INAUGURAL–DISSERTATION

zur

Erlangung der Doktorwürde

der

Naturwissenschaftlich–Mathematischen Gesamtfakultät

der

Ruprecht–Karls–Universität

Heidelberg

vorgelegt von

Diplom–Mathematiker Roman Schefzik

aus Mannheim

Datum der mündlichen Prüfung: .....

Physically coherent probabilistic weather forecasts using multivariate discrete copula-based ensemble postprocessing methods

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

Abstract

Being able to provide accurate forecasts of future quantities has always been a great human desire and is essential in numerous situations in daily life. Meanwhile, it has become routine to work with probabilistic forecasts in the form of full predictive distributions rather than with single deterministic point forecasts in many disciplines, with weather prediction acting as a key example.

Nowadays, probabilistic weather forecasts are usually constructed from ensemble prediction systems, which consist of multiple runs of numerical weather prediction models differing in the initial conditions and/or the parameterized numerical representation of the atmosphere. The raw ensemble forecasts typically reveal biases and dispersion errors and thus call for statistical postprocessing to realize their full potential. Several ensemble postprocessing methods have been developed and are partly recapitulated in this thesis, yet many of them only apply to a single weather quantity at a single location and for a single prediction horizon. In many applications, however, there is a critical need to account for spatial, temporal and inter-variable dependencies.

To address this, a tool called ensemble copula coupling (ECC) is introduced and examined. Essentially, ECC uses the empirical copula induced by the raw ensemble to aggregate samples from predictive distributions for each location, variable and look-ahead time separately, which are obtained via existing univariate postprocessing methods. The ECC ensemble inherits the multivariate rank dependence pattern from the raw ensemble, thereby capturing the flow dependence.

Several variants and modifications of ECC are studied, and it is demonstrated that the ECC concept provides an overarching frame for existing techniques scattered in the literature.

From a mathematical point of view, it is shown that ECC can be considered a copula approach by pointing out relationships to multivariate discrete copulas, which are introduced in this thesis and for which relevant mathematical properties are derived.

A generalization of standard ECC is introduced, which aggregates samples from not necessarily univariate, but general predictive distributions obtained by low-dimensional postprocessing in an ECC-like manner.

Finally, the SimSchaake approach, which combines the notion of similarity-based ensemble methods with that of the so-called Schaake shuffle, is presented as an alternative to ECC. In this technique, the dependence patterns are based on verifying observations rather than on raw ensemble forecasts as in ECC.

The methods and concepts are illustrated and evaluated based on case studies, using real ensemble forecast data of the European Centre for Medium-Range Weather Forecasts. Essentially, the new multivariate approaches developed in this thesis reveal good predictive performances, thus contributing to improved probabilistic forecasts.

Zusammenfassung

Es ist schon immer ein großes menschliches Bedürfnis gewesen und in zahlreichen Situationen des täglichen Lebens unabdingbar, präzise Vorhersagen zukünftiger Größen bereitstellen zu können. Mittlerweile ist es in vielen Disziplinen zur Gewohnheit geworden, mit probabilistischen Vorhersagen in der Form von vollständigen Vorhersageverteilungen zu arbeiten, und nicht mit einzelnen deterministischen Punktvorhersagen, wobei die Wettervorhersage als ein Schlüsselbeispiel fungiert.

Heutzutage werden probabilistische Wettervorhersagen üblicherweise auf der Grundlage von Ensemblevorhersagesystemen erstellt, die aus mehreren Durchläufen numerischer Wettervorhersagemodelle bestehen, welche sich hinsichtlich der Anfangsbedingungen und/oder der parameterisierten numerischen Darstellung der Atmosphäre unterscheiden. Die unbearbeiteten Ensemblevorhersagen offenbaren typischerweise systematische und Dispersionsfehler und benötigen daher eine statistische Nachbereitung, um ihr volles Leistungsvermögen zu verwirklichen. Mehrere Nachbereitungsmethoden für Ensembles sind entwickelt worden und werden in dieser Arbeit zum Teil rekapituliert, wobei viele davon jedoch nur für eine einzelne Wettergröße, an einem einzelnen Ort und für einen einzelnen Vorhersagehorizont gelten. In vielen Anwendungen besteht jedoch ein dringender Bedarf, räumliche, zeitliche und Abhängigkeiten zwischen den Größen zu berücksichtigen.

Um dies zu bewerkstelligen, wird ein Werkzeug namens Ensemble Copula Coupling (ECC) eingeführt und untersucht. Im Wesentlichen verwendet ECC die von dem unbearbeiteten Ensemble induzierte empirische Copula, um Stichproben von getrennten Vorhersageverteilungen für jeden Ort, jede Variable und jeden Vorhersagehorizont zu verbinden, die durch bestehende univariate Nachbereitungsmethoden erhalten werden. Das ECC-Ensemble übernimmt das multivariate Rangabhängigkeitsmuster des unbearbeiteten Ensembles und erfasst dadurch die Datenabhängigkeit.

Mehrere Varianten und Modifikationen von ECC werden untersucht und es wird demonstriert, dass das ECC-Konzept einen übergreifenden Rahmen für vorhandene Methoden liefert, die in der Literatur verstreut sind.

Aus mathematischer Sicht wird gezeigt, dass ECC als ein Copulaansatz angesehen werden kann, indem Beziehungen zu multivariaten diskreten Copulas aufgezeigt werden, die in dieser Arbeit eingeführt und für die relevante mathematische Eigenschaften hergeleitet werden.

Es wird eine Verallgemeinerung des standardmäßigen ECC eingeführt, die Stichproben von nicht notwendigerweise univariaten, sondern allgemeinen Vorhersageverteilungen, die durch niederdimensionale Nachbereitung erhalten werden, auf eine zu ECC ähnliche Weise verbindet.

Schließlich wird die SimSchaake-Methode, welche die Idee der auf Ähnlichkeit basierenden Ensemblemethoden mit der des sogenannten Schaake-Shuffles verbindet, als eine Alternative zu ECC vorgestellt. In diesem Verfahren basieren die Abhängigkeitsmuster auf eingetretenen Beobachtungen und nicht auf unbearbeiteten Ensemblevorhersagen wie bei ECC.

Die Methoden und Konzepte werden anhand von Fallstudien illustriert und bewertet, wobei

reale Ensemblevorhersagedaten vom Europäischen Zentrum für mittelfristige Wettervorhersage verwendet werden. Im Wesentlichen zeigen die neuen multivariaten Methoden, die in dieser Arbeit entwickelt werden, eine gute Vorhersageleistung und tragen somit zu verbesserten probabilistischen Vorhersagen bei.

Acknowledgments

First of all, I thank my advisors Tilmann Gneiting and Thordis Thorarinsdottir for their supervision of this thesis, including numerous discussions, and their dedication to induct me into the scientific sphere with all its facets.

The funding of my work by the Volkswagen Foundation within the project "Mesoscale Weather Extremes: Theory, Spatial Modeling and Prediction (WEX–MOP)" and the German Research Foundation within the Research Training Group (RTG) 1953 "Statistical Modeling of Complex Systems and Processes–Advanced Nonparametric Approaches", respectively, is gratefully acknowledged.

Moreover, the scientific support by the Faculty of Mathematics and Computer Science of Heidelberg University and the doctoral training within the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences are appreciated.

Further, I thank the members in the WEX–MOP project and the RTG 1953 as well as my colleagues at the Institute for Applied Mathematics of Heidelberg University and the Heidelberg Institute for Theoretical Studies, respectively, for creating an enjoyable working atmosphere and for fruitful discussions. I am grateful to Michael Scheuerer and Thordis Thorarinsdottir for providing the forecast and observation data employed for the case studies in this thesis, and further to Elena Scardovi, Michael Scheuerer, Nina Schuhen and Thordis Thorarinsdottir for sharing and providing R code. Besides, I thank Michael Denhard, Martin Leutbecher, Annette Möller, Elena Scardovi and the reviewers of my working papers for useful discussions, comments and hints, and Martin Leutbecher in particular for detecting an error in an earlier version of the data set used in this thesis. Moreover, the support of Marion Münster, Dagmar Neubauer and Milanka Stojkovic in administrative issues is acknowledged.

Last but not least, I am indebted to my parents and family, my friends and my club mates for their support and for providing distraction from science in my leisure time.

Contents

1 Introduction **1**
- 1.1 Uncertainty quantification and probabilistic forecasting . . . 1
- 1.2 Weather prediction and ensemble postprocessing . . . 2
- 1.3 The European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble . . . 7

2 Forecast verification methods **11**
- 2.1 Calibration . . . 12
- 2.2 Sharpness . . . 14
- 2.3 Proper scoring rules . . . 15

3 Statistical ensemble postprocessing **21**
- 3.1 Univariate postprocessing . . . 21
  - 3.1.1 Bayesian model averaging (BMA) . . . 21
  - 3.1.2 Ensemble model output statistics (EMOS) . . . 25
  - 3.1.3 Case study . . . 28
- 3.2 Multivariate postprocessing . . . 33
  - 3.2.1 Dependence modeling via copulas . . . 33
  - 3.2.2 Examples of multivariate postprocessing methods . . . 35

4 Ensemble copula coupling (ECC) **39**
- 4.1 The ensemble copula coupling (ECC) approach . . . 39
- 4.2 The quantization step . . . 49
  - 4.2.1 The sampling methods (R), (T) and (Q) . . . 49
  - 4.2.2 Case study . . . 51
- 4.3 The predictive performance of ECC: Case studies . . . 64
  - 4.3.1 Implementation and reference ensembles . . . 64
  - 4.3.2 Spatial aspects . . . 66
  - 4.3.3 Inter-variable aspects . . . 80
  - 4.3.4 Joint spatial and inter-variable aspects . . . 83
  - 4.3.5 Temporal aspects . . . 86
  - 4.3.6 Conclusions . . . 89
- 4.4 ECC variants when the desired ensemble size after postprocessing exceeds that of the raw ensemble . . . 90
  - 4.4.1 The extended ECC approach . . . 90
  - 4.4.2 Case study . . . 95
  - 4.4.3 Alternative approaches . . . 100
- 4.5 ECC for ensembles with non-exchangeable members . . . 103

5 ECC as an overarching theme **105**
5.1 A member-by-member postprocessing (MBMP) method as an ECC variant . 105
5.1.1 The MBMP approach 105
5.1.2 Case study 107
5.2 Other examples of ECC variants in the extant literature 112
5.2.1 Pinson (2012) 113
5.2.2 Roulin and Vannitsem (2012) 114

6 Multivariate discrete copulas: The theoretical frame **117**
6.1 Multivariate discrete copulas 118
6.2 A characterization of multivariate discrete copulas using stochastic arrays . . 120
6.3 A multivariate discrete version of Sklar's theorem 125
6.4 ECC and the Schaake shuffle as multivariate discrete copula approaches . . . 131

7 Combining low-dimensional postprocessing methods in an ECC-like manner **137**
7.1 Multivariate quantiles 137
7.2 Combining low-dimensional postprocessing methods in an ECC-like manner: The LDP-ECC approach 140
7.3 Case study 148

8 Combining similarity-based ensemble methods with the Schaake shuffle **153**
8.1 Combining similarity-based ensemble methods with the Schaake shuffle: The SimSchaake method 153
8.2 Case study 157

9 Summary and discussion **165**

References **171**

List of Figures **185**

List of Tables **191**

Chapter 1

Introduction

1.1 Uncertainty quantification and probabilistic forecasting

In many applications, decision making relies on potentially high-dimensional output of complex computer simulation models. For instance, this applies to weather and climate predictions and the management of air quality, wildfires, floods, groundwater contamination or disease spread – just to name a few examples. During the last years, the recognition of the need for uncertainty quantification of such model output has been rising considerably, which is for example witnessed by the foundation of interest groups on this topic within the American Statistical Association (ASA) and the Society for Industrial and Applied Mathematics (SIAM), as well as the launch of the SIAM/ASA Journal on Uncertainty Quantification in 2013.

Frequently, the output data are employed to make predictions for uncertain future quantities or events, which has always been a great human desire. Initially, forecasting had been viewed as a purely deterministic issue, in that a prediction used to be a single number. Such point forecasts are partly still issued today, be it for reasons of tradition, reporting requirements or market mechanisms, for instance, and are also of theoretical interest (Gneiting, 2011a,b). However, it is meanwhile clearly established that forecasts should be probabilistic in nature (Dawid, 1984), having the form of full predictive probability distributions over future quantities or events instead of single-valued point forecasts. That is, in place of stating twenty degrees Celsius (°C) as a point forecast for temperature at noon in Heidelberg on a spring day, one should rather issue a predictive distribution, for instance a normal distribution with a mean of twenty degrees Celsius and a standard deviation of one degree Celsius. If ever, probabilistic forecasts in former times were made almost only for binary events (Gigerenzer et al., 2005), such as the chance of rain at noon in Heidelberg on a certain day. Nowadays, also probabilistic forecasts for multi-category or continuous variables are of great importance and are required in a vast range of scientific disciplines comprising weather and climate, hydrology, economics and finance, politics, preventative medicine and epidemiology, among others (Gneiting and Katzfuss, 2014, and references therein).

The aim of probabilistic forecasting is to create predictive distributions of future quantities, from which relevant functionals such as moments, quantiles, prediction intervals or event probabilities can be extracted to quantify the uncertainty of the prediction. In this connection, the concepts of sharpness and calibration play an essential role (Gneiting et al., 2007). Sharpness is a property of the probabilistic forecasts only and concerns the the con-

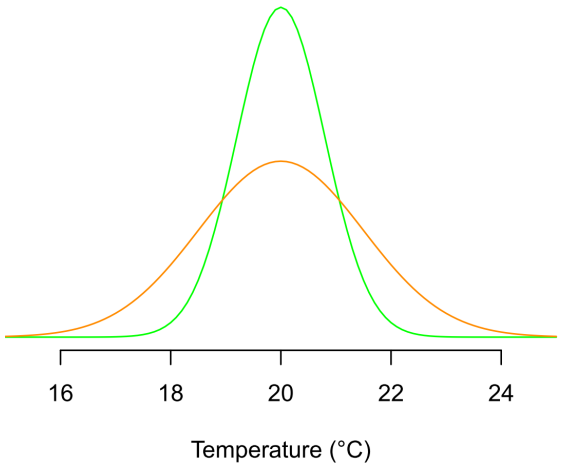


Figure 1.1: Illustration of the goal of probabilistic forecasting: Normal PDFs for temperature in Heidelberg at noon on a spring day made by Forecaster 1 (green curve) and Forecaster 2 (orange curve), respectively. Provided that both predictive distributions are calibrated, the sharper one by Forecaster 1 should be preferred.

centration of the predictive distributions. Calibration, on the other hand, is a joint property of the probabilistic forecasts and the observations, referring to the statistical compatibility between them, in the sense that events predicted to occur with probability p should materialize with empirical frequency p. Gneiting et al. (2007) contend that predictive distributions should be as sharp as possible, subject to calibration. For instance, let the green and orange curves in Figure 1.1 be normal predictive probability density functions (PDFs) for temperature in Heidelberg at noon on a spring day made by Forecaster 1 (green curve) and Forecaster 2 (orange curve), respectively. In this case, the two predictive distributions have the same location, but that issued by Forecaster 1 is sharper, as it obviously reflects a lower degree of uncertainty than that issued by Forecaster 2. The predictive distribution according to Forecaster 1 should thus be preferred, provided that both forecast distributions are calibrated.

From now on, we concentrate on weather prediction as a key application area of probabilistic forecasting for the rest of the thesis. However, it should always be kept in mind that the presented concepts and methods may also apply in much broader settings in which uncertainty is to be quantified.

1.2 Weather prediction and ensemble postprocessing

Accurate predictions for future weather quantities or events are of crucial interest in our society for various reasons. For instance, they become valuable when warnings about extreme events or natural catastrophes such as inundations, storms or droughts are sought. In times in which alternative energy sources become more and more important, they may also affect decisions regarding energy generation through solar technology or wind power plants. Finally, weather forecasts may simply help to facilitate the organization of one's leisure time activities.

Modern weather forecasting originates in the pioneering work of Bjerknes (1904) at the

beginning of the 20th century. Essentially, Bjerknes (1904) proposed the possibility of numerical weather prediction (NWP), stating that the physics of the atmosphere at any point in time can be determined based on seven equations in seven parameters, namely the three hydrodynamic equations of motion (conservation of momentum), the continuity equation (conservation of mass during motion), the equation of state for the atmosphere and the two fundamental laws of thermodynamics (conservation of energy and entropy), comprising three velocity components, density, pressure, temperature and humidity. From a present-day perspective, Bjerknes actually should have rather issued a continuity equation for water than the second thermodynamic law (Lynch, 2008). According to Bjerknes (1904), solving the equation of state eliminates one of the seven unknowns. The remaining equations then build a system of six partial differential equations in six variables, where the initial conditions are set by the observations of the initial state of the atmosphere. As Bjerknes himself did not put his procedure into practice, it was Richardson who derived by hand the first weather prediction more than a decade later, employing a finite differences approach (Richardson, 1922) to simplify the equations in Bjerknes (1904). Unfortunately, his attempts were highly unsatisfactory, both with respect to the totally unrealistic forecast values themselves and the extraordinarily high calculation time needed. However, the advent of the computer sparked hopes to lead Richardson's preparatory work to success, with von Neumann demanding to use computers for weather prediction in 1946. Finally, the first weather forecast made by a computer was issued by the Electronic Numerical Integrator and Computer (ENIAC) of the United States Army in 1950. The first operational forecasts, that is, routine predictions for practical use, were generated by Rossby's group in Sweden in 1954 (Harper et al., 2007). From then on, work on NWP models has continuously intensified, in that new atmospheric models have been introduced and the size of the initial data sets has grown, to take advantage of the increasing computer power in the second half of the 20th century. Moreover, data assimilation systems, which supply the initial conditions describing the current state of the atmosphere on a three-dimensional grid, have become more powerful. For a more detailed overview of the development and the history of NWP, we refer to Harper et al. (2007) and Lynch (2008).

