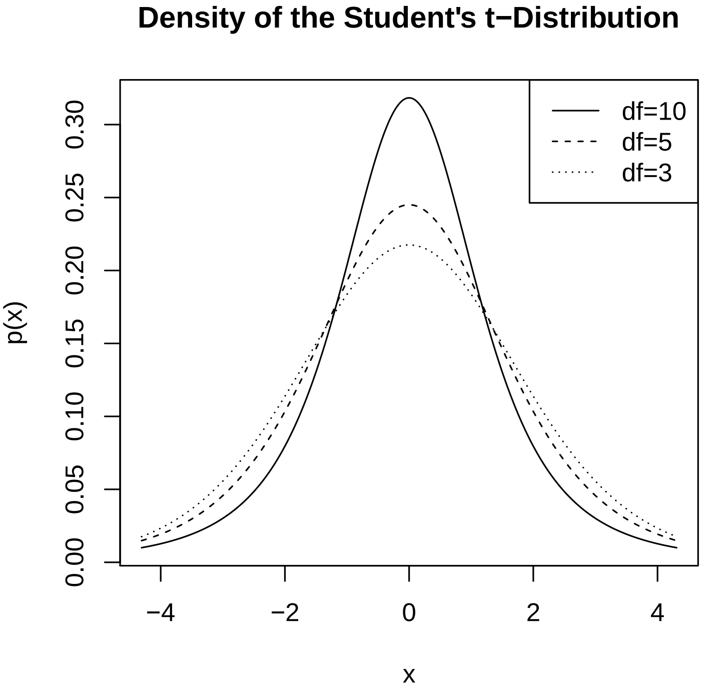

Figure 3.3: Density of Student's t-distribution with varying degrees of freedom (df) ν, while all other parameter are kept equal. The distribution is centered at zero with a variance of four.

Chapter 4

Spatial Postprocessing

This chapter focuses on different models for spatial postprocessing of ensemble temperature forecasts. We start with a summary of the GOP method for point predictions. By modeling the spatial correlation structure, this technique produces sharp and calibrated forecasts for entire weather fields. When combining this procedure with EMOS or BMA, introduced in Chapter 3, we additionally include the information of the ensemble, while still incorporating the spatial structure of the weather field. Subsequently, we present two multivariate reference forecasts, ensemble copula coupling (ECC) and the noise ensemble, ending with a discussion of the application of all techniques to COSMO-DE-EPS.

4.1 Geostatistical Output Perturbation (GOP)

Originally not developed for ensemble outputs, GOP produces sharp and calibrated forecasts, based on one deterministic prediction for weather fields. The technique consists of dressing the multidimensional output of the numerical forecast systems with simulated error fields described by a spatial random process. Thus, GOP perturbates the outputs of numerical weather prediction models, instead of its inputs.

Gel et al. (2004) chose to employ a parametric, stationary, and isotropic geostatistical model, in order to capture the spatial structure of the error fields. The error is defined as the difference between the observation and the bias-corrected forecast. Suppose $\mathcal{S}$ denotes all locations for which forecasts are available. Then, let $\mathbf{Y} = \{y_s : s \in \mathcal{S}\}$ be the vector that describes the weather variable of interest at site $s \in \mathcal{S}$. Further, $\mathbf{F}_m = \{f_{ms} : s \in \mathcal{S}\}$ refers to the weather field forecast by the member m of the ensemble with size M. Note that this procedure is performed for only one member, although we include the subscript for future incorporation of ensemble information.

GOP is based on a statistical model, stating that

$$\mathbf{Y}|\mathbf{F}_m \sim MVN\left(a_m\mathbf{1} + b_m\mathbf{F}_m, \boldsymbol{\Sigma}_m\right),$$

where $\mathbf{1}$ is a vector of length $\#\mathcal{S}$ with all entries equal to 1. Given the forecasts $\mathbf{F}_m$, $\mathbf{Y}$ is multivariate normally distributed with mean equal to the bias corrected forecast, $a_m\mathbf{1} + b_m\mathbf{F}_m$ and the covariance matrix $\boldsymbol{\Sigma}_m$. The entries of $\boldsymbol{\Sigma}_m$ depend on the covariance structure of the error fields. Let $C\left(s_1, s_2\right)$ be a stationary and isotropic correlation function, then the entry (i, j) of $\boldsymbol{\Sigma}_m$ equals

$$\rho_m^2\delta_{ij} + \tau_m^2\, C\left(s_i, s_j\right),$$

where δ_{ij} denotes to the Kronecker delta function. The nugget effect $\rho_m^2 \geq 0$ has two interpretations. On one hand, it can be thought of as the variance of the measurement error. On the other hand, it is a measure of the spatial variation within a distance smaller than the smallest distance between two different sites s_i and s_j, for $i \neq j$. The sum $\rho_m^2 + \tau_m^2$ is called the sill.

There are various ways to model the spatial structure of the weather field with different covariance classes. Gel et al. (2004) suggest the use of the exponential correlation function,

$$C(s_i, s_j) = e^{-\frac{||s_i - s_j||}{r_m}},$$

where $|| \cdot ||$ denotes the Euclidean norm and the range $r_m > 0$ is a parameter in the unit of the distance and determines the rate at which the spatial correlation decays. We also propose a more general approach, where we apply the Matérn correlation function (Matérn, 1986)

$$C(s_i, s_j) = \frac{1}{2^{1-\nu_m}\Gamma\left(\nu_m\right)} \cdot \left(\frac{||s_i - s_j||}{r_m}\right)^{\nu_m} \cdot \mathrm{K}_{\nu_m}\left(\frac{||s_j - s_i||}{r_m}\right).$$

Here, $\Gamma(\cdot)$ denotes the gamma distribution and $\mathrm{K}_\nu(\cdot)$ the modified Bessel function of order $\nu > 0$. The parameter ν regulates the smoothness of the simulated error field. For $\nu = \frac{1}{2}$, the Matérn correlation function coincides with the exponential model above.

The value of $\mathbf{Y}|\mathbf{F}_m$ can also be calculated in terms of realizations of the error field which may be decomposed into two parts $\mathbf{E}_{1m} = \{\varepsilon_{1m}(s) : s \in S\}$ and $\mathbf{E}_{2m} = \{\varepsilon_{2m}(s) : s \in S\}$ (Berrocal et al., 2007):

$$\mathbf{Y}|\mathbf{F}_m = a_m\mathbf{1} + b_m\mathbf{F}_m + \mathbf{E}_{1m} + \mathbf{E}_{2m}.$$

The vectors $\mathbf{E}_{1m}$ and $\mathbf{E}_{2m}$ have a multivariate distribution with mean zero and a covariance structure based on

$$\text{cov}[\varepsilon_{1m}(s_i), \varepsilon_{1m}(s_j)] = \tau_k^2\, C\,(s_i, s_j)$$

and

$$\text{cov}[\varepsilon_{2m}(s_i), \varepsilon_{2m}(s_j)] = \rho_{mk}^2 \delta_{ij},$$

respectively. The term $\mathbf{E}_{1m}$ is referred to as the continuous component of the error field which varies in space, whereas $\mathbf{E}_{2m}$ describes the discontinuous part, as it models a random noise in order to correct measurement errors.

By using the GOP method, an ensemble of any desired size can be obtained. A new member is produced by dressing the bias corrected forecast $a_m\mathbf{1}+b_m\mathbf{F}_m$ with a simulation of the error fields $\mathbf{E}_{1m}$ and $\mathbf{E}_{2m}$. For this simulation we employ the `R` package `RandomFields` by Schlather (2011).

When estimating the parameters, we use a set of training data which contains past forecasts and realized observations. There are several ways to estimate the parameters of the geostatistical model. Gel et al. (2004) mention a fully Bayesian approach, maximum likelihood, and a variogram-based estimation. In order to reduce computational time, the authors chose the last mentioned. Here, we also include a maximum likelihood approach. Independently of the estimation technique for the geostatistical model, the coefficients a_m and b_m are estimated via linear least squares regression over the sliding training period.

In geostatistics, a variogram is a tool which describes the spatial correlation of a stochastic process. Theoretically, it is defined as

$$\gamma_m\,(s_i, s_j) = \frac{1}{2}\text{Var}\,(X(s_i) - X(s_j))\,,$$

where $X(s)$ denotes the value of the stochastic process at location s. Since the underlying model of the GOP method is stationary, the variogram reduces to a function that depends on the distance $d = ||s_i - s_j||$ only. Additionally, the mean and the variance of the error fields are defined as spatially constant so that the theoretical variogram values $\gamma(d)$ of the geostatistical model equal

$$\gamma_m(d) = \rho_m^2 \delta_{ij} + \tau_m^2 \left(1 - C\left(s_i, s_j\right)\right),$$

see e.g. Diggle and Ribeiro Jr. (2007).

For the estimation of the parameters, we calculate an empirical version of the variogram following the approach by Berrocal et al. (2007). After determining a_m and b_m, we compute the errors, which equal the residuals of the linear least squares regression fit. For each day in the training period, we then determine the distances between every possible pair of locations. Additionally, we calculate one-half the squared difference between the corresponding pair of errors,

$$\frac{1}{2}\left(e_{s_i} - e_{s_j}\right)^2,$$

where e_{s_i} and e_{s_j} denote the site-specific errors on one day in the training period. In the following step, the collection of distances are sorted into bins B_l with centers at d_l. The cut points of the bins are chosen by the rule that on average during the entire forecasting period, the same amount of distances should fall in every bin. During the entire forecasting period, the cut points and centers stay constant, as implemented in the `R` package `ProbForecastGOP` by Berrocal et al. (2010). Finally, the empirical variogram value $\hat{\gamma}_m(d_l)$ at distance d_l equals the average of one-half the squared difference of the errors whose distances fall into bin B_l.

When fitting a curve to the empirical variogram values, Berrocal et al. (2007) employ weighted least squares as proposed by Cressie (1985). Let θ_m denote the parameter vector which depends on ρ_m^2, τ_m^2 and r_m, in the case of an exponential correlation function, and additionally on ν for a Matérn model. In order to obtain the optimal value of θ_m, the function

$$S(\theta_m) = \sum_l n_l \frac{\hat{\gamma}_m(x_l) - \gamma(\theta_m, d_l)}{\gamma(\theta_m, d_l)}$$

is minimized. Here, n_l refers to the number of pairs contained in bin B_l. When minimizing the expression, we employ the `R` function `optim` with boundary conditions based on the algorithm by Bryd et al. (1995).

Beside the variogram-based estimation, we also apply a maximum likelihood approach. Since the errors are assumed to be a realization of a Gaussian random field, the likelihood function is

$$L(\theta_m, \mathbf{e}_m) = \frac{1}{(2\pi)^{n/2} \left| \sum(\theta_m) \right|^{1/2}} e^{-\frac{1}{2}\mathbf{e}_m \sum(\theta_m)^{-1} \mathbf{e}_m^t},$$

where $\mathbf{e}_m$ denotes a vector containing the errors over the training period, and $\mathbf{e}_m^t$ its transposite. The covariance parameter θ_m is obtained by maximizing this function. For numerical stability and algebraic simplicity, we replace the likelihood by the log-likelihood function

$$l(\theta_m, \mathbf{e}_m) = -\frac{1}{2}\left(n\log(2\pi) + \log\left(\det\left(\sum(\theta_m)\right) + \mathbf{e}_m \sum(\theta_m)^{-1}\mathbf{e}_m^t\right)\right).$$

In order to cut back on computational cost, we consider the profile log-likelihood, use Cholesky decomposition and normalize the correlation function.

4.2 Spatial BMA

Spatial BMA is a postprocessing method for ensemble forecasts of entire weather fields, which combines the univariate techniques BMA, see Section 3.1, with the aforementioned GOP method. BMA is applied at individual locations, not taking into consideration any spatial correlations of the forecasts, whereas GOP models the spatial structure of the weather field. By combining both methods, the full information of the ensemble is used, and the spatial correlation structures are modeled, resulting in calibrated forecasts for weather fields.