Nowadays, NWP models are still based on a system of partial differential equations, which are discretized and run forward in time to achieve deterministic forecasts of future atmospheric states, and form the basis of modern weather forecasting. However, they exhibit two major sources of uncertainty:

1. The initial conditions might be inaccurate due to incomplete observation data, deficiencies in data assimilation or measurement errors, for example.

2. The model formulation might be inaccurate due to incomplete or inadequate numerical schemes or imperfect knowledge of physical processes including inaccurate parameterizations of sub grid-scale processes, for instance.

Hence, there is an obvious need for uncertainty quantification in weather forecasts. This had already been recognized at the beginning of the 20th century by Cooke (1906, page 23), stating that

> "All those whose duty is to issue regular daily forecasts know that there are times when they feel very confident and other times when they are doubtful as to the coming weather. It seems to me that the condition of confidence or otherwise forms a very important part of the prediction, and ought to find expression."

Lorenz (1963) pointed out the non-linear, chaotic nature of the equations involved in NWP models, in that extremely small deviations in initial inputs might lead to largely differing evolutions of the model – an observation which has later become famous as the "butterfly effect" (Lorenz, 1993). Thus, it becomes impossible to definitely predict the state of the atmosphere. Epstein (1969) stated that it is not appropriate to describe the atmosphere via only a single forecast run and introduced an ensemble of stochastic Monte Carlo simulations to generate means and variances for the atmospheric state, which can be viewed as an early example of probabilistic weather forecasting.

Nevertheless, weather prediction had been considered a deterministic issue through the 1980s, with the idea that for a set of "best" input data, the NWP model leads to one "best" deterministic weather forecast (Gneiting and Raftery, 2005). However, in the early 1990s, a radical change of mind took place in the meteorological community, in that weather forecasts for future quantities or events are now preferred to take the form of full predictive probability distributions rather than single-valued deterministic point forecasts.

The most convenient way to achieve a probabilistic weather prediction is based on so-called ensemble prediction systems of NWP forecasts (Palmer, 2002; Gneiting and Raftery, 2005). An ensemble consists of multiple runs of NWP models differing in the initial conditions and/or the model formulation with respect to the parameterized numerical representation of the atmosphere, thereby addressing the above-mentioned two major sources of uncertainty. Combinations of ensemble member forecasts frequently show more accuracy than any of these forecasts separately (Palmer, 2002). Interpreting ensemble forecasts as a sample from the predictive distribution allows weather forecasts to become probabilistic.

Ensemble prediction systems have been employed operationally since 1992 and can be run either globally or over limited areas. Examples for global ensembles include those run by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Molteni et al., 1996; Buizza, 2006; Leutbecher and Palmer, 2008; ECMWF Directorate, 2012) and the National Centers for Environmental Predictions (NCEP) (Toth and Kalnay, 1997), respectively, while the University of Washington Mesoscale Ensemble (UWME) (Eckel and Mass, 2005) and the COSMO-DE ensemble of the German Weather Service (Gebhardt et al., 2011) are representatives of limited area systems. Moreover, there are single-model and multi-model ensemble prediction systems. A single-model ensemble is based on one particular NWP model, and the different ensemble forecasts are obtained by perturbing the initial conditions and parameterizations. On the contrary, a multi-model ensemble is an aggregation of single-model ensembles, as for instance in the THORPEX Interactive Grand Global Ensemble (TIGGE) database (Bougeault et al., 2010) provided by leading weather centers.

Ensemble members can be regarded as exchangeable (Bröcker and Kantz, 2011) if they differ in random perturbations only, such that they lack individually distinguishable physical features and are statistically indistinguishable. For example, ensembles with exchangeable members can be generated by using bred vectors as in the NCEP ensemble, singular vectors as in the ECMWF ensemble, or ensemble Kalman filter systems (Evensen, 1994; Hamill, 2006). On the other hand, an ensemble prediction system is considered to consist of non-exchangeable members if the NWP inputs or the model parameterizations differ in a systematic rather than a random fashion. An example for an ensemble system with non-exchangeable members is the COSMO-DE ensemble run by the German Weather Service.

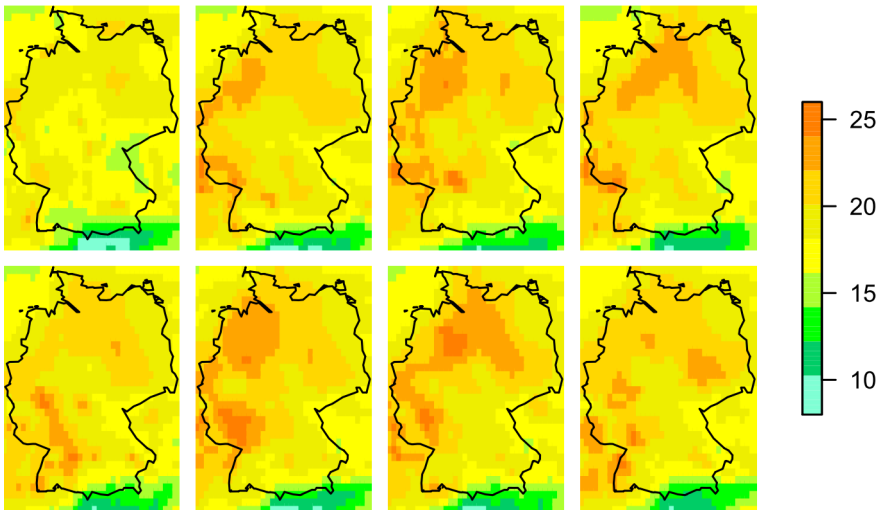


Figure 1.2: 24 hour ahead ensemble forecast for temperature (in °C) over Germany, valid 2:00 am on 3 July 2010. Eight randomly selected members of the ECMWF ensemble are shown.

An illustration of a 24 hour ahead ensemble forecast for temperature over Germany, valid at 2:00 am on 3 July 2010, is given in Figure 1.2, where eight out of the 50 exchangeable members of the ECMWF ensemble are shown. The ECMWF is one of the leading NWP centers worldwide, and its global 50-member ensemble prediction system (Molteni et al., 1996; Buizza, 2006; Leutbecher and Palmer, 2008; ECMWF Directorate, 2012), which will be described extensively in the next section, has been operational since 1992.

The aim of NWP ensemble systems is to address the inherent uncertainty in the prediction. Ensemble forecasts often exhibit a spread-error association, in that a positive correlation between the ensemble range and the forecast error can be observed. The ensemble spread offers an estimate of the uncertainty of the forecast. While on some days, the ensemble spread might be small and the atmosphere thus rather predictable, the ensemble forecasts might diverge drastically on other days, indicating an extremely unpredictable atmosphere. Despite their benefits, the raw NWP ensemble forecasts however tend to reveal model biases and dispersion errors (Hamill and Colucci, 1997). While a model bias refers to systematic errors in the NWP forecast, dispersion errors essentially mean a lack of calibration, in that the observed values fall far too often outside the ensemble ranges, contrary to the desired statistical compatibility between observations and forecasts. An example is given in Figure 1.3, where the 24 hour ahead 50-member ECMWF ensemble forecasts (red dots) for temperature at Hamburg along with the corresponding verifying observations (blue crosses) are shown for the period from 1 April 2011 to 14 April 2011, valid at 2:00 am each day.

To cope with biases and lack of calibration, the NWP raw ensemble forecasts call for statistical postprocessing, which targets at creating calibrated and sharp predictive probability distributions. During the last decade, several univariate ensemble postprocessing methods have been proposed, with Bayesian model averaging (BMA) (Raftery et al., 2005, for instance) and ensemble model output statistics (EMOS) (Gneiting et al., 2005, for example), which is also known as non-homogeneous regression, being two of the most prominent ones. The BMA approach employs mixture distributions, where each ensemble member is linked

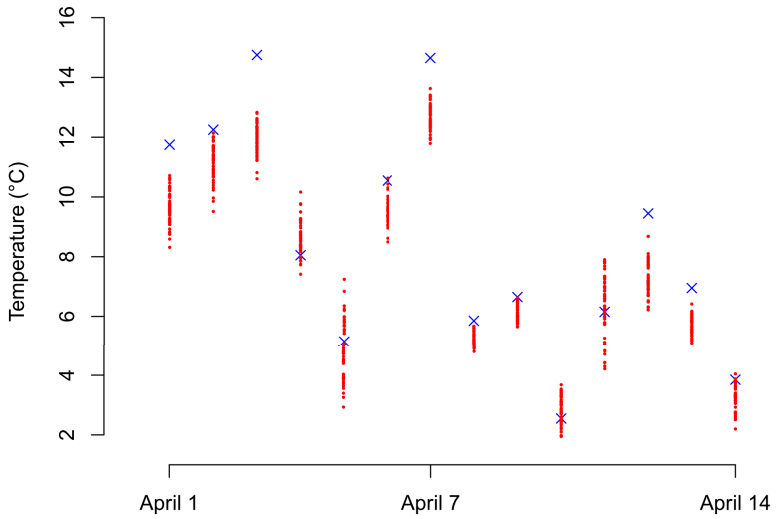


Figure 1.3: 24 hour ahead 50-member ECMWF ensemble forecasts for temperature at Hamburg (red dots) and corresponding verifying observations (blue crosses), valid 2:00 am for the period from 1 April 2011 to 14 April 2011.

to a kernel function, using a weight which reflects the member's relative skill. The EMOS method fits a single, parametric PDF in a regression setting, using summary statistics from the ensemble. Figure 1.4 shows an example of a 24 hour ahead EMOS predictive PDF for temperature at Hamburg, valid at 2:00 am on 1 April 2011. The vertical red lines indicate the 50 ECMWF raw ensemble values, the vertical line of blue crosses the verifying observation, and the vertical black lines the 10th, 50th and 90th percentiles, respectively, of the EMOS distribution. The EMOS predictive distribution corrects both a negative bias and underdispersion. BMA and EMOS postprocessing techniques have been developed for various weather variables and turn out to be well performing, but they mostly apply to a single weather quantity at a single location and a single prediction horizon only.

This is very unfortunate, as in many applications such as flood management, air traffic control, ship routing or winter road maintenance, it is crucially relevant to account for spatial, inter-variable and temporal dependence structures, which cannot be handled by independently postprocessed forecasts. Hence, there is a critical need for multivariate methods that provide physically realistic and coherent probabilistic forecasts for multiple locations, weather quantities and prediction horizons simultaneously. Consequently, much effort has been invested to address this challenge in the last years, and several ensemble postprocessing approaches being able to account for multivariate dependence structures have emerged. Examples include the Spatial BMA approach of Berrocal et al. (2007) and the Spatial EMOS method of Feldmann et al. (2014) for purely spatial settings dealing with temperature and the Gaussian copula method of Möller et al. (2013) to handle inter-variable dependencies. The techniques of Pinson (2012), Schuhen et al. (2012) and Sloughter et al. (2013), respectively, aim particularly at the postprocessing of wind vectors. These methods are parametric and perform well in low-dimensional situations and if specific structure can be exploited.

Frequently, statistical postprocessing of a full NWP ensemble forecast forms a very high-dimensional challenge. For example, one might be confronted with five weather quantities at 500×500 grid boxes and ten vertical levels for 72 lead times, yielding a total of 900 million

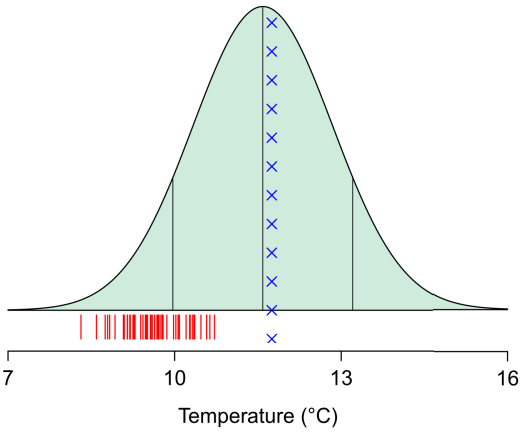


Figure 1.4: 24 hour ahead EMOS predictive PDF for temperature at Hamburg, valid 2:00 am on 1 April 2011. The vertical red lines indicate the 50 ECMWF raw ensemble values, the vertical line of blue crosses the verifying observation, and the vertical black lines the 10th, 50th and 90th percentiles, respectively, of the EMOS distribution.

variables. Even if not all of them might need to be treated simultaneously, many applications comprise much higher dimensions than can be adequately treated by parametric models. Hence, the use of non-parametric techniques, such as the Schaake shuffle (Clark et al., 2004), appears to be most beneficial in high-dimensional settings. Against this background, a multi-stage non-parametric procedure called ensemble copula coupling (ECC) (Schefzik et al., 2013) to handle spatial, inter-variable and temporal dependencies is proposed in this thesis. The key idea is that the postprocessed ECC forecast ensemble inherits the spatial, temporal and inter-variable correlation pattern of the unprocessed raw ensemble, thereby honoring the flow dependence. Basically, this is achieved by using the empirical copula of the raw ensemble to aggregate samples from predictive distributions obtained by univariate postprocessing techniques, which is why ECC can be considered a discrete copula-based approach. ECC requires negligible computational effort once the univariate postprocessing is done, and in concert with its theoretical discrete copula background, it turns out to form an overarching frame for existing techniques scattered in the literature. Moreover, the ECC notion appears to be particularly appealing as it combines analytic, numerical and statistical modeling and can be applied not only to weather prediction, but also in broader settings, in which uncertainty quantification is required. In this thesis, we discuss several variants and modifications of ECC and relate ECC to the Schaake shuffle (Clark et al., 2004).

To illustrate and test our methods, we apply them to the ECMWF ensemble in real-data case studies. For that reason, a detailed description of the ECMWF ensemble dataset as employed in this thesis is provided in the following section.

1.3 The European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble

The European Centre for Medium-Range Weather Forecasts (ECMWF) (`http://www.ecmwf.int`) is one of the world's leading weather centers. It it an independent and intergovernmental organization, which is supported by various European member states and co-operating states, with its headquarters being located in Reading near London in the United Kingdom.

Having been established in 1975, the ECMWF's main objectives comprise the design of numerical methods for medium-range weather forecasting, the distribution of the forecasts to the member states, the conduction of research to improve the forecasts, as well as the collection and storage of weather data, actually providing the world's largest archive of NWP data (Woods, 2006).

Since 1992, the ECMWF has been running an ensemble prediction system operationally. The global ECMWF ensemble consists of 50 members and operates at a horizontal resolution of approximately 32 kilometers on a 0.25×0.25 degree grid, with lead times up to ten days ahead (Molteni et al., 1996; Buizza, 2006; Leutbecher and Palmer, 2008; ECMWF Directorate, 2012). It produces forecasts twice a day at 00:00 Universal Coordinated Time (UTC) and 12:00 UTC, respectively. The 50 ensemble members can be considered exchangeable, as differences between them arise from random perturbations in initial conditions and stochastic model parameterizations. More precisely, the random perturbations in the initial conditions are generated by using singular vectors, while the model uncertainties are represented by stochastically perturbed parameterization tendencies (Buizza et al., 1999; Palmer et al., 2009). Additionally, the ECMWF provides a so-called control run, which is a distinguished NWP run outside the 50-member ensemble, representing the single-valued best estimate of the atmospheric state at the initialization time due to recent and concurrent observations. Essentially, the 50 ECMWF ensemble member forecasts start from slightly different states being close, but not identical, to the best guess of the initial atmospheric state offered by the control run. In some former case studies in the literature, the control run has been included in the ECMWF ensemble, then comprising 51 members. However, we do not proceed so in this thesis and stick to the 50-member ensemble, partly employing the control run as a reference data set as explained later.

In this thesis, we employ the ECMWF ensemble forecasts initialized at 00:00 UTC only, and we confine ourselves to prediction horizons of 24, 48, 72 and 96 hours, respectively. As we focus on forecasts over Germany, the meteorological format of 00:00 UTC corresponds to local times of 1:00 am in Central European Time and 2:00 am in Central European Summer Time, respectively. The weather variables which will be investigated involve temperature, pressure, precipitation and wind vectors. In this context, a wind vector can be either represented by wind speed and wind direction or equivalently by its u (zonal or west-east) and v (meridional or north-south) velocity components. We use the latter representation in our case studies and refer to u- and v-wind, respectively, in what follows. For pressure, precipitation and (u, v)-wind vectors, we use the corresponding ECMWF ensemble forecasts initialized between 1 February 2010 and 30 April 2011 to build our database. In case of temperature, ECMWF ensemble forecast data were available for initializations from 1 February 2010 to 31 December 2012.

Mostly, we focus on the predictive performance of forecasts at the three international airports at Berlin-Tegel, Hamburg-Fuhlsbüttel and Frankfurt am Main. As the ECMWF ensemble forecasts are available on a grid, they need to be bilinearly interpolated to the locations at Berlin, Hamburg and Frankfurt, respectively, before comparing them with the verifying observations, which are directly measured at the three specific observation sites. However, in some cases, we are also interested in the predictive performance on the whole model grid or at least on test regions over the grid. As there are no verifying observations available for all grid points involved, we use the control run for the prediction horizon of 0 hours initialized

at the target date as a grid-based ground truth instead, both as training data and for the assessment of predictive performance. For example, 48 hour ahead ensemble forecasts initialized at 00:00 UTC on 1 May 2010 and thus valid at 00:00 UTC on 3 May 2010 would be compared to the ground truth composed of the 0 hour ahead control run nowcast initialized at 00:00 UTC on 3 May 2010, where the term "nowcast" is generally employed for short-term weather predictions with look-ahead times from zero to six hours. As an alternative to the control runs used in this thesis, so-called analyses could be employed as a grid-based verification data set and ground truth, respectively, as is described in Box A in Hagedorn (2010).

Principally, we use the one-year test period from 1 May 2010 to 30 April 2011 to assess our methods in case studies and employ forecasts and observations prior to 1 May 2010 as training data. When only temperature is involved, and no comparison to other weather variables is intended, we occasionally take advantage of the corresponding larger database and evaluate our approaches over a longer test period. Our real-data case studies should in fact be rather viewed as illustrations or proof-of-concepts, respectively, as we use comparably small test periods and consider forecasts for few locations, weather variables and look-ahead times, which may not suffice for conclusive statements about predictive performance.

The rest of the thesis is organized as follows. In Chapter 2, we give an overview of uni- and multivariate verification methods to assess predictions. Essentially, those tools are presented that are employed later to evaluate the predictive skill in our case studies. Chapter 3 reviews uni- and multivariate statistical ensemble postprocessing methods, with the focus on techniques that are either directly used or needed for comparative reasons in the course of the thesis. The pivotal ensemble copula coupling (ECC) approach, including several variants and modifications, is then introduced in Chapter 4. Chapter 5 shows to what extent ECC can be interpreted as an overarching frame for existing techniques in the literature, whereas the mathematical background of the ECC method in the form of discrete copulas is discussed in Chapter 6. Chapter 7 aims at combining low-dimensional ensemble postprocessing methods in an ECC-like manner. In Chapter 8, an alternative approach to ECC based on a combination of similarity-based ensemble methods and the Schaake shuffle is described. Finally, the dissertation closes with a summary and a discussion of the essential results, as well as an outlook on possible future work, in Chapter 9. The proposed methods will be accompanied by illustrative case studies using the ECMWF ensemble throughout the whole thesis.

A very first version of the ECC approach from Section 4.1, together with initial case studies, had already been presented in the diploma thesis of Schefzik (2011). The ECC notion will be discussed in much more detail in this thesis here, with much more comprehensive case studies. Similarly, the origin of Chapter 6 about discrete copulas lies in Schefzik (2011), but again, the presentation in this thesis will be more detailed, providing new insights, with a focus on general multivariate settings.

This dissertation is partly based on two research papers. Specifically, the already published paper of Schefzik et al. (2013) forms the basis of Chapter 4, while the results of Chapter 6 stem from the working paper of Schefzik (2013), which is available online.

Chapter 2

Forecast verification methods

In this chapter, we address the question how to evaluate predictions. As already hinted at in the introductory chapter, forecasts essentially can either be issued as single-valued point predictions or as ensemble predictions or as probabilistic forecasts taking the form of a full predictive probability density function (PDF) or cumulative distribution function (CDF), respectively. In this context, an M-member ensemble forecast $x_1, \ldots, x_M \in \mathbb{R}$ can also be interpreted as a probabilistic forecast for a real-valued quantity in form of a discrete sample, and each sample value can be regarded as an equally likely potential realization of the future outcome. Then, the ensemble forecast $x_1, \ldots, x_M$ can be identified with its empirical CDF F given by

$$F(z) := \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\{x_m \leq z\}}$$

for $z \in \mathbb{R}$, with $\mathbb{1}_A$ denoting the indicator function of the event A.

The different forecast formats can be converted to some extent into each other. Ensemble predictions or predictive densities can be transformed into point forecasts if necessary or desired by extracting relevant functionals from the corresponding distributions, such as the mean or the median. Moreover, we can always sample from a predictive PDF to obtain an ensemble forecast, and conversely, an ensemble forecast could be replaced by a density estimate, especially if the ensemble size is rather large. Hence, the distinction between ensemble predictions and full forecast PDFs arguably becomes artificial to a certain degree (Gneiting et al., 2008). Other forecast formats, such as interval predictions, are generally feasible, but are not discussed in this thesis.

In the case of point predictions (Gneiting, 2011a), forecast quality is usually measured on the basis of accuracy and association (Fricker et al., 2013). Accuracy describes the correspondence between a forecast $x \in \mathbb{R}$ and an observation $y \in \mathbb{R}$ and is often quantified by some function of the error magnitude. Typical examples of such functions include the absolute error (AE) given by $\mathrm{AE} := |x - y|$ and the squared error (SE) given by $\mathrm{SE} := (x - y)^2$. In contrast, association measures the extent of a given relationship between forecasts and observations. For instance, Pearson's correlation coefficient can be employed to quantify the strength of a linear relationship. As this thesis focuses on ensemble and probabilistic forecasts, we do not provide more detail about point forecasts, although relationships might be discussed occasionally.

For ensemble forecasts and predictive PDFs, the two most important quality features are

calibration, which is sometimes also referred to as reliability, and sharpness. Calibration is a joint property of the predictions and the observations, basically relating to the statistical compatibility between them, whereas sharpness describes the concentration of the predictive distributions, thus being a property of the forecasts only. The goal of probabilistic forecasting is to maximize the sharpness of the predictive distributions, subject to calibration (Gneiting et al., 2007). While we will consider calibration and sharpness from a rather applied point of view (Wilks, 2011) in this thesis, a measure-theoretic handling of these concepts in the frame of prediction spaces is provided in Gneiting and Ranjan (2013).

Several methods to assess the calibration and the sharpness of ensemble or density forecasts have been developed, both for the univariate and the multivariate case. Moreover, proper scoring rules have been shown to offer an appealing tool to evaluate calibration and sharpness simultaneously. In view of their utilization in the case studies throughout the thesis, a selection of those evaluation techniques is reviewed in what follows.

2.1 Calibration

As already stated, calibration relates to the statistical compatibility between forecasts and observations, therefore being a joint property of those. Basically, the forecasts are calibrated if events predicted to occur with probability p materialize with frequency p, or in other words, if the verifying observations can be considered as random draws from the predictive distributions.

In univariate settings, calibration is frequently diagnosed by using the verification rank (VR) (Anderson, 1996; Talagrand et al., 1997; Hamill, 2001) in the case of ensemble forecasts or the probability integral transform (PIT) (Dawid, 1984; Diebold et al., 1998; Gneiting et al., 2007) in case of forecast densities, with reliability diagrams (Wilks, 2011) providing an alternative. The VR is the rank of the materializing observation when pooled with the corresponding M ensemble forecast values. Accordingly, the PIT is the value which the corresponding predictive CDF attains at the verifying observation. Adaptations of the PIT for discrete distributions are proposed in Czado et al. (2009).

Calibration can then be checked empirically by plotting the VR histogram or PIT histogram for aggregated forecast cases. If a predictive distribution is calibrated, the VR or the PIT are uniformly distributed on $\{1, \dots, M+1\}$ or the unit interval $[0, 1]$, respectively, and deviations from uniformity in the histograms indicate miscalibration. VR and PIT histograms can be compared directly, and different forms of deviation from uniformity may hint at different reasons for that. For instance, a skew in the histogram indicates biases, a U-shape points at underdispersion, and an inverse U-shape suggests overdispersion in the forecast distributions. Such VR or PIT histograms have been widely used as tools for assessing calibration. However, their uncritical employment might yield to misinterpretations of the forecast quality (Hamill, 2001). In particular, Hamill (2001) argues that a flat histogram does not necessarily indicate a calibrated ensemble. Hence, uniformity is a necessary condition for calibration, but not a sufficient one.

Concerning multivariate situations of dimension $L \geq 2$, we focus on evaluation tools for M-member ensemble forecasts, as this is the format to be dealt with in the methods and case studies later. Specifically, the multivariate rank (MR) (Gneiting et al., 2008), the band

depth rank (BDR) (Thorarinsdottir et al., 2014) and the average rank (AR) (Thorarinsdottir et al., 2014) provide distinct approaches to rank multivariate data and are employed for calibration verification in this thesis. To describe these concepts, let $S := \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ denote the pooled set consisting of the ensemble forecast $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \in \mathbb{R}^L$ and the respective verifying observation $\boldsymbol{x}_0 := \boldsymbol{y} \in \mathbb{R}^L$, where $\boldsymbol{x}_m := (x_m^1, \ldots, x_m^L) \in \mathbb{R}^L$ for $m \in \{0, 1, \ldots, M\}$. According to Gneiting et al. (2008) and Thorarinsdottir et al. (2014), the rank of the observation in S is then derived in two steps, by

(i) applying a pre-rank function $\rho_S : \mathbb{R}^L \to \mathbb{R}_+$ to compute the pre-rank $\rho_S(\boldsymbol{x}_m)$ for every $m \in \{0, 1, \ldots, M\}$ and

(ii) setting the rank of $\boldsymbol{x}_0$ equal to the rank of $\rho_S(\boldsymbol{x}_0)$ in $\{\rho_S(\boldsymbol{x}_0), \rho_S(\boldsymbol{x}_1), \ldots, \rho_S(\boldsymbol{x}_M)\}$, where ties are resolved at random.

In this context, the forecasts and observations in S do not need to be standardized, as the rankings in the pre-rank functions considered below operate componentwise, thus being invariant to such transformations.

In the case of the multivariate rank (MR) (Gneiting et al., 2008), the pre-rank function is defined by

$$\rho_S^{\mathrm{MR}}(\boldsymbol{x}_m) := \sum_{\mu=0}^{M} \mathbb{1}_{\{\boldsymbol{x}_\mu \preceq \boldsymbol{x}_m\}},$$

with the multivariate partial ordering $\boldsymbol{x}_\mu \preceq \boldsymbol{x}_m$ if and only if $x_\mu^\ell \leq x_m^\ell$ for all $\ell \in \{1, \ldots, L\}$. The final MR is then derived according to (ii), and aggregating the MRs over forecast cases leads to an MR histogram, similarly to the univariate case. Conveniently, the interpretation of the form of the resulting MR histograms coincides with that for the VR or PIT histograms in the univariate case discussed before. The MR histogram is appropriate to assess multivariate probabilistic forecasts for low-dimensional quantities (Gneiting et al., 2008; Schuhen et al., 2012; Möller et al., 2013). However, it loses power in higher dimensions (Thorarinsdottir et al., 2014), as will be observed in our case studies in Section 4.3 later.

To address this shortcoming, Thorarinsdottir et al. (2014) propose to use the band depth rank (BDR) to evaluate calibration in high-dimensional settings. Based on the work of López-Pintado and Romo (2009), they introduce the band depth pre-rank function

$$\begin{aligned}
\rho_S^{\mathrm{BDR}}(\boldsymbol{x}_m) &:= \frac{1}{L} \sum_{\ell=1}^{L} \sum_{0 \leq \mu_1 \leq \mu_2 \leq M} \mathbb{1}_{\{\min\{x_{\mu_1}^\ell, x_{\mu_2}^\ell\} \leq x_m^\ell \leq \max\{x_{\mu_1}^\ell, x_{\mu_2}^\ell\}\}} \\
&= \frac{1}{L} \sum_{\ell=1}^{L} \left[\mathrm{rank}_S(x_m^\ell)[M - \mathrm{rank}_S(x_m^\ell)] + [\mathrm{rank}_S(x_m^\ell) - 1] \sum_{\mu=0}^{M} \mathbb{1}_{\{x_\mu^\ell = x_m^\ell\}} \right],
\end{aligned}$$

with $\mathrm{rank}_S(x_m^\ell) := \sum_{\mu=0}^{M} \mathbb{1}_{\{x_\mu^\ell \leq x_m^\ell\}}$ denoting the rank of the ℓ-th coordinate of $\boldsymbol{x}_m$ in S. Thorarinsdottir et al. (2014) note that if $x_m^\ell \neq x_\mu^\ell$ with probability 1 for all $m, \mu \in \{0, 1, \ldots, M\}$ with $m \neq \mu$ and $\ell \in \{1, \ldots, L\}$, the band depth pre-rank function can be simplified to

$$\rho_S^{\mathrm{BDR}}(\boldsymbol{x}_m) = \frac{1}{L} \sum_{\ell=1}^{L} [M - \mathrm{rank}_S(x_m^\ell)][\mathrm{rank}_S(x_m^\ell) - 1] + (M - 1).$$

Finally, the BDR is computed according to (ii), and an BDR histogram can be obtained by aggregating the BDRs over forecast cases, as ever. Thorarinsdottir et al. (2014) note that the interpretation of a BDR histogram somewhat differs from that of a classical univariate VR or PIT histogram. While a calibrated ensemble still yields a flat BDR histogram corresponding to a uniform distribution on $\{1, \ldots, M+1\}$, a skewed BDR histogram with too many high ranks indicates an overdispersive ensemble, and a skewed BDR histogram with too many low ranks points at either an underdispersive or a biased ensemble. Furthermore, a lack of correlation in the ensemble leads to a U-shaped BDR histogram, whereas an ensemble with too high correlations yields an inverse U-shaped BDR histogram.

In addition to the BDR, Thorarinsdottir et al. (2014) propose the average rank (AR) linked to the pre-rank function

$$\rho_S^{\mathrm{AR}}(\boldsymbol{x}_m) := \frac{1}{L} \sum_{\ell=1}^{L} \mathrm{rank}_S(x_m^\ell),$$

which is just the average over the univariate ranks. Again, the final AR is calculated according to (ii), with an aggregation of the ARs over forecast cases leading to an AR histogram, whose interpretation is similar to that of a VR, PIT or MR histogram. That is, U-shaped AR histograms indicate an underdispersive ensemble, and inverse U-shaped AR histograms point at an overdispersed ensemble, while a bias results in a skewed AR histogram. Similar as for the BDR histogram, under- and overestimation of the correlation structure by the ensemble can lead to U- and inverse U-shaped AR histograms, respectively.

In addition to the concepts described above, the minimum spanning tree rank histogram (Smith, 2001; Smith and Hansen, 2004; Wilks, 2004; Gombos et al., 2007), which is however not used in this thesis, provides another established and popular tool to assess the calibration of multivariate ensemble forecasts.

Checking the calibration of full multivariate predictive distributions rather than ensemble forecasts can be performed by using the copula PIT recently introduced by Ziegel and Gneiting (2013). The copula PIT histogram can be viewed as a generalization and variant of the MR histogram, in that both histograms are inclined to look nearly identical for large-sized ensembles such as the 50-member ECMWF ensemble. An alternative calibration evaluation method applying to multivariate density forecasts is the Box density ordinate transform (Box, 1980; O'Hagan, 2003).

2.2 Sharpness

Contrary to calibration, sharpness is a property of the forecast only, referring to the concentration of the predictive distributions. As stated before, forecast distributions ideally should be as sharp as possible, subject to calibration (Gneiting et al., 2007). Since sharpness strongly depends on the units employed, the forecasts should be standardized if there are components that are incomparable in magnitude (Gneiting et al., 2008).

For univariate ensemble forecasts, the empirical ensemble variance, the empirical ensemble standard deviation or the ensemble range are common measures to quantify sharpness. In the case of univariate density forecasts, sharpness can be assessed by the variance and standard deviation, respectively, of the predictive distribution or by the width of specified

prediction intervals. For instance, for the central 80% prediction interval, the width should be as short as possible, with its empirical coverage lying at the nominal 80% level, while 10% each of the observations are located to its left and right, respectively.

Considering the evaluation of sharpness for multivariate L-dimensional forecasts with $L \geq 2$, we follow Gneiting et al. (2008) and use the determinant sharpness (DS) given by

$$\mathrm{DS} := (\det(\Sigma))^{\frac{1}{2L}}$$

as a convenient measure, with $\Sigma \in \mathbb{R}^{L \times L}$ denoting the covariance matrix of an ensemble or density forecast for an $\mathbb{R}^L$-valued quantity. The DS generalizes the univariate standard deviation and can be applied to both ensembles of size $M > L$ and density forecasts, given that the predictive density has finite second moments.

Alternative sharpness measures for the multivariate case include the root mean squared Euclidean distance between the ensemble members and the ensemble mean vector (Stephenson and Doblas-Reyes, 2000) in the case of ensemble forecasts, or the scatter measures proposed by Bickel and Lehmann (1979), among others.