Let $\mathbf{Y} = \{y(s) : s \in \mathcal{S}\}$ denote a weather field at the set of locations $\mathcal{S}$ and let $\mathbf{F}_1 = \{f_{1s} : s \in \mathcal{S}\}$,...,$\mathbf{F}_M = \{f_{Ms} : s \in \mathcal{S}\}$ denote the corresponding ensemble forecasts. Then, the spatial BMA predictive density equals

$$p(\mathbf{Y}|\mathbf{F}_1, ..., \mathbf{F}_M) = \sum_{m=1}^{M} w_m g_m(\mathbf{Y}|\mathbf{F}_m).$$

As described in Section 3.1, w_m are the weights and $g_m(\mathbf{Y}|\mathbf{F}_m)$ is the conditional density function given that member m is the best forecast in the ensemble. Here, $\mathbf{Y}|\mathbf{F}_m$ has a multivariate normal distribution, centered at the bias-corrected forecasts $a_m\mathbf{1} + b_m\mathbf{F}_m$ with covariance matrix $\Sigma_m^{'}$:

$$\mathbf{Y}|\mathbf{F}_m \sim MVN\left(a_m\mathbf{1} + b_m\mathbf{F}_m, \mathbf{\Sigma}_m^{'}\right).$$

The covariance matrix $\Sigma_m^{'}$ is a fraction of the GOP-based covariance matrix Σ_m,

$$\mathbf{\Sigma}_m^{'} = \frac{\sigma^2}{\rho_m^2 + \tau_m^2}\mathbf{\Sigma}_m,$$

where σ^2 denotes the BMA variance and the parameters ρ_m^2 and τ_m^2 are based on the GOP model. We deflate the covariance matrix Σ_m by the factor

$$\alpha_m = \frac{\sigma^2}{\rho_m^2 + \tau_m^2}$$

as the covariance is overestimated when combining BMA based on mixture densities and GOP.

In analogy to GOP, we can express the value of $\mathbf{Y}|\mathbf{F}_m$ in terms of the continuous $\mathbf{E}_{1m}$ and the discontinuous $\mathbf{E}_{2m}$ error fields,

$$\mathbf{Y}|\mathbf{F}_m = a_m\mathbf{1} + b_m\mathbf{F}_m + \mathbf{E}_{1m} + \mathbf{E}_{2m}.$$

When generating a spatial BMA ensemble, we first draw a sample from the numbers $\{1, ..., M\}$ with probabilities equal to the BMA weights w_m. Then, we dress each of the corresponding forecasts $a_m\mathbf{1} + b_m\mathbf{F}_m$ with simulations of both error fields.

For the parameter estimation of spatial BMA, we use past data within a sliding training period. We fit the BMA model to the forecast ensemble, as presented in Section 3.1, and the GOP model to each member individually (see Section 4.1). Afterwards, the deflation factor α_m is calculated in order to obtain the entries of the covariance matrix $\Sigma_m^{'}$.

Spatial BMA can be viewed as a generalized version of either BMA or GOP. If we consider an ensemble of size one, it reduces to GOP. If we only consider one location, spatial BMA becomes the regular BMA model.

4.3 Spatial EMOS$^+$

Analogously to spatial BMA, we propose a spatial EMOS$^+$ approach, where we combine EMOS$^+$ with the GOP method. However, since the EMOS$^+$ variance parameter σ_s^2 varies spatially, the approach differs slightly from spatial BMA.

For a weather field $\mathbf{Y} = \{y(s) : s \in \mathcal{S}\}$, considered at the set of the locations $\mathcal{S}$, let $\mathbf{F}_1 = \{f_{1s} : s \in \mathcal{S}\}, ..., \mathbf{F}_M = \{f_{Ms} : s \in \mathcal{S}\}$ denote the corresponding ensemble forecasts.

Then, $\mathbf{Y}|\mathbf{F}_m$ has a multivariate normal distribution, centered at the sum of the bias-corrected forecasts with a covariance matrix $\mathbf{\Sigma}^{''}$,

$$\mathbf{Y}|\mathbf{F}_1, ..., \mathbf{F}_M \sim MVN\left(a\mathbf{1} + b_1\mathbf{F}_1 + ... + b_M\mathbf{F}_M, \mathbf{\Sigma}^{''}\right).$$

The covariance matrix $\mathbf{\Sigma}^{''}$ is expressed by

$$\mathbf{\Sigma}^{''} = V\mathbf{\Sigma}^0 V,$$

where $V = \text{diag}(\sqrt{c + dS_1^2}, ..., \sqrt{c + dS_{\#\mathcal{S}}^2})$ is a diagonal matrix with entries equal to the estimated location-specific standard deviations predicted by EMOS$^+$ and $\mathbf{\Sigma}^0$ is a correlation matrix, based on the GOP method.

Of course, we can state the values of $\mathbf{Y}|\mathbf{F}_m$ in terms of the continuous $\mathbf{E}_{1m}$ and the discontinuous $\mathbf{E}_{2m}$ error fields,

$$\mathbf{Y}|\mathbf{F}_1, ..., \mathbf{F}_{Mm} = a\mathbf{1} + b_1\mathbf{F}_1 + ... + b_M\mathbf{F}_M + \mathbf{E}_{1m} + \mathbf{E}_{2m}.$$

For the production of a spatial EMOS$^+$ ensemble, we first calculate the multivariate bias-corrected forecast $a\mathbf{1} + b_1\mathbf{F}_1 + ... + b_M\mathbf{F}_M$. In the following step, the corresponding error fields are simulated and then added to the bias-corrected forecast field.

When estimating the parameters for spatial EMOS$^+$, we first fit EMOS$^+$ to past data in a sliding training period, in order to obtain the parameters a,b_1,...,b_M and the predicted variances $c + dS_s^2$. Then, on a given day, given the error field $\mathbf{e}$, which for simplicity is defined as the difference between the ensemble mean and the verifying observation, we standardize the values by dividing each entry of $\mathbf{e}$ by the site-corresponding predicted standard deviation of EMOS$^+$. To these normed values, accumulated over all days in the training period, we fit a geostatistical model, as described in 4.1, in order to obtain the parameters for $\mathbf{\Sigma}^0$.

4.4 Reference Forecasts

Ensemble Copula Coupling (ECC)

ECC, proposed by Schefzik (2011), is a multivariate postprocessing technique for ensemble forecasts. Based on existing univariate postprocessing methods, ECC models the multivariate dependency structure of the forecasts by incorporating the multivariate rank structure of the original ensemble through a discrete copula. In the current context, we employ ECC based on BMA and EMOS$^+$, discussed in Chapter 3. When generating the ECC forecasts, we proceed according to the following steps:

1. *Univariate postprocessing*

   First, we apply any available postprocessing method to the ensemble output, in order to produce a calibrated predictive distribution. Subsequently, for each location $s \in \mathcal{S}$, we draw a random sample $\hat{f_{1s}}, ..., \hat{f}_{Ms}$ of the original ensemble size M from this distribution.

2. *Combining the results of step 1 with the ensemble's dependency structure*

   Given the ensemble forecasts $f_{1s}, ..., f_{Ms}$, we denote their ranks $\omega(s,1), ..., \omega(s,M)$ at each station. Then, we sort the random sample according to the ensemble's order statistic: $\hat{f}_{\omega(s,1)}, ..., \hat{f}_{\omega(s,M)}$. Finally, one member m of the ECC ensemble equals the vector $(\hat{f}_{\omega(1,m)}, ..., \hat{f}_{\omega(\#\mathcal{S},m)})$.

If a larger ensemble is desired, the steps may be repeated $n \in \mathbb{N}$ times, in order to generate an ensemble of the size nM. ECC provides an easy technique, which in our case produces spatially consistent forecast fields by inheriting the dependence structure of the original ensemble.

Noise Ensemble

In order to account for measurement errors or small scale spatial variations, we include the noise ensemble, which was also considered in Berrocal et al. (2007). To each member m of the raw ensemble, we add a Gaussian noise with mean zero and a variance equal to the corresponding nugget effect ρ_m^2. This task is performed independently at each location. Thus, the noise ensemble does not capture the spatial structure of the weather field, but includes patterns of past site-specific errors, in order to improve the forecasts.

Table 4.1: The results of the multivariate assessment of the surface temperature in Saarland. With different correlation structures, the GOP method is applied to the ensemble mean. All scores are in degrees Celsius and averaged over the time period from 5 January 2011 until 30 November 2011.

	ES	EE
GOP Mean Exponential	3.59	4.92
GOP Mean Matérn	3.59	4.92

Table 4.2: The results of the minimum temperature along a section of the highway A3. The scores MAE, RMSE and the CRPS are in degrees Celsius and averaged over the time period from 5 January 2011 until 30 November 2011. The Brier score for the event that the temperature drops beneath 0°C is only calculated during the winter months, January, February and November.

	CRPS	MAE	RMSE	Brier Score
GOP Mean Exponential	0.87	1.22	1.56	0.078
GOP Mean Matérn	0.89	1.23	1.58	0.084

4.5 Results

We apply the presented methods to 21-h forecasts of surface temperature, issued by COSMO-DE-EPS (see Chapter 2). As discussed in Section 3.5, we use a sliding 25-day training period and the evaluation starts on 5 January 2011, ending on 30 November 2011. For each model, we simulate $10,000$ realizations of the forecasts and then, in order to calculate the scores, we utilize approximation techniques, described in Appendix A. Only the raw and noise ensemble are evaluated with 20 members.

4.5.1 Different Modeling of Spatial Structure for GOP

As mentioned in Section 4.1, there are various possibilities to model the spatial structure and estimate the corresponding parameters for GOP. Before, we have discussed applying a Mátern correlation function, in contrast to the approach by Gel et al. (2004), who use an exponential correlation function. Moreover, we have presented two different ways to estimate the parameters of the GOP model: via a variogram-based method and via maximum likelihood. Thus, before combining GOP with univariate postprocessing methods, we compare a sophisticated approach with a simpler one. On the one hand, we model the spatial structure with a Mátern correlation function and estimate the parameters via maximum likelihood. On the other hand, we base the GOP method on an exponential

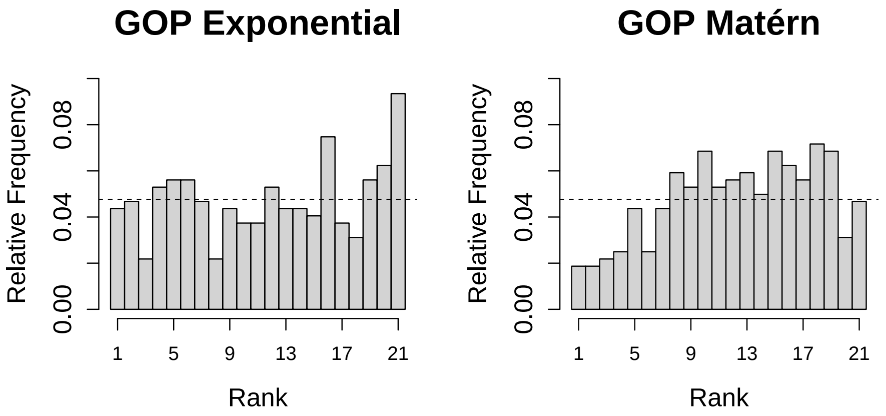

Figure 4.1: Rank histograms for forecasting the minimum temperature along a section of Highway A3. The results are aggregated over the time period from 5 January 2011 until 30 November 2011.

correlation function and utilize the estimation via variograms, constructing the empirical variograms with the `R` package `RandomFields` by Schlather (2011).

We apply these two competing techniques to the ensemble mean, evaluate the performance at seven stations in Saarland, and assess the forecasts for minimum temperature at a section along the highway A3. We will discuss the choices of the evaluation area and the aggregated variable further in the following section. As seen in Tables 4.1, 4.2 and Figure 4.1, the forecasting performance differs only slightly. For Saarland, the two models coincide. Looking at the results along the highway, the simpler model even outperforms the more complex one slightly. Hence, the simple estimation approach via variograms yields good results. Based on these results, we base our geostatistical model on an exponential correlation function with variogram-based estimation in the following. This choice is further supported by the temporal evolution of the parameter ν in the Mátern correlation function, as it averages at 0.41 over the forecasting period. Thus, it is very close to 0.5, for which both correlation functions coincide.