2.3 Proper scoring rules

With the aid of proper scoring rules (Gneiting and Raftery, 2007) as a summarizing measure, calibration and sharpness of probabilistic forecasts can be assessed simultaneously.

A measure-theoretic introduction of scoring rules and their properties, relating to information theory and convex analysis, is given by Gneiting and Raftery (2007). The theoretical interest in scoring rules is also witnessed by the work of Parry et al. (2012) and Ehm and Gneiting (2012). However, we again take an applied point of view and introduce the scoring rule concept to that extent as it is needed for the further development of this thesis.

For our purposes, a scoring rule is a function $s(P, y)$ or $s(P, \boldsymbol{y})$ that assigns a numerical score to the pair (P, y) or $(P, \boldsymbol{y})$ composed of the uni- or multivariate predictive distribution P suggested by the forecaster and the verifying observation $y \in \mathbb{R}$ or observation vector $\boldsymbol{y} \in \mathbb{R}^L$, respectively. In what follows, the predictive distribution P will occasionally be identified with its corresponding predictive CDF F. In this thesis, scores are considered to be negatively oriented penalties, that is, the lower the score the better the predictive performance, with the aim of minimizing them on average. In practice, as in the evaluation of our methods in the case studies later, scores are often reported as averages over forecast cases, that is, over a certain test period.

A very important feature a scoring rule should have is propriety (Gneiting and Raftery, 2007). Assume the predictive distribution Q to be the forecaster's best judgment, and let $s(P, Q)$ denote the expected value of $s(P, \cdot)$ under Q. A scoring rule s is then called proper if it satisfies the expectation inequality

$$s(Q, Q) \leq s(P, Q) \tag{2.1}$$

for all predictive distributions P and Q. It is called strictly proper if (2.1) holds with equality if and only if $P = Q$.

Propriety essentially means that honest and careful assessments of the forecaster are encouraged, in that the forecaster has no incentive to issue any $P \neq Q$ not corresponding to his or her true belief. Thus, it is an important, indispensable property for scoring rules. To some extent, proper scoring rules in a probabilistic prediction setting can be regarded as the analog of consistent scoring functions (Gneiting, 2011a) in the context of point forecasts.

Examples of strictly proper scoring rules that apply to density forecasts only include the logarithmic, quadratic and (pseudo-)spherical score, while the linear score is an example of a scoring rule which is not proper. Variants of these scoring rules are available for both uni- and multivariate settings (Gneiting and Raftery, 2007).

However, we seek to evaluate not only density forecasts, but also predictive distributions expressed in terms of a sample, as is the case for ensemble forecasts, or distributions involving a point mass at zero, as employed in precipitation forecasting, see Chapter 3. Thus, it is more practical to define proper scoring rules directly in terms of predictive CDFs. A very prominent proper scoring rule of this type is the continuous ranked probability score (CRPS), which is in the univariate case given by

$$\begin{aligned} \mathrm{CRPS}(P, y) &:= \int_{-\infty}^{\infty} \left(F(z) - \mathbb{1}_{\{y \leq z\}}\right)^2 \mathrm{d}z && (2.2) \\ &= \mathbb{E}_P[|X - y|] - \frac{1}{2}\,\mathbb{E}_P[|X - X'|], && (2.3) \end{aligned}$$

where X and X' are independent random variables having distribution P with CDF F and finite first moment, and $y \in \mathbb{R}$ denotes the materializing observation. The CRPS as in (2.2) was introduced by Matheson and Winkler (1976), with Gneiting and Raftery (2007) noting its equality to the representation (2.3). It can be reported in the same unit as the verifying observation.

For some distributions, the CRPS can be derived explicitly. In the case of a univariate normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_+$, which will be involved later when modeling temperature, pressure, u- or v-wind, the CRPS is given by

$$\mathrm{CRPS}(\mathcal{N}(\mu, \sigma^2), y) = \sigma \left\{ \frac{y - \mu}{\sigma} \left[2\Phi\left(\frac{y - \mu}{\sigma}\right) - 1 \right] + 2\varphi\left(\frac{y - \mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}} \right\}, \tag{2.4}$$

where Φ and φ denote the CDF and PDF, respectively, of the standard normal distribution with mean 0 and variance 1 (Gneiting et al., 2005).

Friederichs and Thorarinsdottir (2012) derive a closed form expression for the CRPS of a generalized extreme value (GEV) distribution $\mathrm{GEV}(\mu, \sigma, \xi)$ with location, scale and shape parameters μ, σ and ξ, respectively, which they apply to peak wind prediction. Scheuerer (2014) extends these calculations to GEV distributions $\mathrm{GEV}_0(\mu, \sigma, \xi)$ left-censored at zero, for which all mass below zero is assigned to exactly zero. In this case, the CRPS is given by

$$\begin{aligned} \mathrm{CRPS}(\mathrm{GEV}_0(\mu, \sigma, \xi), y) &= (\mu - y)(1 - 2p_y) + \mu p_0^2 - 2\,\frac{\sigma}{\xi}\Big[1 - p_y - \Gamma_l(1 - \xi, -\log(p_y))\Big] \\ &\quad + \frac{\sigma}{\xi}\Big[1 - p_0^2 - 2^{\xi}\,\Gamma_l(1 - \xi, -2\log(p_0))\Big] \qquad (2.5a) \end{aligned}$$

for $\xi \neq 0$, with $p_0 := G(0)$ and $p_y := G(y)$, where G denotes the CDF of the standard GEV distribution, and Γ_l the lower incomplete Gamma function. For $\xi = 0$, an explicit expression

for $\mathrm{CRPS}(\mathrm{GEV}_0(\mu, \sigma, \xi))$ can be derived as well, but Scheuerer (2014) instead employs the approximation

$$\begin{aligned} \mathrm{CRPS}(\mathrm{GEV}_0(\mu, \sigma, \xi), y) &= \frac{\varepsilon - \xi}{2\varepsilon} \mathrm{CRPS}(\mathrm{GEV}_0(\mu, \sigma, \xi = -\varepsilon), y) \\ &\quad + \frac{\varepsilon + \xi}{2\varepsilon} \mathrm{CRPS}(\mathrm{GEV}_0(\mu, \sigma, \xi = \varepsilon), y) \end{aligned} \tag{2.5b}$$

for $\xi \in (-\varepsilon, \varepsilon)$ with $\varepsilon \in \mathbb{R}_+$ reasonably small, where the two scores on the right-hand side are computed according to (2.5a).

If the CRPS cannot be computed in closed form, it can be approximated by using suitable, computationally efficient Monte Carlo methods.

If $P := P_{\mathrm{ens}}$ corresponds to an M-member ensemble forecast $x_1, \ldots, x_M \in \mathbb{R}$, P_{ens} places point mass $1/M$ on the ensemble members, and the CRPS is derived according to (2.3) via

$$\mathrm{CRPS}(P_{\mathrm{ens}}, y) = \frac{1}{M} \sum_{m=1}^{M} |x_m - y| - \frac{1}{2M^2} \sum_{m=1}^{M} \sum_{\mu=1}^{M} |x_m - x_\mu|. \tag{2.6}$$

A decomposition of the CRPS for ensemble forecasts into a reliability, uncertainty and resolution part is presented by Hersbach (2000). Moreover, Bröcker (2012) and Fricker et al. (2013) discuss optimality and fairness aspects, respectively, when evaluating ensembles via the CRPS, depending on different interpretations of an ensemble forecast. These issues will be partly reviewed in Chapter 4.2.

In contrast, if $x \in \mathbb{R}$ is a point forecast and $P := \delta_x$ corresponds to the point measure δ_x, the CRPS reduces to the absolute error $\mathrm{AE} := |x - y| = \mathrm{CRPS}(\delta_x, y)$, thus allowing for a direct comparison between deterministic and probabilistic forecasts. In our case studies, we compute the AE for the point forecast x given by the median of the predictive distribution, which is the Bayes predictor under the absolute error loss function (Gneiting, 2011a).

A modification of the original CRPS is the threshold-weighted continuous ranked probability score (TWCRPS) (Gneiting and Ranjan, 2011), which emphasizes specific regions of interest and can be applied to assess performance in the tails of a predictive distribution P with CDF F, which is useful in case of extreme weather quantities. The TWCRPS is given by

$$\mathrm{TWCRPS}(P, y) := \int_{-\infty}^{\infty} \left(F(z) - \mathbb{1}_{\{y \leq z\}}\right)^2 w(z) \, \mathrm{d}z,$$

where $w(z)$ is a non-negative weight function on $\mathbb{R}$. While for $w(z) := 1$, the TWCRPS is just the original CRPS (2.2), we may set $w(z) := \mathbb{1}_{\{z \geq r\}}$ with $r \in \mathbb{R}$ if we are interested in the right tail of the distribution, for instance (Lerch and Thorarinsdottir, 2013).

A generalization of the CRPS that applies to multivariate quantities is the proper energy score (ES) (Gneiting et al., 2008) given by

$$\mathrm{ES}(P, \boldsymbol{y}) := \mathbb{E}_P[||\boldsymbol{X} - \boldsymbol{y}||] - \frac{1}{2} \mathbb{E}_P[||\boldsymbol{X} - \boldsymbol{X'}||], \tag{2.7}$$

where $||\cdot||$ denotes the Euclidean norm and $\boldsymbol{y} \in \mathbb{R}^L$ the observation vector, and $\boldsymbol{X}$ and $\boldsymbol{X'}$ are independent random vectors with distribution P, with $\mathbb{E}_P[||\boldsymbol{X}||]$ being finite. Variants and generalizations of the ES using other norms are discussed in Gneiting and Raftery (2007), but are not used for the case studies in this thesis, in which we stick to the Euclidean norm.

Especially for density forecasts, the expectations in (2.7) often cannot be computed explicitly, and Monte Carlo methods need to be employed to compute the ES. For example, to derive the ES for a bivariate normal distribution or related predictive densities, we can use the approximation

$$\widehat{\mathrm{ES}}(P, \boldsymbol{y}) := \frac{1}{N} \sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{y}|| - \frac{1}{2(N-1)} \sum_{n=1}^{N-1} ||\boldsymbol{x}_n - \boldsymbol{x}_{n+1}||,$$

with a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^L$ of size $N = 10,000$, for instance, from the predictive density (Gneiting et al., 2008; Schuhen et al., 2012).

If $P := P_{\mathrm{ens}}$ corresponds to an M-member ensemble forecast $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \in \mathbb{R}^L$, the ES can be computed via

$$\mathrm{ES}(P_{\mathrm{ens}}, \boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} ||\boldsymbol{x}_m - \boldsymbol{y}|| - \frac{1}{2M^2} \sum_{m=1}^{M} \sum_{\mu=1}^{M} ||\boldsymbol{x}_m - \boldsymbol{x}_\mu||. \tag{2.8}$$

Besides its appealing properties, the ES is, however, sometimes not able to detect misspecifications of the correlations between the different components of a multivariate quantity (Pinson and Girard, 2012; Pinson and Tastu, 2013). In contrast, the variogram-based proper scoring rules recently proposed by Scheuerer and Hamill (2014) are more discriminative with respect to correlation structures.

For a point forecast $\boldsymbol{x} \in \mathbb{R}^L$ with corresponding point measure $P := \delta_{\boldsymbol{x}}$ in $\boldsymbol{x}$, the ES reduces to the Euclidean error (EE) given by $\mathrm{EE} := ||\boldsymbol{x} - \boldsymbol{y}|| = \mathrm{ES}(\delta_{\boldsymbol{x}}, \boldsymbol{y})$. In the case of an ensemble forecast or a sample from a continuous predictive distribution $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \in \mathbb{R}^L$, we follow Möller et al. (2013) and take the point forecast $\boldsymbol{x}$ to be the multivariate median, which is given by

$$\boldsymbol{x} := \arg\min_{\boldsymbol{\xi} \in \mathbb{R}^L} \sum_{m=1}^{M} ||\boldsymbol{\xi} - \boldsymbol{x}_m||,$$

that is, the vector that minimizes the sum of the Euclidean distance to the single forecast vectors. Practically, the multivariate median $\boldsymbol{x}$ can be derived by the algorithm of Vardi and Zhang (2000), which is implemented in the R (R Core Team, 2013) package `ICSNP`. For a general overview and comparison of algorithms and their implementation to compute the multivariate median, we refer to Fritz et al. (2012). A further discussion of general multivariate quantiles is given in Chapter 7.1.

A further proper scoring rule, that depends on the univariate predictive distribution P only through its mean $\mu_P \in \mathbb{R}$ and its variance $\sigma_P^2 \in \mathbb{R}_+$, is the Dawid-Sebastiani score (DSS) (Dawid and Sebastiani, 1999), which is given by

$$\mathrm{DSS}(P, y) := \log(\sigma_P^2) + \left(\frac{y - \mu_P}{\sigma_P}\right)^2. \tag{2.9}$$

Analogously, a multidimensional generalization of the DSS depending on the multivariate predictive distribution P only through its mean vector $\boldsymbol{\mu}_P \in \mathbb{R}^L$ and its covariance matrix $\Sigma_P \in \mathbb{R}^{L \times L}$ is provided by

$$\mathrm{DSS}(P, \boldsymbol{y}) := \log(\det(\Sigma_P)) + (\boldsymbol{y} - \boldsymbol{\mu}_P)\Sigma_P^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_P)^T. \tag{2.10}$$

The DSS can be applied to both density and ensemble forecasts. For ensemble predictions, the empirical mean and variance or the empirical mean vector and covariance matrix, respectively, are employed for the calculation of the DSS. However, this is reasonable only if the ensemble size M is much larger than the dimension L (Feldmann et al., 2014; Scheuerer and Hamill, 2014).

In multivariate settings, vector-valued quantities should generally be standardized if the components are incomparable in magnitude or incommensurable due to different units. In particular, the Euclidean variant of the ES employed in the thesis does not distinguish between the forecast vector components, such that a standardization may become necessary.

Chapter 3

Statistical ensemble postprocessing

As raw ensembles tend to reveal biases and dispersion errors, statistical postprocessing is required to realize their full potential. There are various state-of-the-art ensemble postprocessing methods, both for univariate and multivariate settings. In this chapter, we review a selection of them, namely those that are relevant for the further development of the thesis and that are employed in our case studies.

3.1 Univariate postprocessing

Univariate postprocessing approaches yield calibrated and sharp predictive distributions valid for single weather variables at single locations and for single look-ahead times. They can roughly be divided into mixture methods, such as Bayesian model averaging (BMA) (Raftery et al., 2005, among others), and regression methods, such as ensemble model output statistics (EMOS), which is also known as non-homogeneous regression (Gneiting et al., 2005, among others). BMA and EMOS are implemented in the packages `ensembleBMA` (Fraley et al., 2011) and `ensembleMOS`, respectively, which are available in the R language and environment (R Core Team, 2013) and can be downloaded at `www.r-project.org`.

In this section, we discuss the BMA and EMOS approaches, where similar reviews have been given in Thorarinsdottir et al. (2012), Schefzik et al. (2013), Gneiting and Katzfuss (2014), Gneiting (2014) and Williams et al. (2014), and assess their predictive performance in a case study using the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble.

3.1.1 Bayesian model averaging (BMA)

For a fixed location and prediction horizon, let y be a weather quantity of interest and $x_1, \ldots, x_M$ the corresponding M ensemble member forecasts. The BMA approach then employs mixture distributions of the form

$$y|x_1 \ldots, x_M \sim \sum_{m=1}^{M} w_m \, f(y|x_m). \tag{3.1}$$

The left-hand side of (3.1) refers to the conditional distribution of y given $x_1, \ldots, x_M$, and $f(y|x_m)$ denotes a parametric distribution or kernel depending on x_m only, where the specific choice of f depends on the weather quantity of interest. The weights $w_1, \ldots, w_M \geq 0$ satisfy $\sum_{m=1}^{M} w_m = 1$ and reflect the respective member's relative predictive skill over a

Table 3.1: BMA settings for univariate weather quantities y, based on an ensemble forecast $x_1, \ldots, x_M$. For precipitation amount, we refer to $y^{1/3} \in \mathbb{R}_+$, as the gamma kernels apply to a cube root transformation. In the case of wind direction, z_m is a bias-corrected ensemble member value on the circle $\mathbb{S}$, and κ_m is a concentration parameter for a fixed ensemble member $m \in \{1, \ldots, M\}$. For the rest, a_m and b_m are mean parameters, and σ_m^2, c_m and d_m are variance parameters. This table follows the similar summary in Table 1 in Schefzik et al. (2013).

Weather quantity	Range	Kernel (f)	Mean	Variance
Temperature	$y \in \mathbb{R}$	Normal	$a_m + b_m x_m$	σ_m^2
Pressure	$y \in \mathbb{R}$	Normal	$a_m + b_m x_m$	σ_m^2
u- and v-Wind	$y \in \mathbb{R}$	Normal	$a_m + b_m x_m$	σ_m^2
Precipitation amount	$y^{1/3} \in \mathbb{R}_+$	Gamma	$a_m + b_m x_m^{1/3}$	$c_m + d_m x_m$
Wind speed	$y \in \mathbb{R}_+$	Gamma	$a_m + b_m x_m$	$c_m + d_m x_m$
	$y \in \mathbb{R}_+$	Truncated normal with cut-off at zero	$a_m + b_m x_m$	σ_m^2
Wind direction	$y \in \mathbb{S}$	von Mises	z_m	κ_m^{-1}
Visibility	$y \in [0, 1]$	Beta	$a_m + b_m x_m^{1/2}$	$c_m + d_m x_m^{1/2}$

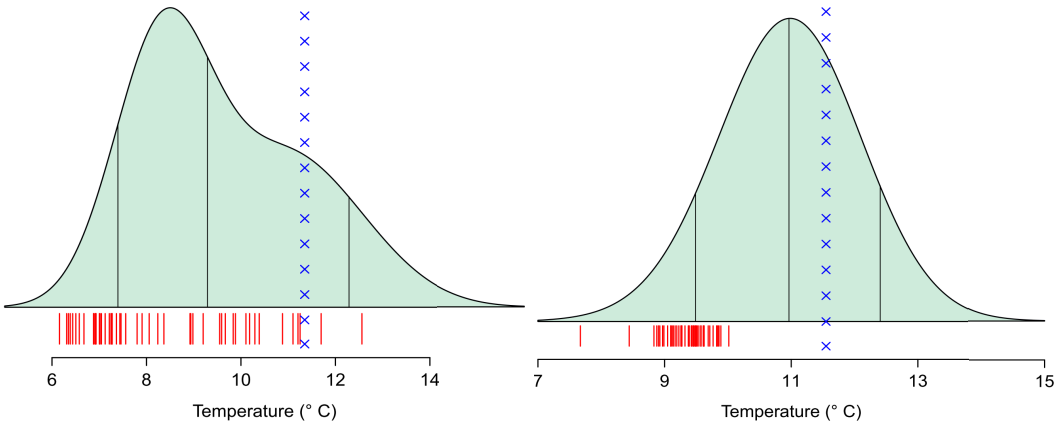
training period.

BMA variants are available for temperature, pressure, u- and v-wind (Raftery et al., 2005), precipitation (Sloughter et al., 2007), wind speed (Sloughter et al., 2010; Baran, 2014), wind direction (Bao et al., 2010), fog (Roquelaure and Bergot, 2008) and visibility and ceiling (Chmielecki and Raftery, 2011). The corresponding BMA settings, especially the specific choice of the kernel f, are documented in Table 3.1.

The mean and variance parameters as described in Table 3.1 are often subject to constraints. Frequently, the variance parameters are supposed to be constant across ensemble members, thus being independent of $m \in \{1, \ldots, M\}$.

The BMA model parameters and weights are typically estimated via regression and optimum score approaches, with maximum likelihood optimization being a special case thereof, based on a training data set consisting of past forecasts and observations. Usually, the training period comprises a sliding training window of the past 20 to 40 days. This continuous update allows for a good adaptation to changes in seasons and weather regimes. Throughout this thesis, we employ a sliding training period of 30 days for estimation in our case studies. Although larger training periods may principally improve the estimation results, they might also introduce biases caused by seasonal effects, such that there is a trade-off made in favor of a shorter training period here. However, several flexible adaptive estimation techniques have also been proposed and studied (Pinson et al., 2009; Raftery et al., 2010; Pinson, 2012), including a recursive maximum likelihood estimation procedure.

Not only the length of the training period, but also the spatial composition of the training data has to be specified, with local and regional approaches being available. Local techniques employ training data from the specific station site or grid box of interest only, such that the parameters vary from site to site, as they are tailored to the local terrain. In contrast, regional approaches composite training data spatially and estimate a single set of parameters, which is then used over entire regions. In the case studies of this thesis, we use local estimation approaches if we deal with the individual locations of Berlin, Hamburg and Frankfurt, respectively, and regional estimation techniques when considering test areas.



(a) 21 April 2011 (b) 25 April 2011

Figure 3.1: 24 hour ahead BMA predictive PDF for temperature at Berlin, valid 2:00 am on (a) 21 April 2011 and (b) 25 April 2011, respectively. The 50-member ECMWF ensemble forecast is shown in red, and the verifying observation in blue, while the black lines indicate the 10th, 50th and 90th percentiles of the BMA predictive distribution.

For ensembles consisting of M exchangeable members, such as the ECMWF ensemble, the BMA weights are assumed to be equal for all ensemble members (Fraley et al., 2010), that is, $w_m = 1/M$ for all m. Moreover, the BMA mean and variance parameters are assumed to be constant across the members in this case, such that they do not depend on m either. Suggestions how to deal with missing data in the context of BMA are given in Fraley et al. (2010).

In view of the case studies, we now describe the BMA approaches for temperature, pressure, u- and v-wind (Raftery et al., 2005) and precipitation (Sloughter et al., 2007) more in detail. For temperature, pressure, u- and v-wind (Raftery et al., 2005), the specific BMA model based on (3.1) is given by

$$y|x_1, \ldots, x_M \sim \sum_{m=1}^{M} w_m \mathcal{N}(a_m + b_m x_m, \sigma_m^2),$$

that is, the kernel f is chosen to be normal with mean $a_m + b_m x_m$ and variance σ_m^2, where $a_1, \ldots, a_M \in \mathbb{R}$ and $b_1, \ldots, b_M \in \mathbb{R}$ are the mean parameters, and $\sigma_1^2, \ldots, \sigma_M^2 \in \mathbb{R}_+$ are the variance parameters. In the standard implementation, the variance parameters are assumed to be constant across the ensemble members, that is, $\sigma^2 := \sigma_1^2 = \cdots = \sigma_M^2$. Based on the corresponding training data, first the mean parameters are estimated for each ensemble member via linear regression. Then, the mean parameters are considered to be known, and the weights $w_1, \ldots, w_M$ and the variance parameter(s) are derived via maximum likelihood estimation, using the expectation-maximization algorithm.

Figure 3.1 shows two examples of BMA predictive PDFs for temperature. Being a mixture, a BMA predictive PDF for a specific weather quantity can conceptionally look quite different for different test cases, depending on the corresponding amount of dispersion within the raw ensemble. In the examples here, the raw ensemble in Figure 3.1 (a) has a rather high spread, while the raw ensemble in Figure 3.1 (b) is more concentrated.

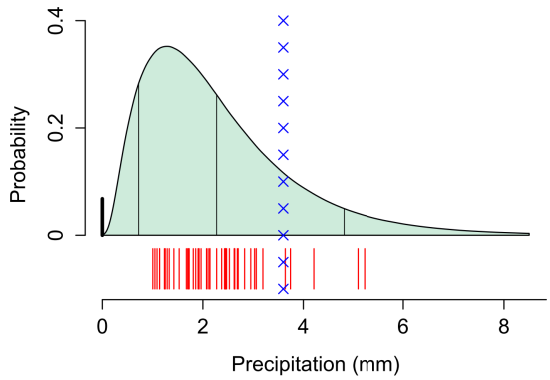


Figure 3.2: 24 hour ahead BMA predictive PDF for six hour precipitation accumulation at Hamburg, valid 1:00 am on 11 December 2010. The 50-member ECMWF ensemble forecast is shown in red, and the verifying observation in blue, while the thin black lines indicate the 10th, 50th and 90th percentiles of the mixed discrete-continuous BMA predictive distribution. The thick black bar represents the point mass of 0.068 at zero, and the density at positive accumulations thus has a mass of 0.932.

In the case of precipitation, the modeling is somewhat more involved, in that the predictive distribution is mixed discrete-continuous, consisting of a point mass at zero, as there is a positive probability that no rain occurs, and a possibly highly skewed density on the positive real axis. Specifically, Sloughter et al. (2007) take the kernel f to be a Bernoulli-Gamma mixture. The Bernoulli part provides a point mass at zero based on a logistic regression link, in that

$$\text{logit}\left(\mathbb{P}(y=0|x_m)\right) = \log\left(\frac{\mathbb{P}(y=0|x_m)}{\mathbb{P}(y>0|x_m)}\right) = \alpha_m + \beta_m x_m^{1/3} + \gamma_m \delta_m,$$

with parameters α_m, β_m and γ_m and a predictor δ_m satisfying $\delta_m = 1$ if $x_m = 0$ and $\delta_m = 0$ otherwise, for $m \in \{1, \ldots, M\}$. Here, $\mathbb{P}(y > 0|x_m)$ is the probability of non-zero precipitation given the forecast x_m. In other words, the Bernoulli component is specified by

$$\mathbb{P}(y=0|x_m) = \frac{\exp(\alpha_m + \beta_m x_m^{1/3} + \gamma_m \delta_m)}{1 + \exp(\alpha_m + \beta_m x_m^{1/3} + \gamma_m \delta_m)}.$$

The continuous part, dealing with the precipitation amount given that it is not zero, is given by a gamma distribution in terms of the cube root transformation $y^{1/3}$ of the precipitation accumulation. Hence, the specific BMA model of Sloughter et al. (2007) in case of precipitation is

$$y^{1/3}|x_1, \ldots, x_M \sim \sum_{m=1}^{M} w_m[\mathbb{P}(y=0|x_m)\mathbb{1}_{\{y=0\}} + \mathbb{P}(y>0|x_m)h_m(y^{1/3}|x_m)\mathbb{1}_{\{y>0\}}],$$

with h_m denoting a gamma distribution with mean $\mu_m = a_m + b_m x_m^{1/3}$ and variance $\sigma_m^2 = c_m + d_m x_m$, including parameters a_m, b_m, c_m and d_m. The member-specific parameters α_m, β_m and γ_m are estimated via logistic regression, while the member-specific coefficients a_m and b_m are determined via linear regression. The parameters c_m and d_m, which are usually assumed to be constant across the ensemble members, and the member-specific weights

$w_1, \ldots, w_M$ are estimated via the maximum likelihood method, employing the expectation-maximization algorithm.

Exemplarily, a BMA predictive distribution for precipitation in terms of the non-transformed precipitation accumulation y is given in Figure 3.2.

For temperature and precipitation, BMA has been applied in real time to generate forecasts over the Pacific Northwest regions of the United States based on the University of Washington Mesoscale Ensemble (UWME) (Eckel and Mass, 2005), which are available at `www.probcast.com` (Mass et al., 2009). Moreover, BMA has been employed to postprocess temperature predictions in Canada (Wilson et al., 2007), Iran (Soltanzadeh et al., 2011) and Hungary (Baran et al., 2014b), respectively, while Courtney et al. (2013) apply it to wind energy over Ireland.

Critical reviews and suggestions for improvement of BMA are provided by Hamill (2007) and Bishop and Shanley (2008), among others.

In heterogeneous terrain, for instance in an area comprising ocean, mountains and lowlands, it might be preferable to employ locally adaptive parameters that vary spatially. Spatially adaptive so-called geostatistical model averaging approaches, which fit a BMA model using Bayesian regulization at each observation site separately and then interpolate the estimated parameters to locations where no observations are available, have been proposed by Kleiber et al. (2011a) and Kleiber et al. (2011b) for temperature and precipitation, respectively, and extend BMA directly.

3.1.2 Ensemble model output statistics (EMOS)

Contrary to BMA, being based on mixture distribution, the ensemble model output statistics (EMOS) approach, which is also known as non-homogeneous regression, fits a single parametric distribution g employing summary statistics from the ensemble. Again, for a fixed location and prediction horizon, let y be a weather quantity of interest and $x_1, \ldots, x_M$ the corresponding ensemble member forecasts. The EMOS predictive distribution then has the general form

$$y|x_1, \ldots, x_M \sim g(y|x_1, \ldots, x_M), \tag{3.2}$$

where the parameters of the parametric family of probability distributions on the right-hand side of (3.2) depend on all ensemble members simultaneously.

EMOS implementations are available for temperature, pressure, u- and v-wind (Gneiting et al., 2005), precipitation (Wilks, 2009; Scheuerer, 2014), wind speed (Thorarinsdottir and Gneiting, 2010; Möller, 2013; Lerch and Thorarinsdottir, 2013; Baran and Lerch, 2014) and wind gusts (Thorarinsdottir and Johnson, 2012), respectively. An overview is given in Table 3.2.

Concerning the choice of the training period and the composition of the training data, respectively, the same comments hold as for BMA.

In the case of temperature, pressure, u- and v-wind, we follow Gneiting et al. (2005) and

Table 3.2: EMOS settings for univariate weather quantities y. In the case of precipitation amount, we refer to $y^{1/2} \in \mathbb{R}_+$ for the truncated logistic distribution approach of Wilks (2009), as it applies to root-transformed precipitation accumulations. This table follows the similar summary in Table 2 in Schefzik et al. (2013).

Weather quantity	Range	Distribution (g)
Temperature	$y \in \mathbb{R}$	Normal
Pressure	$y \in \mathbb{R}$	Normal
u- and v-Wind	$y \in \mathbb{R}$	Normal
Precipitation amount	$y^{1/2} \in \mathbb{R}_+$	Truncated logistic
	$y \in \mathbb{R}_+$	Generalized extreme value left-censored at zero
Wind speed	$y \in \mathbb{R}_+$	Truncated normal with cut-off at zero
	$y \in \mathbb{R}_+$	Gamma
	$y \in \mathbb{R}_+$	Generalized extreme value
	$y \in \mathbb{R}_+$	Log-normal

use the regression model

$$y = a + b_1 x_1 + \ldots + b_M x_M + \varepsilon \text{ with } \varepsilon \sim \mathcal{N}(0, c + ds^2), \tag{3.3}$$

where $a, b_1, \ldots, b_M \in \mathbb{R}$ and $c, d \in \mathbb{R}_+^0$ are parameters that need to be estimated, and $s^2 := \frac{1}{M} \sum_{m=1}^{M} (x_m - \bar{x})^2$ denotes the empirical ensemble variance, with the empirical ensemble mean $\bar{x} := \frac{1}{M} \sum_{m=1}^{M} x_m$. The corresponding Gaussian predictive EMOS distribution is then given by

$$y|x_1, \ldots, x_M \sim \mathcal{N}(a + b_1 x_1 + \ldots + b_M x_M, c + ds^2), \tag{3.4}$$

having mean $a + b_1 x_1 + \ldots + b_M x_M$ and variance $c + ds^2$. While a is a bias correction term, the regression coefficients $b_1, \ldots, b_M$ reflect the performance of the ensemble members over the training period relative to the other members, as well as the correlations between the ensemble members. If the ensemble members can be considered exchangeable, it needs to be assumed that the regression coefficients are equal, that is, $b := b_1 = \cdots = b_M$. In this case,

$$\begin{aligned} y|x_1, \ldots, x_M &\sim \mathcal{N}(a + b(x_1 + \cdots + x_M), c + ds^2) \\ &= \mathcal{N}(a + b\ M\ \bar{x}, c + ds^2). \end{aligned} \tag{3.5}$$

The parameters $a, b_1, \ldots, b_M, c$ and d are estimated by minimizing the training CRPS, which is based on Formula (2.4) and expressed as an analytic function of the parameters. This estimation technique comes within an optimum score frame, which generalizes the classical maximum likelihood notion (Gneiting et al., 2005).

An EMOS predictive PDF in the case of temperature has been presented in Figure 1.4 in the introductory chapter, and another example for pressure is shown in Figure 3.3.

For precipitation, Scheuerer (2014) provides an EMOS variant using the generalized extreme value (GEV) distribution with CDF

$$G(y) := \begin{cases} \exp\Big[-\Big\{1 + \xi\Big(\frac{y-\mu}{\sigma}\Big)\Big\}^{-1/\xi}\Big] & \text{for } \xi \neq 0, \\ \exp\Big[-\exp\Big(-\frac{(y-\mu)}{\sigma}\Big)\Big] & \text{for } \xi = 0, \end{cases}$$

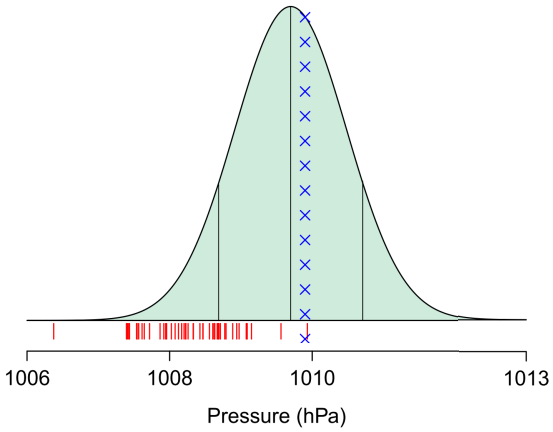


Figure 3.3: 24 hour ahead EMOS predictive normal PDF for pressure at Frankfurt, valid 2:00 am on 20 June 2010. The 50-member ECMWF ensemble forecast is shown in red, and the verifying observation in blue, while the black lines indicate the 10th, 50th and 90th percentiles of the EMOS predictive distribution.

where the parameters μ, σ and ξ characterize location, scale and shape, respectively, of the GEV distribution. For $\xi < 0$, $y > \mu - (\sigma/\xi)$, one sets $G(y) := 1$, while for $\xi > 0$, $y < \mu - (\sigma/\xi)$, one defines $G(y) := 0$. Scheuerer (2014) assumes $\xi \in (-0.278, 1)$ because then, the GEV has positive skew and its mean η exists, where

$$\eta = \begin{cases} \mu + \sigma\frac{\Gamma(1-\xi)-1}{\xi} & \text{for } \xi \neq 0, \\ \mu + \sigma\gamma & \text{for } \xi = 0, \end{cases}$$

with the Gamma function Γ and the Euler-Mascheroni constant $\gamma \approx 0.5772$. In order to be appropriate for modeling precipitation amounts, Scheuerer (2014) considers the GEV distribution to be left-censored at zero, such that all mass below zero is assigned to exactly zero, with predictive CDF

$$\tilde{G}(y) := \begin{cases} G(y) & \text{for } y \geq 0, \\ 0 & \text{for } y < 0. \end{cases}$$

According to Scheuerer (2014), this distribution is non-negative and exactly zero with positive probability if either $\xi \leq 0$ or $\xi > 0$ and $\mu < \sigma/\xi$. In his model, Scheuerer (2014) links the parameters η and σ from the GEV distribution left-censored at zero to the raw ensemble forecast $\boldsymbol{x} := (x_1, \ldots, x_M)$ by setting

$$\eta := \alpha_0 + \alpha_1\bar{x} + \alpha_2\overline{\mathbb{1}_{\{\boldsymbol{x}=0\}}}$$

and

$$\sigma := \beta_0 + \beta_1\,\mathrm{MD}(\boldsymbol{x}),$$

where $\bar{x}$ is the ensemble mean,

$$\overline{\mathbb{1}_{\{\boldsymbol{x}=0\}}} := \frac{1}{M}\sum_{m=1}^{M}\mathbb{1}_{\{x_m=0\}}$$

the fraction of zero precipitation and

$$\mathrm{MD}(\boldsymbol{x}) := \frac{1}{M^2}\sum_{m=1}^{M}\sum_{\mu=1}^{M}|x_m - x_\mu|$$

the ensemble mean difference. The parameters $\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1$ and ξ are then estimated by minimizing the training CRPS according to the formulas (2.5a) and (2.5b), respectively. Details about the implementation can be found in Scheuerer (2014).