4.5.2 Overall Performance of the Models

Having established the methodology for the GOP technique, we compare the performance of all aforementioned models: spatial BMA, ECC based on BMA, univariate BMA, spatial EMOS$^+$, ECC for EMOS$^+$, regular EMOS, GOP, as well as the raw and noise ensemble. Since GOP is not an ensemble-based method, we apply it to member 15 of the COSMO-DE-

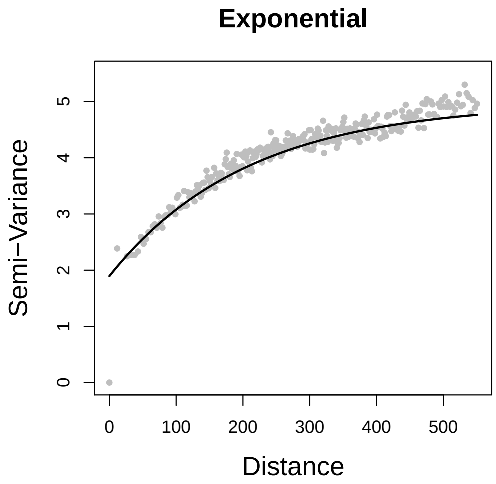

Figure 4.2: Empirical variogram of errors for member E15 on 28 November 2011 with a fitted exponential variogram

EPS, which was assigned the highest BMA weight of 0.31 over the course of the forecasting period.

Figure 4.2 shows an empirical variogram of the error field for member 15 on 28 November 2011 with a fitted exponential variogram. According to the figure, the exponential correlation function represents the spatial structure well, which additionally supports the choice of a simpler geostatistical model in Section 4.5.1.

Given forecast fields with 514 dimensions, assessing the predictive performance of vector-valued quantities in such high dimensions is challenging. So, we employ different evaluation approaches. On the one hand, to check if the forecast method captures the correlation structure of the weather field well, we apply a variogram-based approach. On the other hand, in order to reduce computational cost, we consider a subset of our data set and investigate the predictive performance for minimum temperature along the highway A3. Additionally, we check the forecasting performance for all seven observation sites in Saarland with multivariate techniques. More evaluation results can be found in Appendix C.

In order to evaluate how well the models reproduce the spatial correlation structure, we compute empirical variograms of the verifying observations and calculate the $19/21 \approx 90.5\%$ point-wise prediction intervals for the variogram values of the different methods, discussed in Appendix A. Figure 4.4 shows an example for 28 January 2011. All multivariate techniques capture the spatial structure and the variogram values of the observations fall mainly within the boundaries of the prediction interval, while the raw and noise ensemble as well as the regular BMA approach fail to describe the dependencies of the weather field. However, the raw ensemble seems to capture the spatial structure, but underestimates the variance.

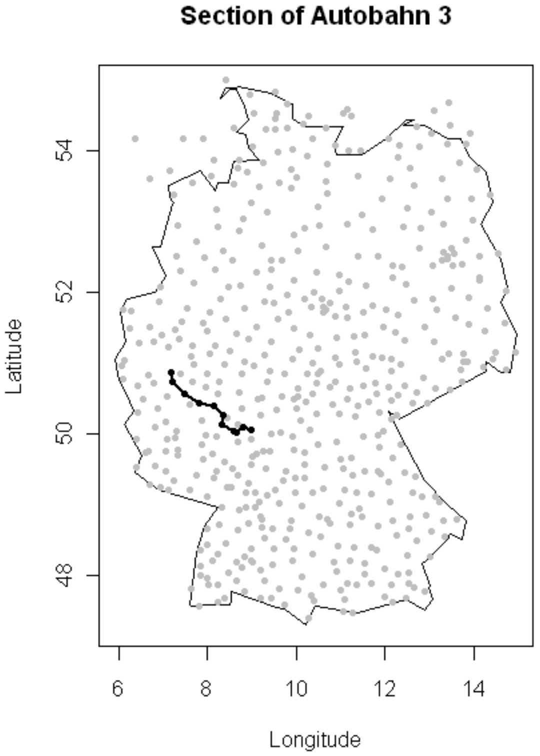

Figure 4.3: The map of Germany shows a section of Highway A3. We evaluated the predictive performance of the minimum temperature at eleven nearby stations, which are marked in black. The gray points represent all stations for which forecasts are available.

This explains the good results for both ECC approaches, which still rely on the structure of the ensemble, but correct the variance term. The good performance of EMOS$^+$ on this specific day is probably a coincidence, as Table 4.3, where the averaged coverage over the forecasting period is shown, indicates, that only multivariate postprocessing techniques reach results higher than 50%. Only methods based on a more sophisticated approach, which model the dependency structure with a geostatistical model, as spatial BMA, spatial EMOS$^+$ and GOP member 15, yield values in the 60% range.

Spatial correlation has a huge impact on composite quantities such as minimum temperature or average temperature. When only considering these quantities, the verification process reduces to a one-dimensional one, which gives us the opportunity to employ well-established univariate verification methods. We consider a section with eleven observation sites along Highway A3, which connects the two large German cities of Cologne and Frankfurt and thus is one of the busiest highways in Germany (Figure 4.3). Hence, reliable weather predictions are crucial for maintenance operations of the highway.

Figure 4.5 shows histograms of the accumulated ranks over the forecasting period. Again, the methods based on modeling the dependency structure with a geostatistical model

Table 4.3: Coverage of the $19/21 \approx 90.5\%$ point-wise prediction intervals for variogram values. The results are averaged over the time period from 5 January 2011 until 30 November 2011.

Model	Coverage in %
Raw Ensemble	19.38
Noise Ensemble	37.11
GOP Member 15	63.73
Spatial BMA	67.64
ECC BMA	55.37
BMA	34.40
Spatial EMOS$^+$	64.82
ECC EMOS$^+$	55.69
EMOS$^+$	25.80

Table 4.4: Assessment results for forecasting the minimum temperature along the section of Highway A3. MAE, RMSE and the CRPS in degrees Celsius are averaged over all stations and the time period from 5 January 2011 until 30 November 2011. The Brier score for the event that the temperature drops beneath 0°C is only calculated during the winter months, January, February and November.

Model	CRPS	MAE	RMSE	Brier Score
Raw Ensemble	1.72	1.92	2.33	0.120
Noise Ensemble	1.21	1.56	1.91	0.107
Spatial BMA	0.86	1.21	1.55	0.081
ECC BMA	0.95	1.28	1.64	0.102
BMA	1.08	1.41	1.81	0.120
Spatial EMOS$^+$	0.87	1.22	1.56	0.083
ECC EMOS$^+$	0.92	1.25	1.61	0.094
EMOS$^+$	1.05	1.37	1.77	0.114
GOP Member 15	0.88	1.22	1.57	0.086

perform best, as the corresponding histograms of spatial BMA, spatial EMOS$^+$ and GOP member 15 appear almost uniform. For the ECC versions and the univariate postprocessing methods, however, the forecasts are biased, as many observations are higher than the predictions. Additionally, the raw ensemble shows a strong bias and often overestimates the minimum temperature. The same pattern applies to the noise ensemble, but not in such a predominant way. All non-geostatistical models underestimate uncertainty, as many observations obtain ranks on the edge of the ensemble.

The scores in Table 4.4 further support these results, as spatial BMA, spatial EMOS$^+$ and GOP member 15 yield the lowest CRPS scores. The corresponding values differ only slightly, which might suggest that the ensemble information, gained by BMA or EMOS$^+$, is

Table 4.5: Multivariate and univariate assessment of the surface temperature in Saarland. The scores are all in degrees Celsius and are averaged over all stations and the time period from 5 January 2011 until 30 November 2011.

Model	ES	EE
Raw Ensemble	5.62	5.86
Noise Ensemble	5.65	5.90
Spatial BMA	3.56	4.89
ECC BMA	3.61	4.90
BMA	3.58	4.90
Spatial EMOS$^+$	3.57	4.91
ECC EMOS$^+$	3.61	4.91
EMOS$^+$	3.59	4.90
GOP Member 15	3.62	4.94

not that beneficial for the overall performance. The ECC techniques, which incorporate the original structure of the ensemble, perform well, whereas the raw and noise ensemble, as well as the univariate forecasting methods fail to deliver reliable predictions. These findings are reflected by the other scores, MAE and RMSE, as well. For the Brier score, only the geostatistical methods yield good results, followed by ECC for EMOS$^+$. All other methods perform poorly when forecasting the probability of the temperature to drop beneath 0°C.

Considering a subsection of the overall data, we look at seven stations, covering the state of Saarland. As seen in Table 4.5, we apply multivariate assessment tools. The energy score and Euclidean error are lower for all postprocessed forecast, compared to the raw and noise ensemble. However, we find that the scores seem to be insensitive to slight variations in the multidimensional structure and thus, the difference between the univariate and spatial techniques vanishes. Figure 4.6 shows the minimum spanning tree histograms of the competing forecast models. For the raw ensemble, many low MST ranks are observed, resulting in uncalibrated forecasts. The histograms of the remaining models appear close to uniform, fulfilling the necessary condition for calibrated forecasts.

This chapter has presented several approaches to multivariate forecasting. We have demonstrated the need to model spatial correlation with a geostatistical approach, in order to obtain spatially consistent, sharp and calibrated temperature forecasts. The benefits of these approaches can be recognized in the forecasting performance, illustrated by the scores.

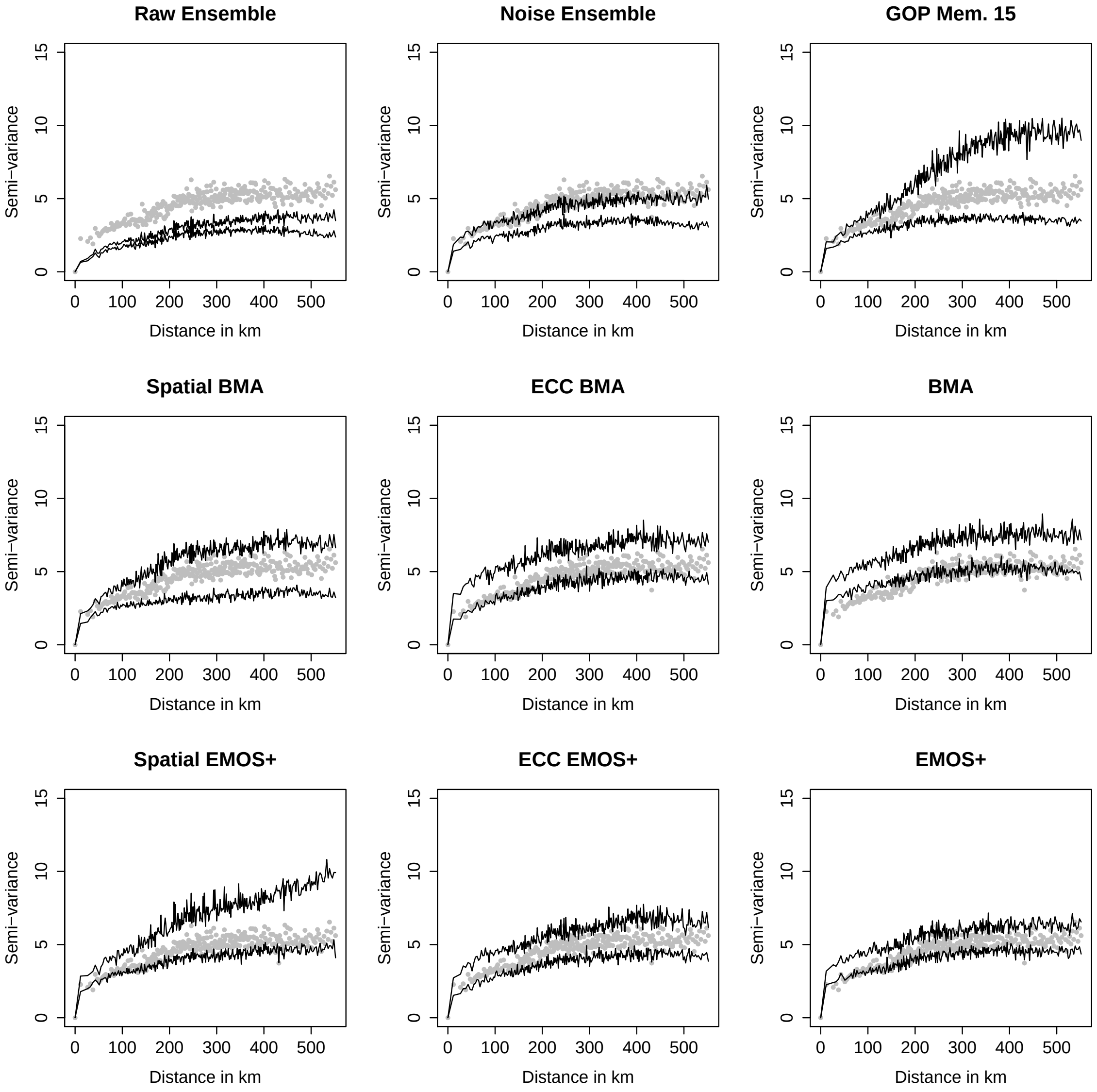

Figure 4.4: Empirical variogram coverage: The gray points show the variogram values of the verifying observations on 28 January 2011. The black lines are the limits of the $19/21 \approx 90.5\%$ point-wise prediction intervals for a 20 member ensemble of the respective models.