Alternatively to the approach of Scheuerer (2014), truncated logistic distributions can be employed to model precipitation, motivated as follows. Predictions for the probability of the precipitation amount exceeding a certain threshold have been constructed using logistic regression (Wilks and Hamill, 2007; Hamill et al., 2008). However, if a full predictive distribution is desired, this approach fails because it is typically inconsistent across thresholds, violating the monotonicity restriction for CDFs. To address this shortcoming, Wilks (2009) introduced a technique, in which the postprocessed predictive CDF takes the form

$$G(y|x_1, \ldots, x_M) = \frac{\exp(a + b_1 x_1 + \ldots + b_M x_M + h(y))}{1 + \exp(a + b_1 x_1 + \ldots + b_M x_M + h(y))}, \tag{3.6}$$

with parameters $a, b_1, \ldots, b_M$, and with h increasing strictly monotonically and without bounds as a function of the precipitation accumulation $y \geq 0$, whereas $G(y|x_1, \ldots, x_M) = 0$ for $y < 0$. If h is chosen to be linear, mixtures of a point mass at zero and a truncated logistic distribution are obtained. In view of the parametric family in (3.6), the approach of Wilks (2009) can be considered an EMOS method. In fact, Wilks (2009) applies the truncated logistic distributions to root-transformed precipitation amounts $y^{1/2}$. More generally, the use of transformations $x_1^\alpha, \ldots, x_M^\alpha$ and y^β, respectively, with $\alpha, \beta \in \mathbb{R}_+$, in model (3.6) is feasible (Gneiting, 2014). The coefficients $a, b_1, \ldots, b_M$ and those included in the model for h need to be fitted from appropriate training data. Generalizations allowing for interaction terms are discussed by Ben Bouallègue (2013).

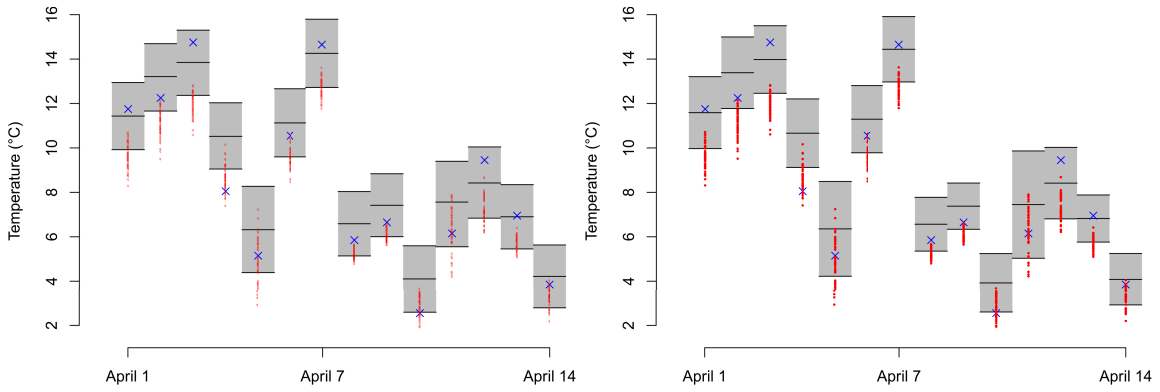
The EMOS approach for temperature was employed by Hagedorn et al. (2008) and Kann et al. (2009), among others, while Lerch and Thorarinsdottir (2013) provide a review and comparison of EMOS regression models for wind speed.

Locally adaptive methods for estimating EMOS parameters in the case of temperature are provided by Scheuerer and Büermann (2014) and Scheuerer and König (2014), respectively, who introduce further refinements of EMOS employing geostatistical methods, in which the postprocessing at individual stations varies in space. Similar locally adaptive EMOS models for wind speed are considered in Möller (2013).

In addition, Möller (2014) proposes a spatially adaptive extension of univariate EMOS for temperature called Markovian EMOS due to the Markovian dependence structure induced by the Gaussian Markov random field representations of the Gaussian fields used to model the parameters in the approach.

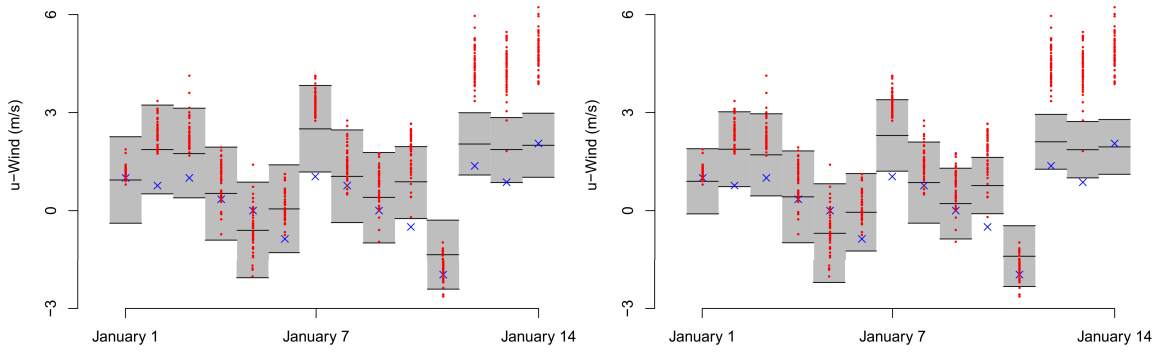
3.1.3 Case study

We now apply the univariate BMA and EMOS approaches with a sliding training window of 30 days to the ECMWF raw ensemble forecasts and assess their predictive performances over the one-year test period from 1 May 2010 to 30 April 2011 by using some of the univariate evaluation methods presented in Chapter 2. The focus is on 24 hour ahead predictions for temperature, pressure, u- and v-wind and precipitation at the three locations of Berlin, Hamburg and Frankfurt individually, valid at 00:00 UTC, corresponding to the local time of 1:00 am in winter and 2:00 am in summer, respectively.



(a) BMA postprocessing (b) EMOS postprocessing

Figure 3.4: 24 hour ahead ensemble forecasts of temperature at Hamburg, valid 2:00 am for the test period from 1 April 2011 to 14 April 2011, and characteristic quantities of the corresponding postprocessed (a) BMA and (b) EMOS predictive distribution, respectively. The ECMWF raw ensemble forecast is indicated by the red dots, and the verifying observation by the blue cross, while the black lines show the 10th, 50th and 90th percentiles of the (a) BMA and (b) EMOS predictive distribution. The 80% BMA/EMOS prediction intervals are represented by the segments shaded in gray.



(a) BMA postprocessing (b) EMOS postprocessing

Figure 3.5: 24 hour ahead ensemble forecasts of u-wind at Frankfurt, valid 1:00 am for the test period from 1 January 2011 to 14 January 2011, and characteristic quantities of the corresponding postprocessed (a) BMA and (b) EMOS predictive distribution, respectively. The ECMWF raw ensemble forecast is indicated by the red dots, and the verifying observation by the blue cross, while the black lines show the 10th, 50th and 90th percentiles of the (a) BMA and (b) EMOS predictive distribution. The 80% BMA/EMOS prediction intervals are represented by the segments shaded in gray.

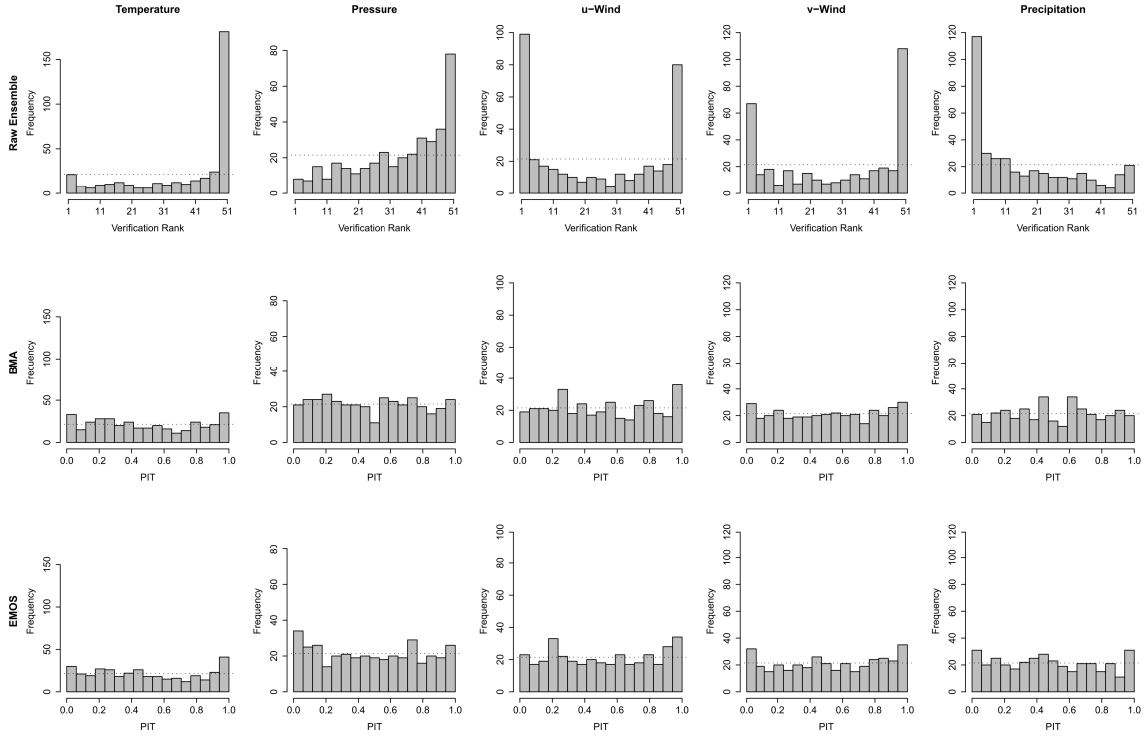


Figure 3.6: Calibration check for 24 hour ahead forecasts of temperature, pressure, u- and v-wind and precipitation at Berlin over the test period from 1 May 2010 to 30 April 2011. Top row: VR histograms for the ECMWF raw ensemble. Middle row: PIT histograms for BMA postprocessed predictive distributions. Bottom row: PIT histograms for EMOS postprocessed predictive distributions

Table 3.3: Average CRPS and AE for univariate 24 hour ahead forecasts of temperature, pressure, u- and v-wind and precipitation at Berlin, Hamburg and Frankfurt. The results are averaged over the test period from 1 May 2010 to 30 April 2011.

		CRPS			AE		
		Berlin	Hamburg	Frankfurt	Berlin	Hamburg	Frankfurt
Temperature	Raw Ensemble	1.21	1.01	1.23	1.50	1.26	1.53
(°C)	BMA	0.90	0.79	0.88	1.27	1.10	1.23
	EMOS	0.90	0.78	0.87	1.24	1.09	1.22
Pressure	Raw Ensemble	0.54	0.51	0.55	0.75	0.71	0.75
(hPa)	BMA	0.43	0.39	0.43	0.62	0.54	0.61
	EMOS	0.44	0.40	0.44	0.60	0.54	0.60
u-Wind	Raw Ensemble	0.83	0.89	0.96	1.06	1.11	1.19
(m/s)	BMA	0.69	0.69	0.59	0.97	0.97	0.80
	EMOS	0.70	0.68	0.60	0.97	0.96	0.81
v-Wind	Raw Ensemble	0.74	0.76	1.25	0.95	0.95	1.49
(m/s)	BMA	0.58	0.64	1.06	0.82	0.90	1.49
	EMOS	0.59	0.65	1.07	0.82	0.89	1.50
Precipitation	Raw Ensemble	0.25	0.31	0.41	0.32	0.39	0.51
(mm)	BMA	0.23	0.37	0.40	0.30	0.44	0.49
	EMOS	0.27	0.36	0.43	0.32	0.43	0.50

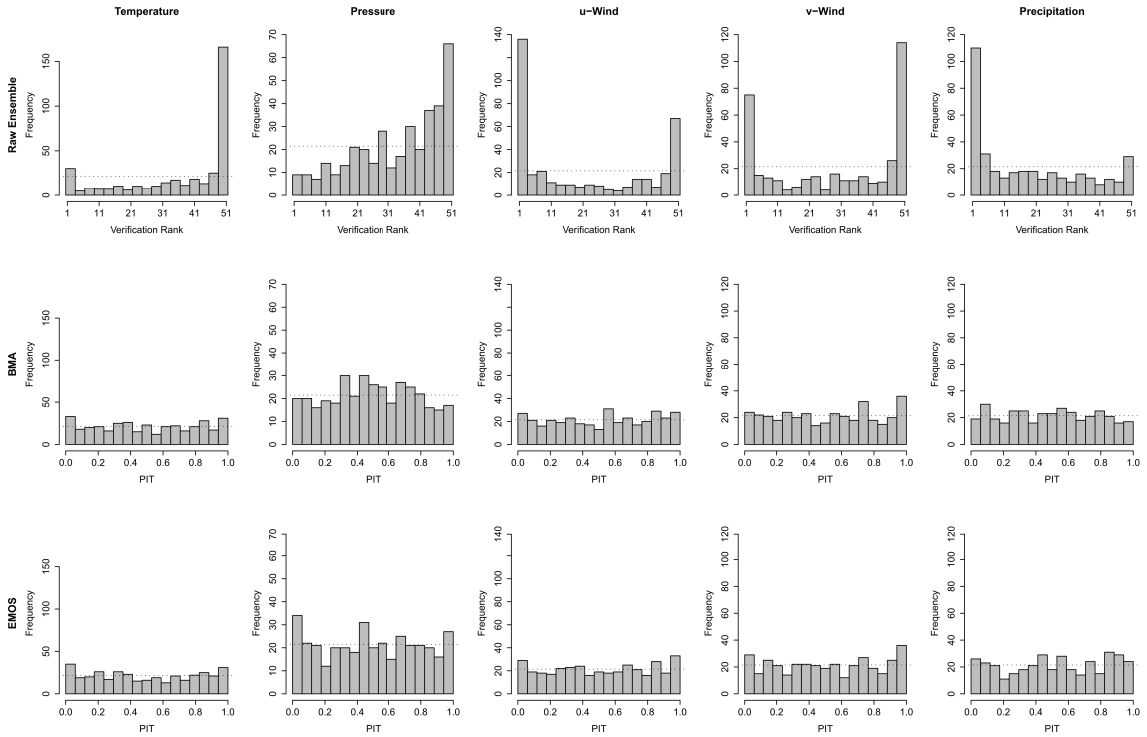


Figure 3.7: Calibration check for 24 hour ahead forecasts of temperature, pressure, u- and v-wind and precipitation at Hamburg over the test period from 1 May 2010 to 30 April 2011. Top row: VR histograms for the ECMWF raw ensemble. Middle row: PIT histograms for BMA postprocessed predictive distributions. Bottom row: PIT histograms for EMOS postprocessed predictive distributions.

The positive impact of ensemble postprocessing via BMA or EMOS is demonstrated in Figures 3.4 and 3.5 comprising examples for temperature at Hamburg, supplementing Figure 1.4 in the introductory chapter, and u-wind at Frankfurt, respectively, over selected two-week test periods. The 50-member ECMWF raw ensemble forecasts are marked by the red dots, and the verifying observations by the blue crosses, while the black horizontal lines indicate the 10th, 50th and 90th percentiles, respectively, of the corresponding BMA or EMOS predictive distributions. In far too many cases, the observations obviously fall outside the raw ensemble range. However, they mostly lie within the 80% prediction intervals of the BMA or EMOS predictive distributions, which are indicated by the segments shaded in gray.