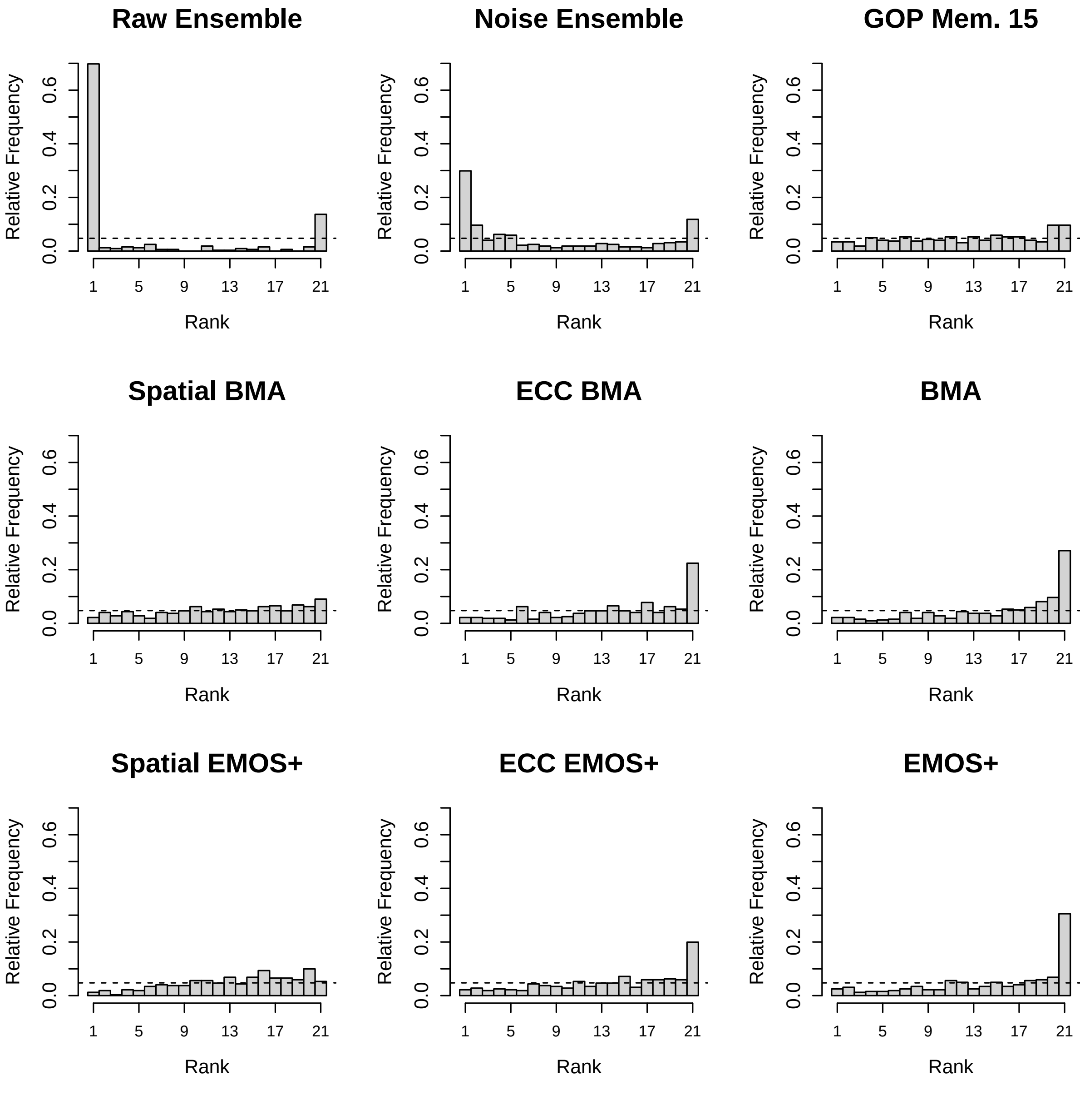

Figure 4.5: Rank histograms for forecasting the minimum temperature along a section of Highway A3. The results are aggregated over the time period from 5 January 2011 until 30 November 2011.

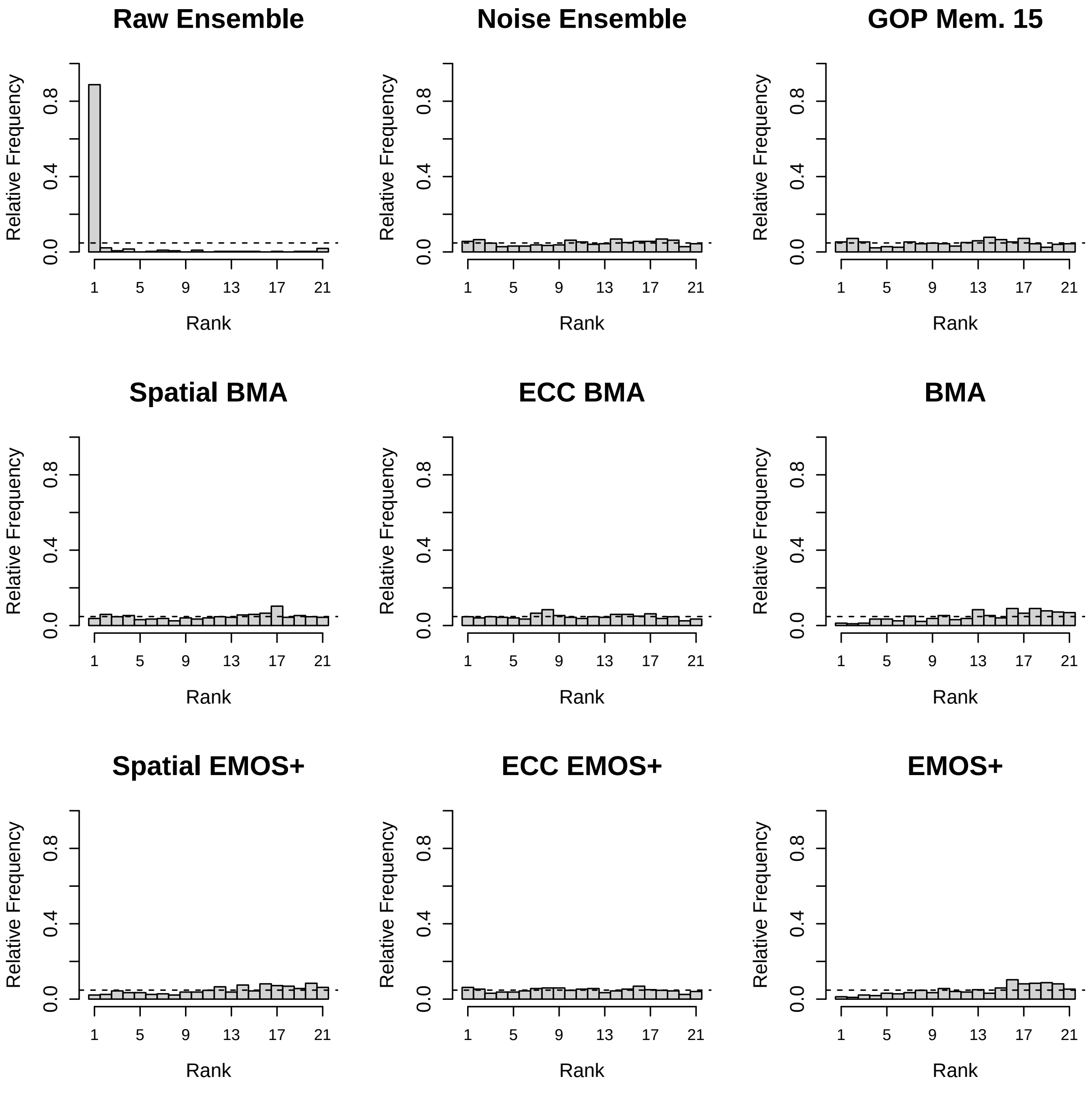

Figure 4.6: Minimum spanning tree rank histograms for forecasts of temperature at seven stations in Saarland. The results are aggregated over the time period from 5 January 2011 until 30 November 2011.

Chapter 5

Discussion

In this thesis, we have demonstrated the importance of spatial modeling in ensemble postprocessing. Besides reviewing the existing method spatial BMA (Berrocal et al., 2007), we have introduced a similar approach, based on EMOS$^+$. Both models are obtained by combining state of the art univariate postprocessing methods with GOP. Thus, the predictive distributions are based on spatial statistical and ensemble information. For spatial EMOS$^+$, the predictive density for a weather field is a multivariate normal density centered at the bias-corrected ensemble, whereas in the case of BMA the predictive density is a weighted average over the multivariate Gaussian density of each bias-corrected member.

In a case study with COSMO-DE-EPS, the spatial EMOS$^+$ and spatial BMA outperformed all other methods including the multivariate ECC approaches. In particular, when evaluating the composite quantities of minimum and average temperatures, the multivariate models showed their superiority.

However, there remain several possibilities for improvements and further investigation. The presented methods, spatial BMA and EMOS$^+$, are only applicable to weather variables whose distribution of forecast errors can be approximated by a normal distribution. Berrocal et al. (2008) proposed a two-stage spatial model for producing correlated probabilistic forecasts of precipitation accumulation. Further development is needed in order to address variables such as wind gust or even wind speed.

Additionally, there are options for advances within spatial BMA and EMOS$^+$. For both methods, the bias correction is based on linear least squares regression, which does not account for differences in terrain, such as altitude or land use. A more sophisticated approach might yield better results.

Moreover, local parameter estimation such as proposed by Kleiber et al. (2011) could further improve the forecast performance. In particular considering the histogram with a uniform center and two bars at the sides, discussed in Section 3.6, a model with local parameters might resolve this phenomenon. If the forecasts were calibrated locally, the structure of the error fields would be more realistic and the overall parameter estimation for the GOP method would yield better results.

In the case of GOP, there are several ways to reduce the computational cost of parameter estimation. We used a rolling training period and estimated the covariance parameters daily. If the variable of interest is temporal stationary over the years and a larger data set is available, the estimation can be based on a previous year or season. For spatial EMOS$^+$, we only estimated the covariance parameters for the ensemble mean, which performed as well as spatial BMA, for which the parameters are estimated individually for each member. Hence, in order to speed up computational time for spatial BMA, the member-specific parameter estimation can be replaced as well.

When modeling the covariance structure with the GOP method, we have shown that an exponential correlation function with parameters based on simple variogram estimation yields slightly better results than the modeling with a Mátern correlation function, whose parameters are obtained by maximum likelihood estimation. However, there might still be room for improvement with different covariance structures, as discussed in e.g. Gneiting (1999), which represents the atmosphere more accurately.

Obviously, our choice of the length of the training period is subject to debate. However, due to shortage of data, out-of-sample comparison of different lengths was not possible and so we followed the proposal by Berrocal et al. (2007). Given data sets over longer time periods, data from the same season of previous years could be included into the training period (Hamill et al., 2004).

When evaluating the forecasting performance of the multivariate postprocessing methods, we found the improvements over univariate postprocessing to be quite significant in the assessment of forecasts for aggregated quantities. However, they were not visible in the multivariate energy score and Euclidean error. We suspect that these scores are not sufficiently sensitive to slight variations in the multidimensional structure, and thus are not a good verification tool for this task, but further research on this topic is needed.

Motivated by the high correlation of the forecast errors within COSMO-DE-EPS, which was neither accounted for by the BMA weights nor the EMOS coefficients, as the values varied drastically over time, we also investigated a ridge regression approach. When calculating the predictive mean, we applied linear least squares, but also introduced a penalty term, which should force the coefficients to be similar, if the members were strongly correlated. We proposed that two members should be assigned similar coefficients, if they were based on the same global model or were produced by the same model formulation of COSMO-DE-EPS. However, our attempt failed to yield better results than original EMOS or BMA in terms of MAE and consequently this line of research was not pursued any further.

In conclusion, we have demonstrated that spatial EMOS$^+$ and spatial BMA perform very well, despite the fact that there are still many possibilities for improvement or changes in either method. Our verification results for composite quantities have demonstrated once again the need for spatial modeling in ensemble postprocessing. Also, we have addressed some of the issues that arise in the statistical postprocessing of COSMO-DE-EPS and hope that further research will provide solutions to these.

Appendix A

Verification Methods

The goal of probabilistic forecasts is to "maximize sharpness subject to calibration" (Gneiting et al., 2007). Calibration depends on the observations as well as the forecasts, measuring the statistical consistency between both. This means that a predicted event should on average occur as often as its forecast probability indicates. Sharpness is a characteristic of the forecast only and measures the concentration of the predictive distribution. Illustrating the concept, a small standard deviation yields a sharp forecast, whereas a large dispersion around the distribution's mean does not reflect a sharp forecast.

There are several techniques to assess the predictive performance of forecasts, considering sharpness and calibration individually as well as simultaneously. Here, we present a selection of tools which we have used in the proceeding chapters to evaluate the performance of the competing forecasters. On account of the fact that we produce one-dimensional forecasts as well as forecast fields with up to 514 dimensions, the shown methods not only cover the univariate, but also the multivariate case. They can be applied to ensemble forecasts as well as predictive distributions. However, the distinction between both is rather artificial, because by sampling from the predictive distribution, an ensemble is obtained or, on the other hand, an ensemble can be replaced by an appropriate distribution. Thus, both types of forecasts can be transformed into each another.

Here, we first present tools to check calibration, followed by a method to evaluate the sharpness of forecasts. Then, we introduce proper scoring rules, which address sharpness and calibration simultaneously. Finally, we present the empirical variogram coverage for forecasting fields.

A.1 Assessing Calibration

As mentioned before, calibration is a measure for the statistical consistency between the forecasts and the verifying observations, which means that an event with a certain predicted probability should on average occur as many times as has been predicted. Depending on the type of the forecast, there are several ways to assess calibration. Given a probabilistic forecast with a predictive cumulative distribution function (CDF) for a univariate quantity, usually a probability integral transform (PIT) histogram (Dawid, 1984; Diebold et al., 1998) is employed. In case of an ensemble forecast, the corresponding counterpart is the verification rank histogram (VRH) or Talagrand diagram (Anderson, 1996; Hamill and Colucci, 1997; Talagrand et al., 1997). A method to assess the calibration of a multivariate quantity is the minimum spanning tree (Smith, 2001; Gneiting et al., 2008).

Probability Integral Transform Histogram

In order to assess the calibration of a univariate forecast distribution F, the PIT histogram (Dawid, 1984; Diebold et al., 1998) is frequently used. Its idea is based on the assumption that nature samples the materializing event y from an unknown, true distribution G. Then, for all forecasts and observations available, the PIT values $p = F(y)$ are determined and collected. For F to equal G and therefore an ideal forecaster, it is a necessary condition that these values have a uniform distribution on $[0, 1]$ (Gneiting et al., 2007).

In practice, for every forecast the PIT value is computed, collected and afterwards sorted into bins in order to plot a histogram. When interpreting the outcome of the histogram, we take a look at its shape. If the histogram shows uniformity, it indicates calibration. A U-shape reveals that the predictive distributions are underdispersive, meaning that they are too narrow. On the other hand, a hump-shaped histogram suggests that too many observations lie in the center of the distributions and therefore the prediction intervals are too wide, resulting in overdispersion. If the histograms resemble a triangle, the predictive distributions are biased.

Rank Histogram

Considering a discrete ensemble forecast, the VRH or Talagrand diagram (Anderson, 1996; Hamill and Colucci, 1997; Talagrand et al., 1997) replaces the PIT histogram. The interpretation of the resulting histogram stays the same, but the construction differs. Given an ensemble of size M, for each forecast available, the rank of the observation y in the set of all forecasts $x_1, ..., x_M$, combined with y, is determined. Afterwards, a histogram of all aggregated ranks, which range from 1 to $m + 1$, is plotted.

Minimum Spanning Tree Rank Histogram

In order to check the calibration of multivariate ensemble forecasts, the minimum spanning tree (MST) rank histogram (Smith, 2001) is an appropriate tool. Evaluating an ensemble forecast $\{\mathbf{x}_i \in \mathbb{R}^d | i = 1, ..., M\}$, $d \in \mathbb{N}$, with the observation $\mathbf{x}_0 \in \mathbb{R}^d$, the construction of the MST histogram follows these steps:

1. *Standardize*

   Sometimes, it can be useful to apply a principal component transform to the set of forecasts and observations $\{\mathbf{x}_i \in \mathbb{R}^d | i = 0, ..., M\}$, in order to generate standardized quantities $\{\mathbf{x}_i^\star \in \mathbb{R}^d | i = 0, ..., M\}$.

2. *Compute minimum spanning tree*

   Considering every subset $\{\mathbf{x}_i^\star \in \mathbb{R}^d | i \in \{0, ..., M\} \setminus j\}$, for $j = 0, ...M$, where either the observation or one of the ensemble members has been removed, determine the minimum spanning tree and calculate its length, $l_j > 0$.

3. *MST rank*

   The MST rank r equals the rank of l_0 in the pool of all lengths l_i, $i = 0, ..., M$. If ties occur, they are resolved at random. Let

   $$s^< = \sum_{i=0}^{M} \mathbb{I}(l_i < l_0)$$

   be the number of MST lengths that are smaller than the length of the MST without the observation, and

   $$s^= = \sum_{i=0}^{M} \mathbb{I}(l_i = l_0)$$

   the number of MST lengths which equal the length of the MST without the observation. Then the MST rank r is chosen from a discrete uniform distribution on the set $\{s^< + 1, ..., s^< + s^=\}$, so that $r \in \{1, ..., M + 1\}$.

4. *Aggregate ranks and plot histogram*

   Finally, we collect all MST ranks over all dates and locations available and plot the corresponding histogram.

For the calculation of the minimum spanning trees, we employ the `R` package `vegan` by Oksanen et al. (2011).

The interpretation of this tool differs from that of other histogram-based methods. An underdispersed or biased ensemble results in many low MST ranks. In contrast, if the ensemble is overdispersed, the higher ranks are overpopulated. Given an ideal forecaster, so that the ensemble members and the observation can be considered exchangeable, the MST rank histogram appears uniform.

A.2 Assessing Sharpness

Sharpness measures the concentration of the predictive distribution and is therefore a feature of the forecast only. Following the principle of "maximizing sharpness subject to calibration" (Gneiting et al., 2007), the more concentrated the distribution, the sharper it is, the better we rate it, subject to calibration. When evaluating the sharpness, we determine the width of the prediction intervals. To facilitate comparability of the postprocessed forecasts with the M member ensemble, often the nominal $\frac{M-1}{M+1} \cdot 100\%$ prediction interval is used, because its coverage corresponds to that of the ensemble range.

A.3 Proper Scoring Rules

A very important tool to assess a forecaster's performance are proper scoring rules, which evaluate sharpness and calibration concurrently. Scores are negatively oriented and can be interpreted as a penalty, which the forecaster wants to reduce. They are based on the predictive distribution F and the verifying observation y. If we assume y to be drawn from a true, but unknown distribution G, the expected value is written as $s(F, G)$. We call a scoring rule proper if it is minimized for $F = G$; it is strictly proper if $s(F, G) > s(G, G)$ for all $F \neq G$. The theory of strictly proper scoring rules is discussed further in Gneiting and Raftery (2007).

Here, we present the equations for each score independently of time and space. In practice, we calculate the scores for all times and locations available and then determine the average score

$$\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s(F_i, y_i),$$

where n denotes the number of available forecasts.

Brier Score

The Brier score (BS), also referred to as quadratic score, is a well-known scoring rule for univariate quantities with a CDF, which was first presented by Brier (1950), and is defined as

$$\mathrm{bs}(z) = \left(F(z) - \mathbb{I}(z \geq y)\right)^2,$$

where $z \in \mathbb{R}$ is a threshold value, F represents the predictive distribution and y the verifying observation.

Continuous Rank Probability Score

The continuous rank probability score (CRPS), also applicable to univarite quantities, corresponds to the integral of the BS over all threshold values $z \in \mathbb{R}$ (Toth and Kalnay, 1997). It was proposed by Matheson and Winkler (1976) and further developed in Hersbach (2000), Gneiting et al. (2005) and Wilks (2006):

$$\mathrm{crps}(F, y) = \int_{-\infty}^{\infty} (F(z)) - \mathbb{I}(z \geq y))^2 dz = \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_P|X - X'|,$$

where F is the predictive distribution, y represents the verifying observation, and X and X' are independent random variables with distribution function F and finite first moment. Gneiting and Raftery (2007) showed the second equality.

Energy Score

A multivariate generalization of the CRPS is the energy score (ES), described in Gneiting and Raftery (2007), which is defined as

$$\mathrm{es}(F, \mathbf{y}) = \mathbb{E}_F||\mathbf{X} - \mathbf{y}|| - \frac{1}{2}\mathbb{E}_F||\mathbf{X} - \mathbf{X}'||,$$

where $||\cdot||$ is the Euclidean norm, F represents the predictive distribution, $\mathbf{y} \in \mathbb{R}^d$ is the observation and $\mathbf{X}\in \mathbb{R}^d$ and $\mathbf{X}'\in \mathbb{R}^d$ are independent random vectors with distribution F. For $d = 1$, the energy score reduces to the CRPS. In order to facilitate computability, we replace the ES for normal or related predictive densities with a Monte Carlo approximation. In this case, we draw a sample $\mathbf{x}_1, ..., \mathbf{x}_k \in \mathbb{R}^d$ of size $k = 10,000$ from the predictive distribution F and estimate the ES via

$$\hat{\text{es}}(F, \mathbf{y}) = \frac{1}{k}\sum_{i=1}^{k} ||\mathbf{x}_i - \mathbf{y}|| - \frac{1}{2(k-1)}\sum_{i=1}^{k-1} ||\mathbf{x}_i - \mathbf{x}_{i+1}||.$$

We follow the same technique when calculating the CRPS. Given a forecast ensemble, instead of a predictive distribution, the ES can be calculated, using

$$\text{es}(F_{ens}, \mathbf{y}) = \frac{1}{M}\sum_{i=1}^{M} ||\mathbf{x}_i - \mathbf{y}|| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} ||\mathbf{x}_i - \mathbf{x}_j||,$$

where point masses $\frac{1}{M}$ are placed on each ensemble member $\mathbf{x}_1, ..., \mathbf{x}_M \in \mathbb{R}^d$.

Mean Absolute Error

In the case of a deterministic forecast, the CRPS can be reduced to the absolute error (Gneiting et al., 2005):

$$\text{ae}(\mu, y) = |\mu - y|,$$

with μ being the median of F.

Euclidean Error

The multivariate generalization of the absolute error is the Euclidean error

$$\text{ee}(F, \mathbf{y}) = ||\text{smed}_\text{F} - \mathbf{y}||,$$

where smed_F defines the spatial median of the predictive distribution F, which can be defined as (Gneiting, 2011; Vardi and Zhang, 2000)

$$\mathrm{smed}_F = \arg\min\nolimits_{\mathbf{x}\in\mathbb{R}^d} \mathbb{E}_F ||\mathbf{x} - \mathbf{X}||,$$

with $\mathbf{X}$ a random vector with distribution F. For the calculation of the spatial median, we employ the `R` package `ICSNP` by Nordhausen et al. (2010).

A.4 Empirical Variogram Coverage

This is an alternative approach to evaluate the spatial performance of forecasting ensembles, which is also used in Berrocal et al. (2007). Considering one forecast field, we calculate the empirical variogram values for the observed temperature as well as all members of the ensemble individually. Then, at each distance we determine the maximum and minimum of the semi-variance within the values of the ensemble, in order to generate the borders of the point-wise $\frac{M-1}{M+1} \cdot 100\%$ prediction intervals. If the forecaster simulates the true spatial structure, $\frac{M-1}{M+1} \cdot 100\%$ of the empirical variogram values of the observations should fall within the prediction interval. Finally, this technique is repeated over all forecast fields available and averaged.