Calibration is evaluated via the verification rank (VR) histograms for the raw ensemble and the probability integral transform (PIT) histograms for the BMA and EMOS method, respectively. The corresponding histograms can be compared directly, and the results for the different weather variables at Berlin, Hamburg and Frankfurt are shown in Figures 3.6 to 3.8 for BMA and EMOS, respectively. Throughout, the raw ensemble predictions reveal either skewed or U-shaped VR histograms, pointing at biases and underdispersion, respectively. On the other hand, the BMA or EMOS forecast densities are clearly better calibrated than the raw ensemble, with mostly nearly uniform PIT histograms. Note that in the case of the PIT histograms for precipitation, observations of zero precipitation are randomized within their probabilistics range to avoid a false impression of bias, as suggested in the manual to

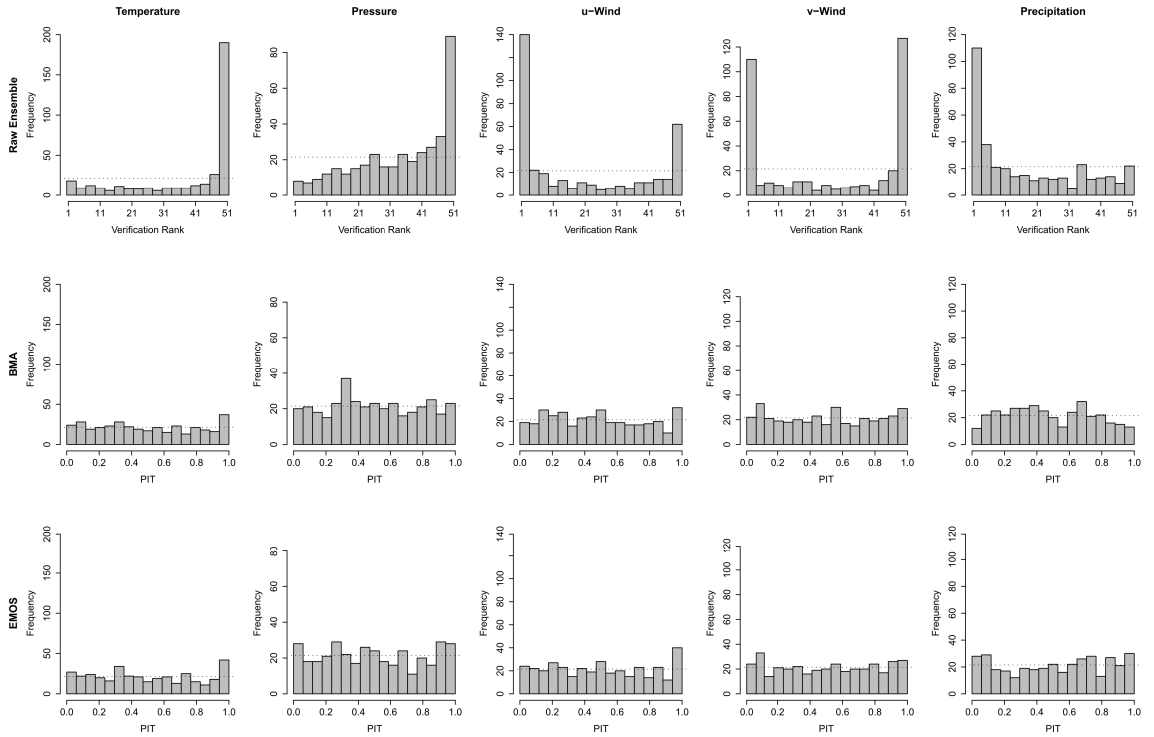


Figure 3.8: Calibration check for 24 hour ahead forecasts of temperature, pressure, u- and v-wind and precipitation at Frankfurt over the test period from 1 May 2010 to 30 April 2011. Top row: VR histograms for the ECMWF raw ensemble. Middle row: PIT histograms for BMA postprocessed predictive distributions. Bottom row: PIT histograms for EMOS postprocessed predictive distributions

the R package `ensembleBMA`.

Table 3.3 shows the average results for the continuous ranked probability score (CRPS) and the absolute error (AE) as overall performance measures, where in the case of the AE, the reference point forecast is given by the median, that is, the 50th percentile, of the corresponding predictive BMA or EMOS distribution. The CRPS is computed according to the formulas in Section 2.3. In the case of temperature, pressure, u- and v-wind, respectively, both the BMA and the EMOS postprocessing significantly improve the predictive skill, yielding lower scores than the unprocessed raw ensemble. The only exception can be observed for v-wind at Frankfurt, where BMA and EMOS at least fail to outperform the raw ensemble in terms of the AE. As far as precipitation is concerned, the situation is somewhat more mixed. For this variable, BMA outperforms the raw ensemble in terms of both the CRPS and the AE at Berlin and Frankfurt, respectively, while it fails to do so at Hamburg. EMOS even performs worse than the raw ensemble in terms of the CRPS at all three locations. At least, EMOS does not perform worse than the raw ensemble with respect to the AE at Berlin and Frankfurt. Although calibration of BMA and EMOS for precipitation appears to be rather fine at each of the three locations, the partial failures in terms of the scores, especially the CRPS, might be due to potential deficiencies in the details of the two postprocessing approaches and their implementation. A deeper analysis of the corresponding BMA and EMOS forecasts involving the (half-) Brier score (Brier, 1950; Wilks, 2011) reveals that the two postprocessing techniques generally work well in predicting precipitation or no precipitation. In case they are not able to outperform the raw ensemble

with respect to the scores, they on average fail in predicting extreme precipitation events dealing with higher thresholds. All this witnesses that precipitation generally appears to be one of the most difficult and complex weather variables to handle.

In a nutshell, the case study underlines previous results, in that both BMA and EMOS usually outperform the unprocessed raw ensemble, while showing comparable, nearly equal predictive skill. Generally, BMA is more flexible, while EMOS is more parsimonious.
A similar comparison study of the impact of BMA and EMOS postprocessing, respectively, for temperature and wind speed has been conducted by Baran et al. (2014a), using forecast data over Hungary.

3.2 Multivariate postprocessing

The univariate BMA and EMOS approaches presented in Section 3.1 only apply to a single location, a single weather quantity and a single prediction horizon. This is rather unfortunate, as independently postprocessed forecasts fail to account for spatial, inter-variable and temporal dependence structures. However, in many applications, such as air traffic control (Chaloulos and Lygeros, 2007), flood warning (Schaake et al., 2010), winter road maintenance (Berrocal et al., 2010) or the management of renewable energy sources (Pinson, 2013), it is crucial that such dependencies are restored by the postprocessing methods. Hence, the development of postprocessing techniques leading to physically realistic and coherent probabilistic forecasts of spatio-temporal weather trajectories is of tremendous interest, with the aim to handle multiple locations, weather variables and look-ahead times simultaneously.
Meanwhile, several multivariate ensemble postprocessing approaches have thus been designed, yielding truly multivariate predictive distributions. For example, extensions of the BMA and EMOS approaches for temperature in Section 3.1 being able to address spatial dependence structures have been introduced by Berrocal et al. (2007) and Feldmann et al. (2014) within the frame of Spatial BMA and Spatial EMOS, respectively.

Copulas (Joe, 1997; Nelsen, 2006) play an important part in the context of dependence modeling, and they turn out to be well suited to face the challenge of constructing multivariate postprocessing approaches. In this section, the importance of copulas for handling dependencies is described. Moreover, several selected examples of multivariate ensemble postprocessing approaches are presented, where the focus is on those methods that are either relevant for the whole context of the thesis or which can be interpreted in a copula framework.

3.2.1 Dependence modeling via copulas

Originally introduced by Sklar (1959), copulas play an important part in probability and statistics whenever modeling of stochastic dependence is required. Essentially, a copula is a multivariate CDF with standard uniform margins. A precise mathematical definition is given in the following (Nelsen, 2006).

Definition 3.1. A function $C : [0,1]^L \to [0,1]$ is an L-dimensional copula or L-copula if it satisfies the following conditions.

(C1) C is grounded, in that $C(u_1, \ldots, u_L) = 0$ if $u_\ell = 0$ for at least one $\ell \in \{1, \ldots, L\}$.

(C2) $C(1, \ldots, 1, u_\ell, 1, \ldots, 1) = u_\ell$ for all $\ell \in \{1, \ldots, L\}$.

(C3) C is L-increasing, in that

$$\Delta_{a_L}^{b_L} \cdots \Delta_{a_1}^{b_1} C(u_1, \ldots, u_L) \geq 0$$

for all $a_\ell, b_\ell \in [0,1]$ such that $a_\ell \leq b_\ell$ for all $\ell \in \{1, \ldots, L\}$, where

$$\begin{aligned} \Delta_{a_\ell}^{b_\ell} C(u_1, \ldots, u_L) &:= C(u_1, \ldots, u_{\ell-1}, b_\ell, u_{\ell+1}, \ldots, u_L) \\ &\quad - C(u_1, \ldots, u_{\ell-1}, a_\ell, u_{\ell+1}, \ldots, u_L). \end{aligned}$$

Definition 3.2. Let $J_1, \ldots, J_L$ be subsets of $[0,1]$ containing at least the points 0 and 1. A function $C^* : J_1 \times \cdots \times J_L \to [0,1]$ is an L-dimensional subcopula or L-subcopula if it analogously satisfies the conditions (C1), (C2) and (C3) in Definition 3.1, but is defined on $J_1 \times \cdots \times J_L$ rather than on the entire L-cube $[0,1]^L$.

The field of copulas has been developing rapidly over the last decades, and copulas have been applied to a wide range of problems in various areas, such as climatology, meteorology and hydrology (Möller et al., 2013; Genest and Favre, 2007; Schoelzel and Friederichs, 2008; Zhang et al., 2012) or econometrics, insurance and mathematical finance (Cherubini et al., 2004; Embrechts et al., 2003; Pfeifer and Nešlehová, 2003; Genest et al., 2009), with Mikosch (2006) providing a critical review on their use. However, copulas are also of immense theoretical interest, due to their appealing mathematical properties. For a general overview of the mathematical theory of copulas, we refer to the textbooks by Joe (1997) and Nelsen (2006), as well as to the survey paper by Sempi (2011).

The relevance of copulas in multivariate dependence modeling is based on the famous theorem of Sklar (1959), which states that any multivariate CDF can be linked to its univariate marginal CDFs via a copula function.

For the purposes of this thesis, let us suppose that we have a predictive CDF F_ℓ for each univariate weather variable Y_ℓ for $\ell \in \{1, \ldots, L\}$, where the multi-index $\ell := (i, j, k)$ refers to weather quantity $i \in \{1, \ldots, I\}$, location $j \in \{1, \ldots, J\}$ and prediction horizon $k \in \{1, \ldots, K\}$, with $L := I \times J \times K$. Our goal is then to provide a physically realistic multivariate joint predictive CDF F with margins $F_1, \ldots, F_L$. With this setting in mind, Sklar's theorem can be formulated as follows.

Theorem 3.3. (Sklar) (Sklar, 1959; Nelsen, 2006)

1. For any multivariate CDF H with marginal CDFs $F_1, \ldots, F_L$, there exists a copula C such that
$$H(y_1, \ldots, y_L) = C(F_1(y_1), \ldots, F_L(y_L)) \tag{3.7}$$
for $y_1, \ldots, y_L \in \overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$. Moreover, C is uniquely determined if $F_1, \ldots, F_L$ are continuous; otherwise, C is uniquely determined on $\mathrm{Ran}(F_1) \times \cdots \times \mathrm{Ran}(F_L)$.

2. Conversely, if C is a copula and $F_1, \ldots, F_L$ are univariate CDFs, then the function H as defined in (3.7) is a multivariate CDF with margins $F_1, \ldots, F_L$.

Due to Sklar's theorem, the desired multivariate distribution can thus be constructed by combining margins obtained by univariate postprocessing and a multivariate dependence structure contained in a suitable copula function. Hence, the main challenge now is to choose and fit the copula C appropriately.

Various different types of copulas are available, including but not limited to Gaussian, elliptical, Archimedean, vine or pair, extremal and discrete or empirical copulas.

The most common type used in parametric approaches is that of a Gaussian copula (Embrechts et al., 2003), under which the joint multivariate CDF H is given by

$$H(y_1, \ldots, y_L|R) := \Phi_L(\Phi^{-1}(F_1(y_1)), \ldots, \Phi^{-1}(F_L(y_L))|R), \tag{3.8}$$

where $\Phi_L(\cdot|R)$ denotes the CDF of an L-dimensional normal distribution with mean vector $\mathbf{0} := (0, \ldots, 0)$ and correlation matrix R, and Φ^{-1} the inverse of the CDF of the univariate standard normal distribution. A major advantage of Gaussian copula models is that only the margins $F_1, \ldots, F_L$ and the correlation matrix R are required. Some examples for the use of Gaussian copulas in multivariate postprocessing are discussed in the next subsection, while parametric or semi-parametric alternatives comprise elliptical (Demarta and McNeil, 2005), Archimedean (Nelsen, 2006; McNeil and Nešlehová, 2009), extremal (Davison et al., 2012) and vine or pair copulas (Aas et al., 2009; Erhardt et al., 2014), for instance. All these types of copula are typically employed if the dimension L is rather small, or if a specific structure can be exploited.

By contrast, for large L, and if no specific structure can be utilized, it is reasonable to use non-parametric approaches, which rely on discrete or empirical copulas (Rüschendorf, 1976; Kolesárová et al., 2006; Rüschendorf, 2009), which are also known under the term "empirical dependence functions" (Deheuvels, 1979). Letting $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ with $\boldsymbol{x}_m := (x_m^1, \ldots, x_m^L) \in \mathbb{R}^L$ for $m \in \{1, \ldots, M\}$ be a data set of size M with values in $\mathbb{R}^L$ and assuming for simplicity that there are no ties, the corresponding empirical copula E_M is given by

$$E_M\left(\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right) := \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\{\text{rank}(x_m^1) \leq i_1, \ldots, \text{rank}(x_m^L) \leq i_L\}} \tag{3.9}$$

for integers $i_1, \ldots, i_L \in \{0, \ldots, M\}$, where $\text{rank}(x_m^\ell)$ denotes the rank of x_m^ℓ in $\{x_1^\ell, \ldots, x_M^\ell\}$ for $\ell \in \{1, \ldots, L\}$ and $m \in \{1, \ldots, M\}$. Detailed theoretical aspects of discrete and empirical copulas, respectively, are presented in Chapter 6. Examples of multivariate postprocessing approaches based on empirical copulas include the Schaake shuffle (Clark et al., 2004), which is discussed in the next subsection, and ensemble copula coupling (Schefzik et al., 2013), which forms the main subject of this thesis and is introduced in Chapter 4.

3.2.2 Examples of multivariate postprocessing methods

Now we discuss some examples of multivariate postprocessing methods and point out relationships to copulas.

(a) Addressing spatial dependence structures via parametric methods

Spatial BMA and Spatial EMOS

To account for spatial dependence patterns, an established strategy is to fuse univariate ensemble postprocessing techniques on the one hand and geostatistical models from spatial statistics on the other hand. Examples of multivariate postprocessing methods which are able to address spatial dependence structures in the case of temperature include Spatial BMA (Berrocal et al., 2007) and Spatial EMOS (Feldmann et al., 2014). These two

approaches combine the standard BMA and EMOS notions for univariate ensemble postprocessing, respectively, with the geostatistical output perturbation (GOP) technique proposed by Gel et al. (2004), which models the spatial correlation structure. The final output is a mixture of multivariate normal distributions in the case of Spatial BMA and a single multivariate normal distribution in the case of Spatial EMOS. Thus, an interpretation of the two approaches in the Gaussian copula setting of Equation (3.8) is available, where in this spatial setting, the correlation matrix R is taken to be highly structured and satisfies requirements auch as spatial stationarity and/or isotropy. In general, the use of Gaussian copula techniques is very common in geostatistics, where this notion is referred to as anamorphosis (Chilès and Delfiner, 2012). A counterpart of GOP which can be applied to precipitation fields is the two-stage spatial (TSS) model introduced by Berrocal et al. (2008). Ideas and first attempts for a combination of the TSS model with BMA for precipitation (Sloughter et al., 2007) can be found in Scheuerer and Gneiting (2011), while a combination of the TSS model with an appropriate EMOS method has not been implemented up to now to my knowledge.

(b) Addressing inter-variable dependence structures via parametric methods

BMA and Gaussian copulas

Möller et al. (2013) propose the use of Gaussian copulas to retain the inter-variable dependence pattern of forecasts at individual locations and for fixed prediction horizons, where BMA is employed to generate the postprocessed marginal predictive distributions for each weather quantity separately. The method is an application of Gaussian copulas except that precipitation forecasts require special treatment due to the point mass at zero.