Appendix B

Plots of the Empirical Variograms

In R, there are several packages provided for calculating empirical variograms. Berrocal et al. (2007) suggest the use of the package ProbForecastGOP (Berrocal et al., 2010). We also compute variograms based on the package RandomFields by Schlather (2011). As Figure B.1 shows, the results for the same member differ substantially on certain days, depending on the package. Since models based on RandomFields yield better forecasting results, we use this package for the entire thesis.

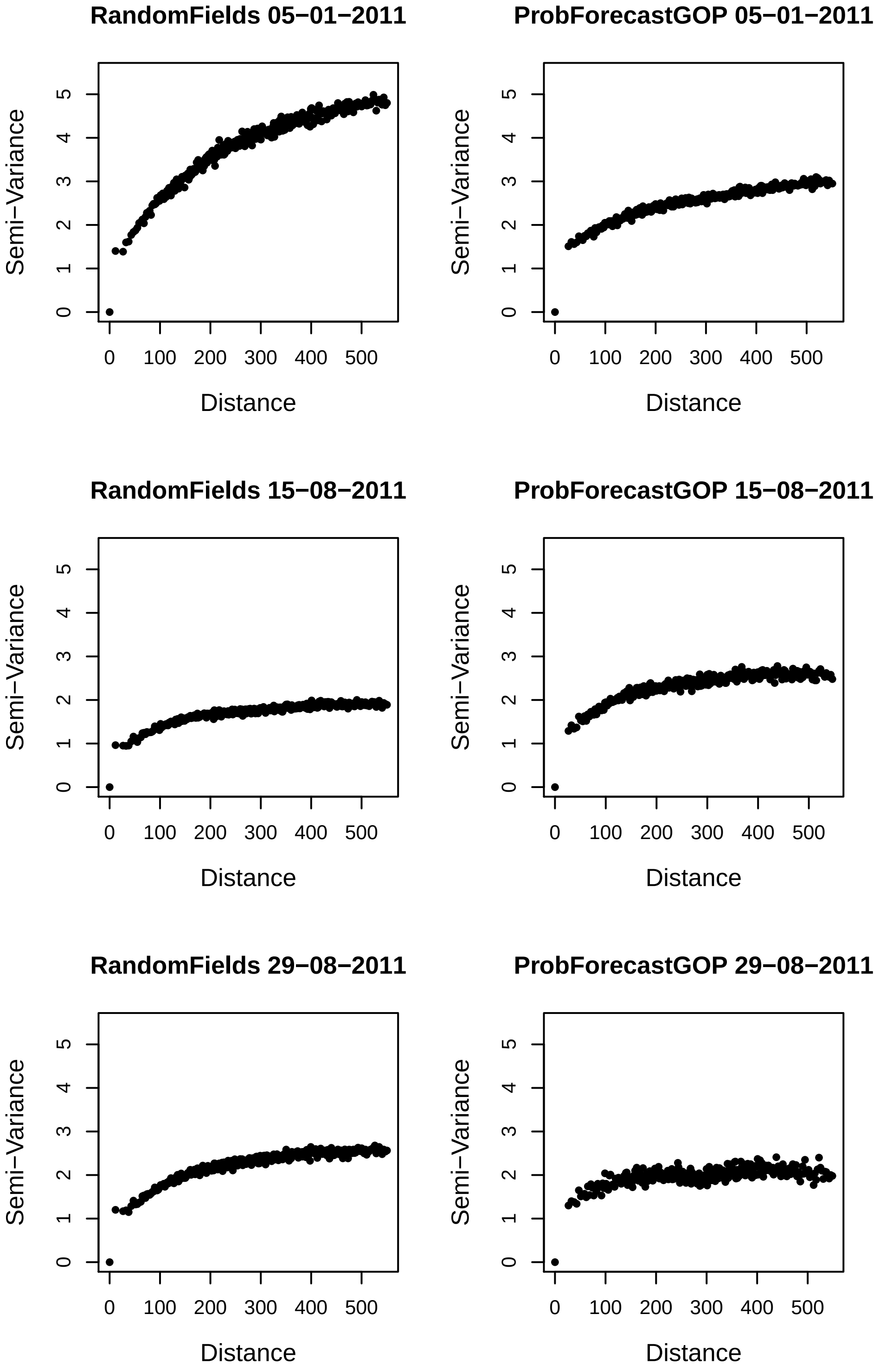

Figure B.1: Comparison of empirical variograms for member 15 of COSMO-DE-EPS on different days, calculated by the `R` packages `RandomFields` and `ProbForecastGOP`. Each variogram is based on 300 bins, for which the cut points are kept equal over all days and for both packages.

Appendix C

Further Verification Results

In addition to the results presented in Subsection 4.5.2, we also consider multivariate assessment of different subsets of the data, another aggregated variable, and the univariate performance of the forecasts. The interpretation of these results does not differ from the discussion in 4.5.2. However, for the sake of completeness, we add our findings.

Table C.1 presents the scores for the prediction of the average temperature in every German state individually; the results are then averaged over all states. Figure C.1 exemplary shows the rank histograms for Hessen.

Furthermore, we consider two subsets of the data. One contains the observation sites at the airports in Hamburg, Berlin and Frankfurt, separated by a great distance, and the other consists of three closely located stations in Berlin. The corresponding scores can be seen in the Tables C.2 and C.3, as well as the minimum spanning tree rank histograms in the Figures C.2 and C.3.

For forecasts at individual sites, spatial BMA or ECC BMA are equal to the original BMA. Thus, the one-dimensional assessment of the spatial techniques coincides with their univariate counterparts, as can be seen in Table C.4 and Figure C.4. The slight differences are due to variability in drawing the forecast samples. Theoretically, the same applies to EMOS$^+$. However, we found that for spatial EMOS$^+$ the predicted site-specific variance is greater than for regular EMOS$^+$. The reason for this might be rooted in the normalization process of the error field.

Table C.1: Verification results for forecasting the average temperature in 14 of Germany's states. The MAE, RMSE, CRPS and the width of the nominal 19/21 ≈ 90.5% prediction interval in degrees Celsius, as well as its coverage are aggregated over all states and the time period from 5 January 2011 until 30 November 2011.

				Prediction Intervals	
	CRPS (°C)	MAE (°C)	RMSE (°C)	Width (°C)	Coverage (%)
Raw Ensemble	1.26	1.43	1.73	1.18	23.34
Noise Ensemble	1.23	1.43	1.73	1.51	29.97
Spatial BMA	0.71	1.00	1.26	3.78	86.29
ECC BMA	0.78	1.00	1.27	2.27	60.04
BMA	0.82	1.00	1.26	1.41	42.30
Spatial EMOS$^+$	0.72	1.00	1.26	4.66	92.68
ECC EMOS$^+$	0.78	1.01	1.27	2.21	58.77
EMOS$^+$	0.85	1.00	1.26	1.15	34.89
GOP Member 15	0.73	1.01	1.28	3.39	81.02

Table C.2: Multivariate assessment of the temperature at airports in Hamburg, Berlin and Frankfurt. The scores are all in degrees Celsius and are averaged over the time period from 5 January 2011 until 30 November 2011.

Model	ES	EE
Raw Ensemble	2.48	3.63
Noise Ensemble	2.61	3.67
Spatial BMA	2.12	2.99
ECC BMA	2.15	2.99
BMA	2.13	2.99
Spatial EMOS$^+$	2.13	2.99
ECC EMOS$^+$	2.10	2.96
EMOS$^+$	2.11	2.96
GOP Member 15	2.15	3.02

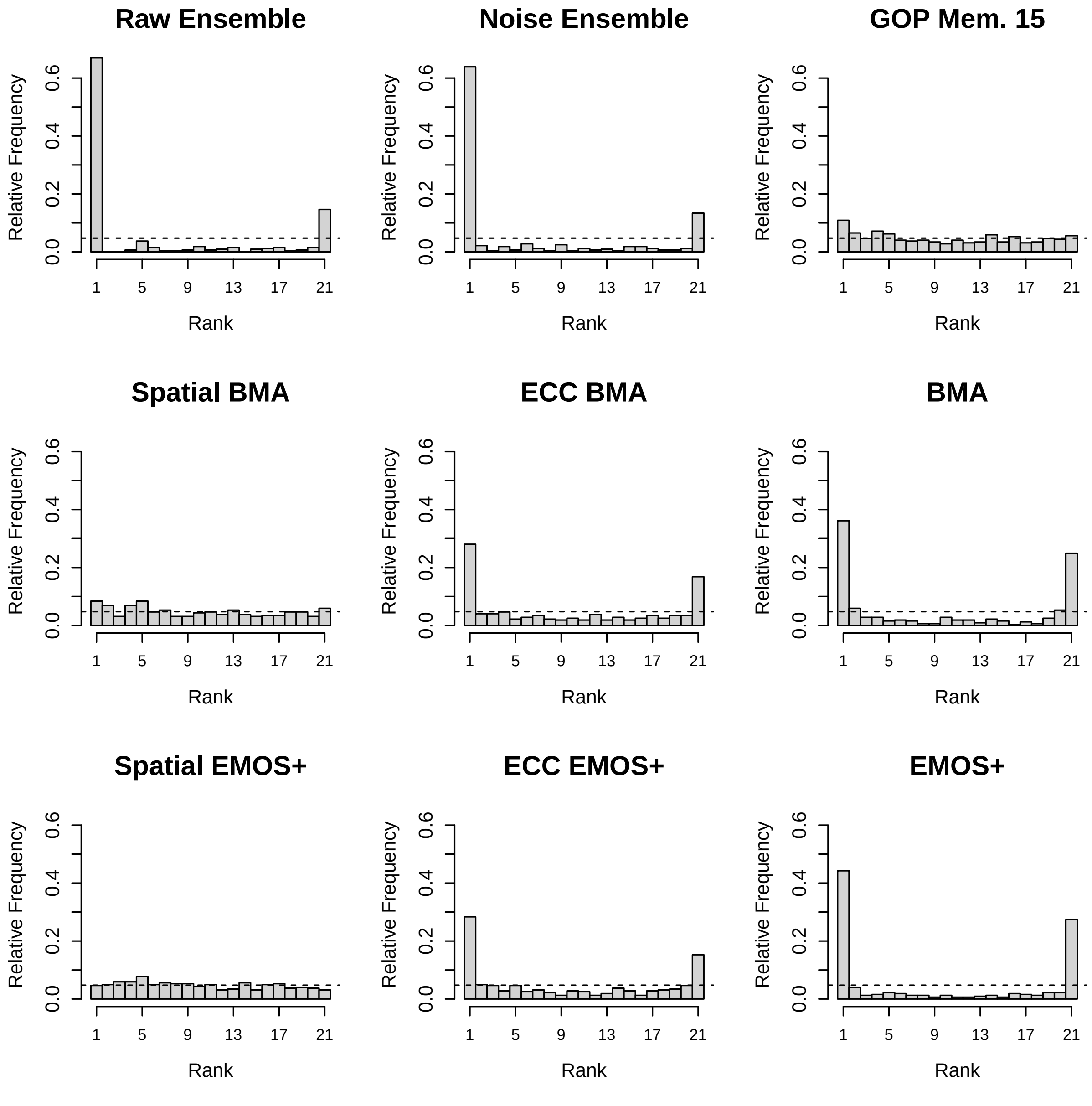

Figure C.1: Rank histograms for forecasting the average temperature in the German state Hessen. The results are aggregated over the time period from 5 January 2011 until 30 November 2011.

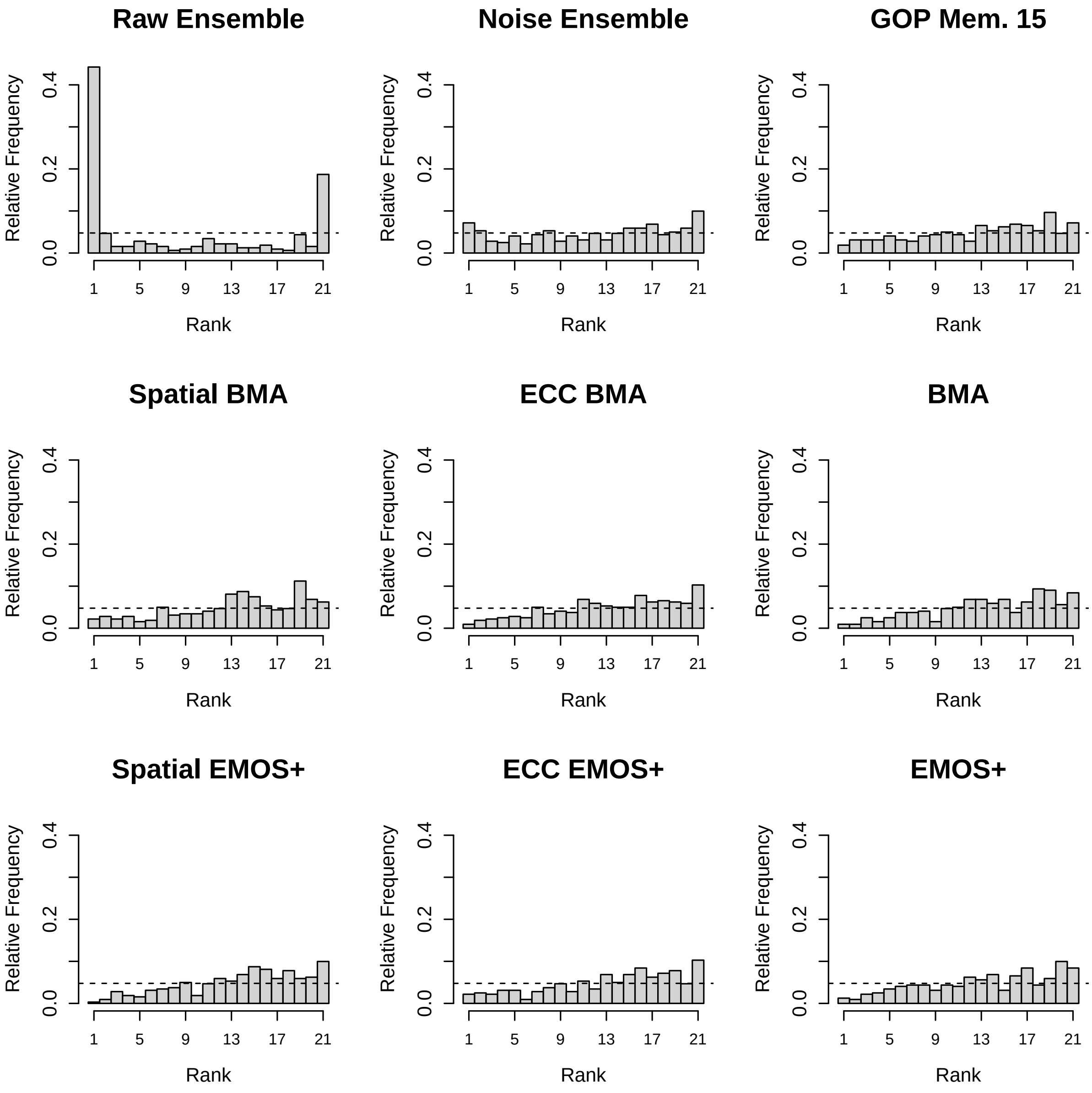

Figure C.2: Minimum spanning tree rank histograms for forecasts of surface temperature at airports in Hamburg, Berlin, and Frankfurt. The results are aggregated over the time period from 5 January 2011 until 30 November 2011.

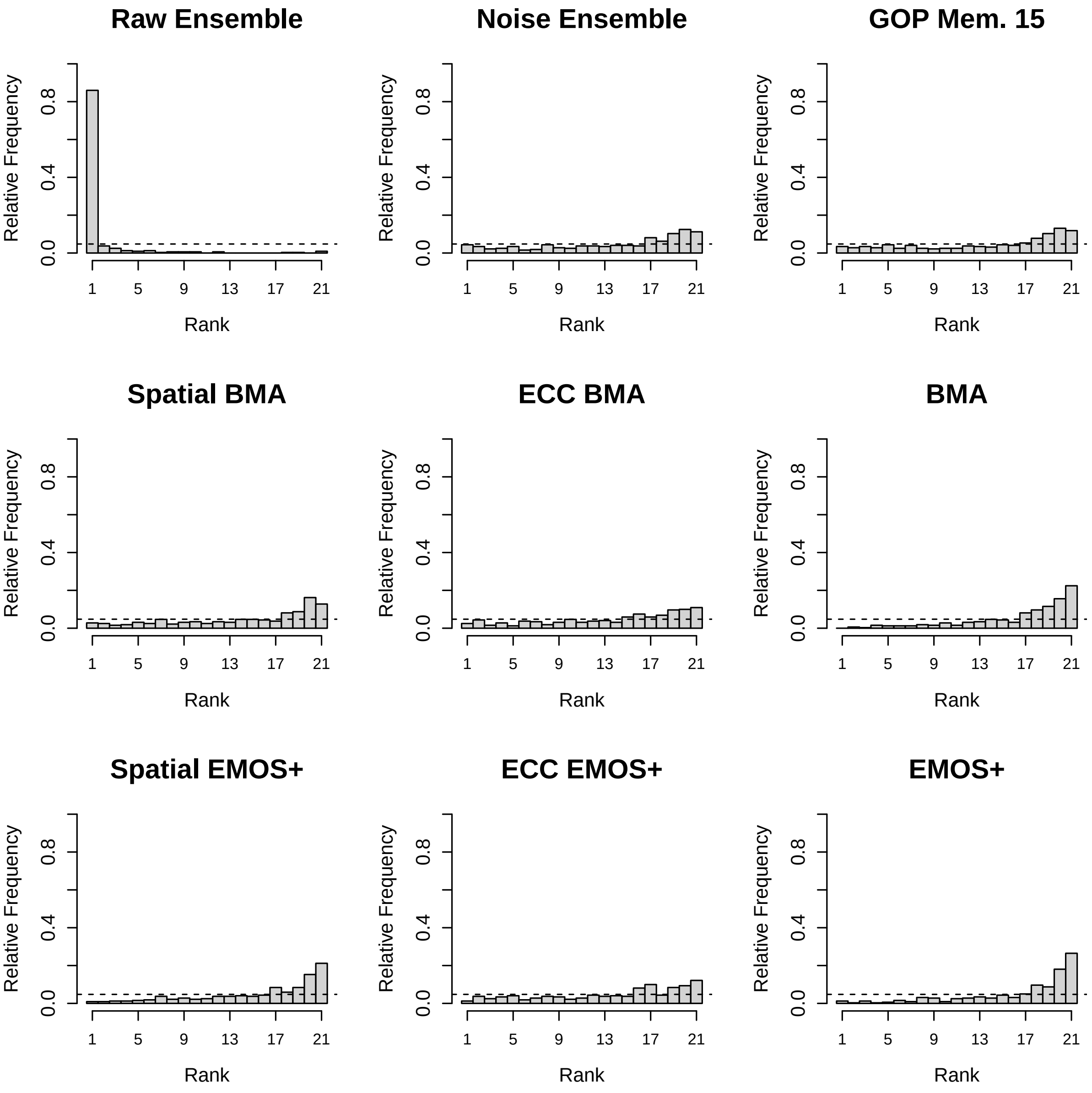

Figure C.3: Minimum spanning tree rank histograms for forecasts of temperature at three stations in Berlin. The results are aggregated over the time period from 5 January 2011 until 30 November 2011.

Table C.3: Multivariate assessment of the surface temperature at three stations in Berlin. The scores are all in degrees Celsius and are averaged over the time period from 5 January 2011 until 30 November 2011.

Model	ES	EE
Raw Ensemble	2.50	3.64
Noise Ensemble	2.66	3.68
Spatial BMA	2.01	2.82
ECC BMA	2.04	2.83
BMA	2.06	2.83
Spatial EMOS$^+$	2.00	2.79
ECC EMOS$^+$	2.02	2.80
EMOS$^+$	2.05	2.79
GOP Member 15	2.01	2.81

Table C.4: Univariate assessment results for forecasting the surface temperature in Germany: The MAE, RMSE, CRPS and the width of the nominal 19/21 ≈ 90.5% prediction interval in degrees Celsius, as well as its coverage are aggregated over all stations and the time period from 5 January 2011 until 30 November 2011.

				Prediction Intervals	
	CRPS (°C)	MAE (°C)	RMSE (°C)	Width (°C)	Coverage (%)
Raw Ensemble	1.57	1.77	2.27	1.50	25.97
Noise Ensemble	1.37	1.79	2.29	4.43	66.79
Spatial BMA	1.04	1.46	1.86	5.91	88.87
ECC BMA	1.04	1.46	1.86	5.91	88.83
BMA	1.04	1.46	1.86	5.91	88.82
Spatial EMOS$^+$	1.05	1.46	1.87	6.14	94.08
ECC EMOS$^+$	1.05	1.46	1.87	5.76	87.99
EMOS$^+$	1.04	1.46	1.87	5.76	87.99
GOP Member 15	1.06	1.48	1.89	5.38	85.20

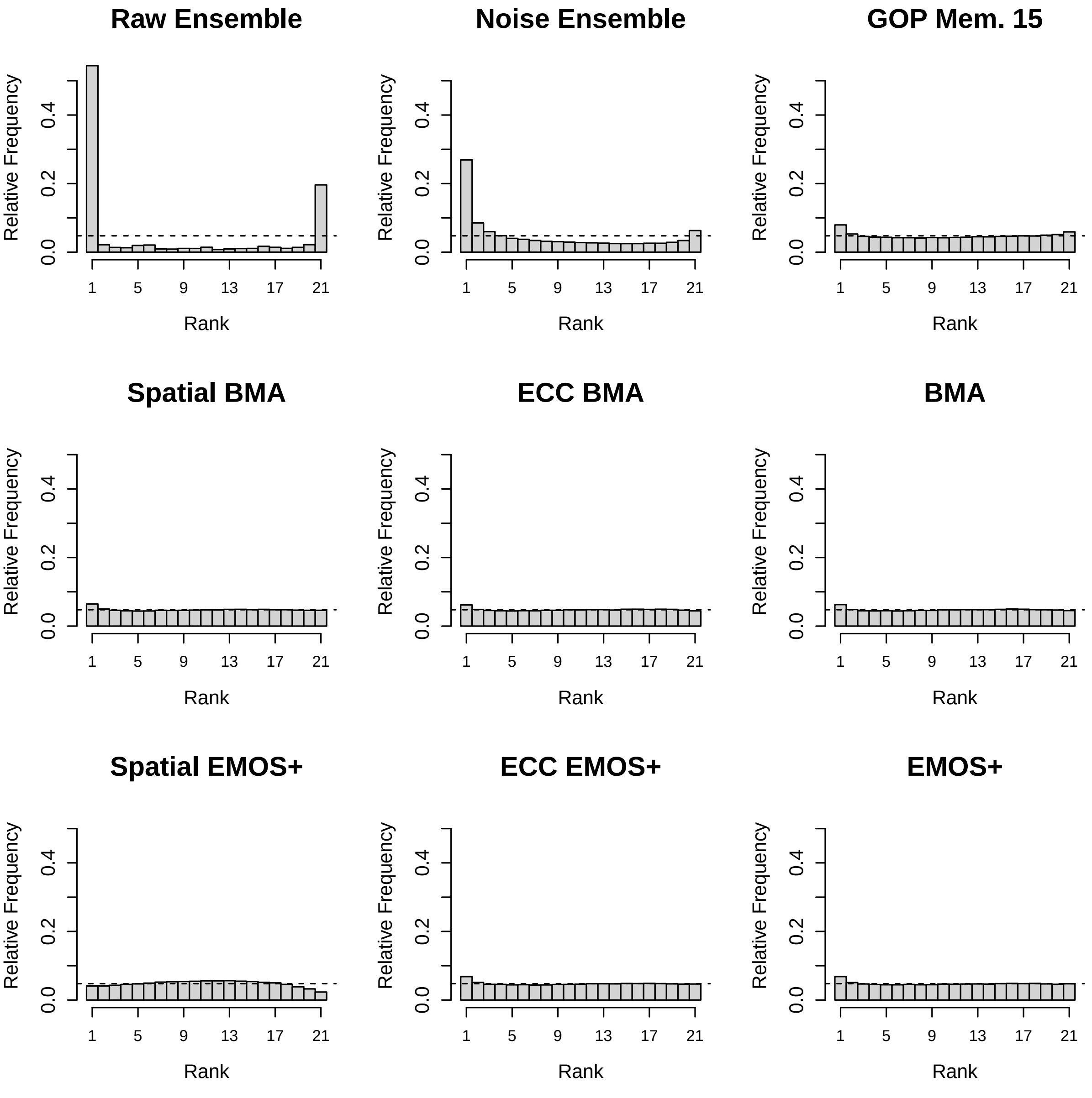

Figure C.4: Rank histograms for forecasting surface temperature in Germany. The results are aggregated over all stations and the time period from 5 January 2011 until 30 November 2011.

Bibliography

Anderson, J. L. (1996) A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations. *Journal of Climate*, **9**, 1518–1530.

Baldauf, M., A. Seifert, J. Förstner, D. Majewski, M. Raschendorfer and T. Reinhardt (2011) Operational Convective-Scale Numerical Weather Prediction with the COSMO Model: Description and Sensitivities. *Monthly Weather Review*, **139**, 3887–3905.

Bao, L., T. Gneiting, E. P. Grimit, P. Guttorp and A. E. Raftery (2010) Bias Correction and Bayesian Model Averaging for Ensemble Forecasts of Surface Wind Direction. *Monthly Weather Review*, **138**, 1811–1821.

Berrocal, V. J., Y. Gel, A. E. Raftery and T. Gneiting (2010) ProbForecastGOP: Probabilistic weather forecast using the GOP method. URL `http://cran.r-project.org/package=ProbForecastGOP`.

Berrocal, V. J., A. E. Raftery and T. Gneiting (2007) Combining Spatial Statistical and Ensemble Information in Probabilistic Weather Forecasts. *Monthly Weather Review*, **135**, 1386–1402.

Berrocal, V. J., A. E. Raftery and T. Gneiting (2008) Probabilistic Quantitative Precipitation Field Forecasting using a Two-Stage Spatial Model. *The Annals of Applied Statistics*, **2**, 1170–1193.

Bishop, C. M. (2006) *Pattern Recognition and Machine Learning.* Springer.

Brier, G. W. (1950) Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, **78**, 1–3.

Bryd, R. H., P. Lu, J. Nocedal and C. Zhu (1995) A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, **16**, 1190–1208.

Buizza, R. (1997) Potential Forecast Skill of Ensemble Prediction and Spread and Skill Distributions of the ECMWF Ensemble Prediction System. *Monthly Weather Review*, **125**, 99–119.

Cressie, N. A. C. (1985) Fitting Variogram Models by Weighted Least Squares. *Mathematical Geology*, **17**, 563–586.

Dawid, A. P. (1984) Statistical theory: The Prequential Approach (with Discussion and Rejoinder). *Journal of the Royal Statistical Society Series A*, **147**, 278–292.

Diebold, F. X., T. A. Gunther and A. S. Tay (1998) Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review*, **39**, 863–883.

Diggle, P. J. and P. J. Ribeiro Jr. (2007) *Model-based Geostatistics*. Springer.

Eckel, F. A. and F. M. Clifford (2005) Aspects of Effective Mesoscale, Short-Range Ensemble Forecasting. *Weather and Forecasting*, **20**, 328–350.

Environmental Modeling Center (2003) The GFS Atmospheric Model. *NCEP Off. Note 442*, Natl. Cent. for Environ. Protection, Camp Spring.

Fraley, C., A. E. Raftery, J. M. Sloughter, T. Gneiting and U. of Washington (2011) ensembleBMA: Probabilistic Forecasting using Ensembles and Bayesian Model Averaging. URL `http://cran.r-project.org/package=ensembleBMA`.

Gebhardt, C., S. E. Theis, M. Paulat and Z. Ben-Bouallègue (2011) Uncertainties in COSMO-DE Precipitation Forecasts Introduced by Model Perturbations and Variation of Lateral Boundaries. *Atmospheric Research*, **100**, 168–177.

Gel, Y., A. E. Raftery and T. Gneiting (2004) Calibrated Probabilistic Mesoscale Weather Field Forecasting: The Geostatistical Output Perturbation Method. *Journal of the American Statistical Association*, **99**, 575–583.

Gneiting, T. (1999) Correlation Functions for Atmospheric Data Analysis. *Quarterly Journal of the Royal Meteorological Society*, **125**, 2449–2464.

Gneiting, T. (2011) Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, **106**, 746–762.

Gneiting, T., F. Balabdaoui and A. E. Raftery (2007) Probabilistic Forecasts, Calibration and Sharpness. *Royal Statistical Society*, **69**, 243–268.

Gneiting, T., A. E. Raftery, A. H. Westveld and T. Goldman (2005) Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, **133**, 1098–1118.

Gneiting, T., L. I. Stanberry, E. P. Grimit, L. Held and N. A. Johnson (2008) Assessing Probabilistic Forecasts of Multivariate Quantities, with an Application to Ensemble Predictions of Surface Winds. *Test*, **17**, 211–235.

Gneiting, T. G. and A. E. Raftery (2007) Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, **102**, 359–378.

Gosset, W. S. (1908) The Probable Error of a Mean. *Biometrika*, **6**, 1–25.

Hamill, T. M. and S. J. Colucci (1997) Verification of Eta-RSM Short-Range Ensemble Forecasts. *Monthly Weather Review*, **125**, 1312–1327.

Hamill, T. M., J. S. Whitaker and X. Wei (2004) Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts. *Monthly Weather Review*, **132**, 1434–1447.

Hersbach, H. (2000) Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, **15**, 559–570.

Hoeting, J. A., D. Madigan, A. E. Raftery and C. T. Volinsky (1999) Bayesian Model Averaging: A Tutorial. *Statistical Science*, **14**, 382–417.

Janssen, P. and J. Bidlot (2002) ECMWF Wave Model (CY25R1). IFS Documentation Cycle CY25r1, Report, ECMWF, Reading, U.K.

Kleiber, W., A. E. Raftery, J. Baars, T. Gneiting, C. F. Mass and E. Grimit (2011) Locally Calibrated Probabilistic Temperature Forecasting Using Geostatistical Model Averaging and Local Bayesian Model Averaging. *Monthly Weather Review*, **139**, 2630–2649.

Leith, C. E. (1974) Theoretical Skill of Monte-Carlo Forecasts. *Monthly Weather Review*, **104**, 409–418.

Leutbecher, M. and T. N. Palmer (2008) Ensemble Forecasting. *Journal of Computational Physics*, **227**, 3515–3539.

Lewis, J. M. (2005) Roots of Ensemble Forecasting. *Monthly Weather Review*, **133**, 1865–1885.

Majewski, D., D. Liermann, P. Prohl, B. Ritter, M. Buchhod, T. Hanisch, P. Gerhard and W. Wergen (2002) The Operational Global Icoshedral-Hexagonal Gridpoint Model GME: Description and High-Resolution Tests. *Monthly Weather Review*, **130**, 319–338.

Matérn, B. (1986) *Spatial variation.* Springer.

Matheson, J. E. and R. L. Winkler (1976) Scoring Rules for Continuous Probability Distributions. *Management Science*, **22**, 1087–1096.

Nordhausen, K., S. Sirkia, H. Oja and D. E. Tyler (2010) ICSNP: Tools for Multivariate Nonparametrics. URL `http://cran.r-project.org/package=ICSNP`.

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens and H. Wagner (2011) vegan: Community Ecology Package. URL `http://cran.r-project.org/package=vegan`.

Peralta, C., Z. Ben Bouallègue, S. E. Theis, C. Gebhardt and M. Buchhold (2012) Accounting for Initial Condition Uncertainties in COSMO-DE-EPS. *Journal of Geophysical Research*, **117**, D07108.

Peralta, C. and M. Buchhold (2011) Initial Condition Perturbations for the COSMO-DE-EPS. *COSMO Newsletter*, **11**, 115–123.

Raftery, A. E., T. Gneiting, F. Balabdaoui and M. Polakowski (2005) Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, **133**, 1155–1174.

Schefzik, R. (2011) *Ensemble Copula Coupling.* Diploma thesis, Faculty of Mathematics and Computer Science, Heidelberg University.

Schlather, M. (2011) RandomFields: Simulation and Analysis of Random Fields. URL `http://cran.r-project.org/package=RandomFields`.

Schuhen, N., T. L. Thorarinsdottir and T. Gneiting (2012) Ensemble Model Output Statistics for Wind Vectors. *to appear in Monthly Weather Review.*

Sloughter, J. M., T. Gneiting and A. E. Raftery (2010) Probabilistic Wind Speed Forecasting Using Ensembles and Bayesian Model Averaging. *Journal of the American Statistical Association*, **105**, 25–35.

Sloughter, J. M. L., A. E. Raftery, T. Gneiting and C. Fraley (2007) Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging. *Monthly Weather Review*, **135**, 3209–3220.

Smith, L. A. (2001) Disentangling Uncertainty and Error: on the Predictability of Nonlinear Systems. *In: A. Mess (ed) Nonlinear dynamics and statistics*, 31–64, Birkhäuser, Boston.

Steppeler, J., G. Doms, U. Schättler, H. W. Bitzer, A. Gassmann, U. Damrath and G. Gregoric (2003) Meso-Gamma Scale Forecasts using the Nonhydrostatic Model LM. *Meteorology and Atmospheric Physics*, **82**, 75–96.

Talagrand, O., R. Vautard and B. Strauss (1997) Evaluation of Probabilistic Prediction Systems. *Proc. Workshop on Predictability*, 1–25, Reading, UK, European Centre for Medium-Range Weather Forecasts.

Theis, S. and C. Gebhardt (2012) Operationelles NWV-System. URL `http://www.dwd.de/bvbw/generator/DWDWWW/Content/Forschung/FE1/Aenderungen__NWV__System/COSMO__DE__EPS/PDF__2012__2014/PDF__COSMO__DE__EPS__22__05__2012,templateId=raw,property=publicationFile.pdf/PDF_COSMO_DE_EPS_22_05_2012.pdf`.

Theis, S., C. Gebhardt, M. Buchhold, Z. Ben-Bouallègue, R. Ohl, M. Paulat and C. Peralta (2011) COSMO-DE-EPS: Start of Pre-Operational Phase. URL `http://www.dwd.de/bvbw/generator/DWDWWW/Content/Forschung/FE1/Seminare__Tagungen/2011/COSMO__User2011/Presentations/06__Ensembles/Theis2011,templateId=raw,property=publicationFile.pdf/Theis2011.pdf`.

Thorarinsdottir, T. L. and T. Gneiting (2010) Probabilistic Forecasts of Wind Speed: Ensemble Model Output Statistics by using Heteroscedastic Censored Regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **173**, 371–388.

Thorarinsdottir, T. L. and M. S. Johnson (2012) Probabilistic Wind Gust Forecasting using Non-Homogeneous Gaussian Regression. *Monthly Weather Review*, **140**, 889–897.

Thorarinsdottir, T. L., M. Scheuerer and K. Feldmann (2012) Statistical Post-Processing of Ensemble Forecasts. *to appear in Promet.*

Toth, Z. and E. Kalnay (1997) Ensemble Forecasting at NCEP and the Breeding Method. *Monthly Weather Review*, **125**, 3297–3319.

Wilks, D. and T. Hamill (2007) Comparison of Ensemble-MOS Methods using GFS Reforecasts. *Monthly Weather Review*, **135**, 2379–2390.

Wilks, D. S. (2006) *Statistical Methods in the Atmospheric Sciences.* Elsevier Academic Press, Amsterdam, 2nd edition.

Zhan, D., J. Li, L. Ji, B. Huang, W. Wu, J. Chen and Z. Song (1995) A Global Spectral Model and Test of its Performance. *Advances in Atmospheric Sciences*, **12**, 67–78.

Erklärung

Hiermit versichere ich, dass ich meine Arbeit selbstständig unter Anleitung verfasst habe, dass ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, und dass ich alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entlehnt sind, durch die Angabe der Quellen als Entlehnungen kenntlich gemacht habe.

14. Juni 2012