Joint postprocessing of temperature and wind speed using BMA

The approach of Baran and Möller (2014) uses a bivariate BMA model based on bivariate truncated normal distributions for the joint postprocessing of temperature and wind speed. It is a direct extension of the univariate BMA methods for temperature using normal distributions (Raftery et al., 2005) and wind speed using truncated normal distributions (Baran, 2014), in that the joint modeling of the two weather variables is performed via a bivariate normal BMA kernel with wind coordinate truncated from below at zero, as wind speed is a non-negative quantity.

Postprocessing of (u, v)-wind vectors

One of the most common and important examples of handling inter-variable dependence only is the joint postprocessing of u- and v-wind. To this end, Sloughter et al. (2013) propose a BMA variant for bivariate wind vectors using mixtures of suitably transformed bivariate normal densities, where each mixture component can be linked to a Gaussian copula, allowing for a copula interpretation of this approach.

In view of the further development of the thesis, we concentrate on discussing the alternative approaches of Schuhen et al. (2012) and Pinson (2012) in what follows.

To model dependence patterns between u- and v- wind at a fixed location and for a fixed look-ahead time, Schuhen et al. (2012) propose an EMOS approach for wind vectors, in

which the postprocessed predictive distribution is bivariate normal, taking the form

$$\mathcal{N}_2\left(\begin{pmatrix} \mu_u \\ \mu_v \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}\right)$$

with predictive PDF

$$\begin{aligned} f(u,v) &:= \frac{1}{2\pi\sigma_u\sigma_v\sqrt{1-\rho^2}} \\ &\quad \times \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(u-\mu_u)^2}{\sigma_u^2} - 2\rho\frac{(u-\mu_u)(v-\mu_v)}{\sigma_u\sigma_v} + \frac{(v-\mu_v)^2}{\sigma_v^2}\right)\right). \end{aligned}$$

In the method of Schuhen et al. (2012) for ensembles consisting of exchangeable members, the predictive means $\mu_u = a_u + b_u\bar{u}$ and $\mu_v = a_v + b_v\bar{v}$ are linearly bias-corrected versions of the empirical means $\bar{u}$ and $\bar{v}$ of the u- and v-wind raw ensemble forecasts $u_1, \ldots, u_M$ and $v_1, \ldots, v_M$, respectively. Moreover, the predictive variances $\sigma_u^2 = c_u + d_u s_u^2$ and $\sigma_v^2 = c_v + d_v s_v^2$ are linear functions of the empirical raw ensemble variances s_u^2 and s_v^2, respectively. In this context, the parameters $a_u, a_v, b_u, b_v, c_u, c_v, d_u$ and d_v have to be estimated from training data. Finally, the predictive correlation coefficient

$$\rho := r\cos\left(\frac{2\pi}{360}(\nu\theta + \psi)\right) + q \tag{3.10}$$

is considered to be a trigonometric function of the mean ensemble wind direction θ, where the coefficients q and r concern the overall magnitude of ρ and have to satisfy $|q| + |r| \leq 1$. Moreover, while ψ encodes phase information, $\nu \in \mathbb{N}$ corresponds to the number of periods of the trigonometric function. In Schuhen et al. (2012), ν is constrained to be either 1, 2 or 3. In view of the Gaussian copula frame fixed in Equation (3.8), the method of Schuhen et al. (2012) deals with the special case of $L = 2$ and invokes a bivariate normal predictive distribution H with univariate normal margins F_1 and F_2 for u- and v-wind, respectively.

Pinson (2012) proposes a bivariate postprocessing method for (u, v)-wind vectors which directly transfers a u- and v-wind raw ensemble forecast $u_1, \ldots, u_M$ and $v_1, \ldots, v_M$, respectively, at a fixed location and for a fixed prediction horizon into a calibrated ensemble via a member-by-member postprocessing approach, instead of fitting a full predictive postprocessed PDF as in Schuhen et al. (2012). Essentially, Pinson (2012) assumes that the raw ensemble members sample a bivariate Gaussian distribution and introduces a two-dimensional translation and dilation of the sets of raw ensemble predictions. The translation and dilation factors are obtained by models for the mean and the variance of the bivariate Gaussian densities, while Pinson (2012) does not consider a potential correction of the correlation coefficient, as opposed to Schuhen et al. (2012). Precisely, the new calibrated u- and v-wind ensemble members $\hat{u}_m$ and $\hat{v}_m$ in Pinson's approach are given by

$$\begin{aligned} \hat{u}_m &:= \bar{u} + \tau_u + \xi_u(u_m - \bar{u}), \\ \hat{v}_m &:= \bar{v} + \tau_v + \xi_v(v_m - \bar{v}) \end{aligned}$$

for $m \in \{1, \ldots, M\}$. In this context, u_m and v_m denote the raw ensemble u- and v-wind forecasts, with empirical means $\bar{u}$ and $\bar{v}$, respectively. Moreover, τ_u and τ_v are certain translation factors and ξ_u and ξ_v certain dilation factors, which depend on $\bar{u}$ and $\bar{v}$ and the empirical raw ensemble standard deviations s_u and s_v, as well as on model parameters. The work of Pinson (2012) can be interpreted as a special case of the ensemble copula coupling technique (Schefzik et al., 2013), which turns out to be an empirical copula method and is introduced in the next chapter. Relationships are discussed in detail in Chapter 5.

(c) Addressing temporal dependence structures via parametric methods

Pinson et al. (2009) and Schoelzel and Hense (2011) use Gaussian copulas to capture dependence structures over consecutive prediction horizons in postprocessed predictive distributions.

(d) Addressing dependence structures via the Schaake shuffle as a non-parametric method

The multivariate parametric approaches presented before work particularly well in low dimensions or if specific structure can be exploited. However, these parametric methods are bound to fail in high-dimensional settings. In contrast, the Schaake shuffle (Clark et al., 2004) offers an alternative non-parametric method for reconstructing physically coherent spatio-temporal structures. It was introduced by Clark et al. (2004) as a reordering technique and has been interpreted by Schefzik et al. (2013) as an empirical copula approach as follows.

For a given weather quantity $i \in \{1, \dots, I\}$ at a given location $j \in \{1, \dots, J\}$, summarized in the multi-index $\ell^* := (i, j)$, let $\tilde{x}_1^{\ell^*}, \dots, \tilde{x}_N^{\ell^*}$ be a postprocessed N-member ensemble forecast valid on date t, with $L^* := I \times J$ being the dimension of the overall output of a single ensemble member forecast. Here, we think of $\tilde{x}_1^{\ell^*}, \dots, \tilde{x}_N^{\ell^*}$ as postprocessed ensemble forecasts in the form of samples from corresponding univariate predictive distributions $F_1, \dots, F_L$ at each location for each weather quantity separately, which could be obtained via BMA or EMOS, for instance. Moreover, let $y_1^{\ell^*}, \dots, y_N^{\ell^*}$ denote corresponding selected historical observations from N dates $t_1, \dots, t_N$ in the past of t. In this context, the same N dates are employed for all locations and weather variables throughout the whole procedure of the Schaake shuffle. Following the specific implementation of Clark et al. (2004), the dates $t_1, \dots, t_N$ can be chosen from the whole historical record, except for the year of the forecast of interest, and should lie within seven days before and after the date t, regardless of the year. For each ℓ^*, the corresponding order statistics $y_{(1)}^{\ell^*} \leq \dots \leq y_{(N)}^{\ell^*}$ then induce a permutation π_{ℓ^*} of $\{1, \dots, N\}$ defined by $\pi_{\ell^*}(n) := \text{rank}(y_n^{\ell^*})$ for $n \in \{1, \dots, N\}$, with any ties resolved at random. The final postprocessed Schaake shuffle ensemble $\hat{x}_1^{\ell^*}, \dots, \hat{x}_N^{\ell^*}$ is then given by $\hat{x}_1^{\ell^*} := \tilde{x}_{(\pi_{\ell^*}(1))}^{\ell^*}, \dots, \hat{x}_N^{\ell^*} := \tilde{x}_{(\pi_{\ell^*}(N))}^{\ell^*}$ for each $\ell^* \in \{1, \dots, L^*\}$. Hence, the postprocessed ensemble forecasts are just reordered with respect to the rank dependence structure of the selected historical observations. From a copula-based point of view, the empirical copula defined by the past observations according to (3.9) is applied to the postprocessed ensemble forecasts, such that both the historical materializations and the final forecasts reveal the same multivariate rank dependence pattern. A detailed study of empirical and discrete copulas, respectively, is given in Chapter 6, where also their relationships to the Schaake shuffle and the ensemble copula coupling (ECC) approach are explicitly discussed.

Even though the Schaake shuffle has performed well in meteorological and hydrologic applications (Clark et al., 2004; Schaake et al., 2007; Voisin et al., 2011; Vrac and Friederichs, 2014), a major shortcoming of its standard variant is the failure to condition the multivariate dependence pattern on current or predicted atmospheric states. This could be addressed by combining the notion of the Schaake shuffle with the search for so-called analog ensembles and is further investigated in Chapter 8.

Chapter 4

Ensemble copula coupling (ECC)

As noted, the parametric multivariate postprocessing methods presented in Section 3.2.2 are not suitable when dealing with very high-dimensional settings. Moreover, they are capable to model either spatial or inter-variable or temporal dependencies, but combinations of all types of dependence are not addressed. As seen, the Schaake shuffle (Clark et al., 2004) as an empirical copula approach can help to resolve this shortcoming. However, it does not make use of the multivariate dependence structure of the NWP ensemble.

In this chapter, we introduce and illustrate an empirical copula method, called ensemble copula coupling (ECC) (Schefzik et al., 2013), that is related to the Schaake shuffle, but in which the multivariate dependence structure of the postprocessed forecast is adopted from the original unprocessed NWP ensemble, rather than from historical observations, thereby capturing the atmospheric flow dependence. Moreover, quantization aspects in the context of ECC are investigated, and ECC is tested and assessed in a comprehensive case study using the ECMWF ensemble. Finally, we discuss modifications of ECC, which are able to handle some of the limitations arising in the standard ECC implementation.

4.1 The ensemble copula coupling (ECC) approach

In this section, we present and illustrate the standard ensemble copula coupling (ECC) approach (Schefzik et al., 2013) as a multivariate ensemble postprocessing tool. ECC has been originally hinted at by Bremnes (2007) and Krzysztofowicz and Toth (2008), and initial investigations have been conducted in Schefzik (2011), before Schefzik et al. (2013) introduced the general ECC frame exposed in what follows.

Essentially, ECC uses both univariate postprocessing methods and the rank order information available in the NWP raw ensemble forecast to produce calibrated and sharp, physically coherent spatio-temporal weather trajectories. As NWP models are based on discretizations of the equations governing the physics of the atmosphere, observed multivariate dependence structures can be expected to be reasonably well represented in the raw ensemble prediction system. However, it is generally advisable to make diagnostic checks in order to verify whether dependence structures within the raw ensemble are compatible with historical observations. For the ECMWF ensemble data set employed in the case studies throughout this thesis, the resemblance of the corresponding dependence patterns mostly turns out to be sufficiently large, in particular in the case of pressure.

Assuming that the raw ensemble members are exchangeable, ECC is a multi-stage approach, proceeding as follows (Schefzik et al., 2013).

1. *Obtaining an NWP raw ensemble:* Initialized at a fixed date, multiple runs of an NWP model differing in the inputs or model parameters in suitable ways supply an M-member ensemble forecast, where each member attains values in $\mathbb{R}^L$. As an output, we are thus given the univariate margins $x_1^\ell, \ldots, x_M^\ell$ of the raw ensemble, with the multi-index $\ell := (i, j, k)$ referring to weather variable $i \in \{1, \ldots, I\}$, location $j \in \{1, \ldots, J\}$ and prediction horizon $k \in \{1, \ldots, K\}$, where $L := I \times J \times K$.

2. *Derivation of the univariate order statistics of the raw ensemble:* For each margin ℓ, compute the order statistics $x_{(1)}^\ell \leq \ldots \leq x_{(M)}^\ell$ of the raw ensemble values. These induce a permutation σ_ℓ of the integers $\{1, \ldots, M\}$ via $\sigma_\ell(m) := \text{rank}(x_m^\ell)$ for $m \in \{1, \ldots, M\}$. If there are ties among the ensemble values, they are resolved at random, which is a natural procedure. Other allocation methods are feasible and do not entail major technical problems.

3. *Univariate postprocessing:* For each margin ℓ, apply a univariate postprocessing method, such as BMA or EMOS as presented in Section 3.1, to the raw ensemble output $x_1^\ell, \ldots, x_M^\ell$ and obtain a univariate postprocessed predictive CDF, F_ℓ.

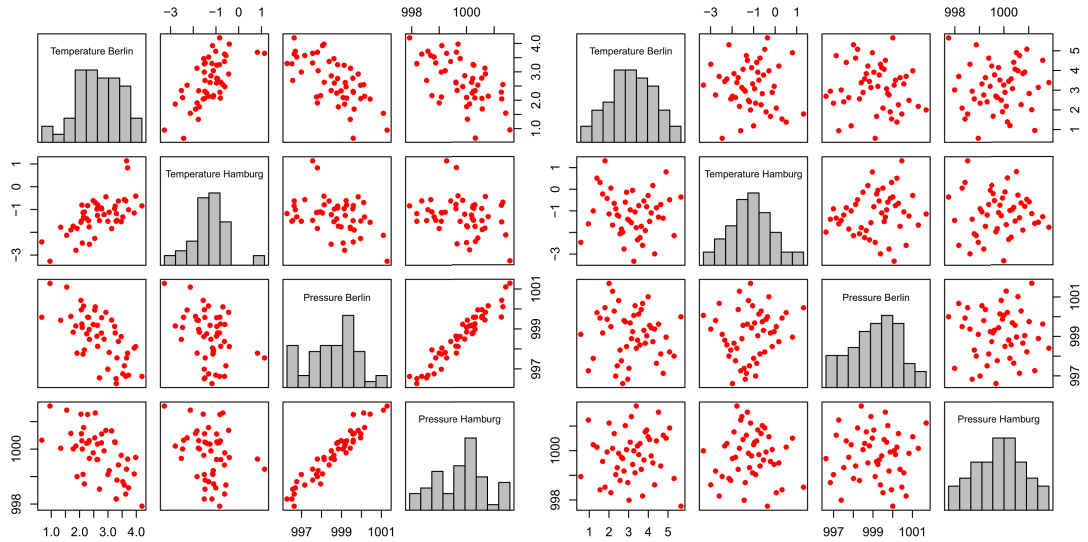
4. *Quantization/Sampling:* Represent each univariate postprocessed predictive distribution F_ℓ by a discrete sample $\tilde{x}_1^\ell, \ldots, \tilde{x}_M^\ell$ of size M. This sample can be generated in various ways, as will be described and discussed in Section 4.2, where we consider the sampling methods (R), (T) and (Q) referring to samples obtained by employing a Random quantization, a Transformation approach or equidistant Quantiles, respectively.

5. *Reordering:* For each margin ℓ, the final postprocessed ECC ensemble $\hat{x}_1^\ell, \ldots, \hat{x}_M^\ell$ is given by
$$\hat{x}_1^\ell := \tilde{x}_{(\sigma_\ell(1))}^\ell, \ldots, \hat{x}_M^\ell := \tilde{x}_{(\sigma_\ell(M))}^\ell.$$
That is, the sample from each univariate marginal postprocessed distribution is reordered with respect to the rank dependence structure of the raw ensemble, thereby preserving the flow dependence.

ECC in the form presented above is basically a reordering technique. The necessity of this reordering notion to capture the rank dependence pattern of the raw ensemble is not new and has been recognized in the extant literature (Bremnes, 2007; Flowerdew, 2012, 2014, for instance). Alternatively, member-by-member postprocessing approaches (Johnson and Bowler, 2009; Pinson, 2012, among others) directly retain the rank dependence pattern of the raw ensemble by construction, such that the reordering step becomes obsolete. Examples of such approaches, which turn out to be special cases of ECC, will be discussed in Chapter 5.

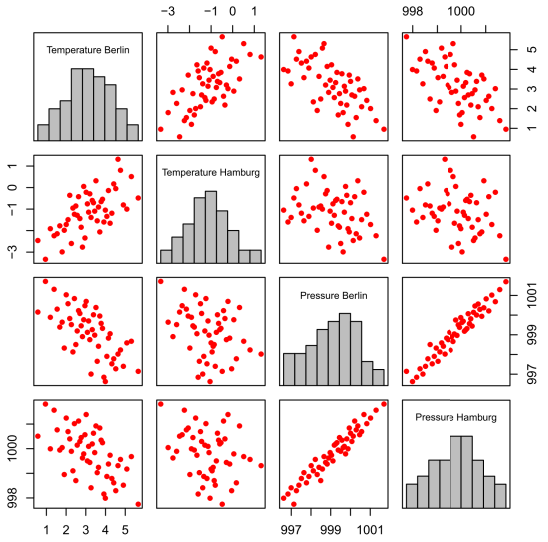
However, ECC can alternatively be interpreted as an empirical copula approach (Schefzik et al., 2013), in that the empirical copula defined by the raw ensemble is applied to the discrete samples from the univariate postprocessed marginal distributions to obtain the ECC ensemble. This justifies the term "ensemble copula coupling". Details in this context will be provided in Chapter 6.

The basic notion of ECC is illustrated for an $L = 4$-dimensional setting in Figure 4.1,



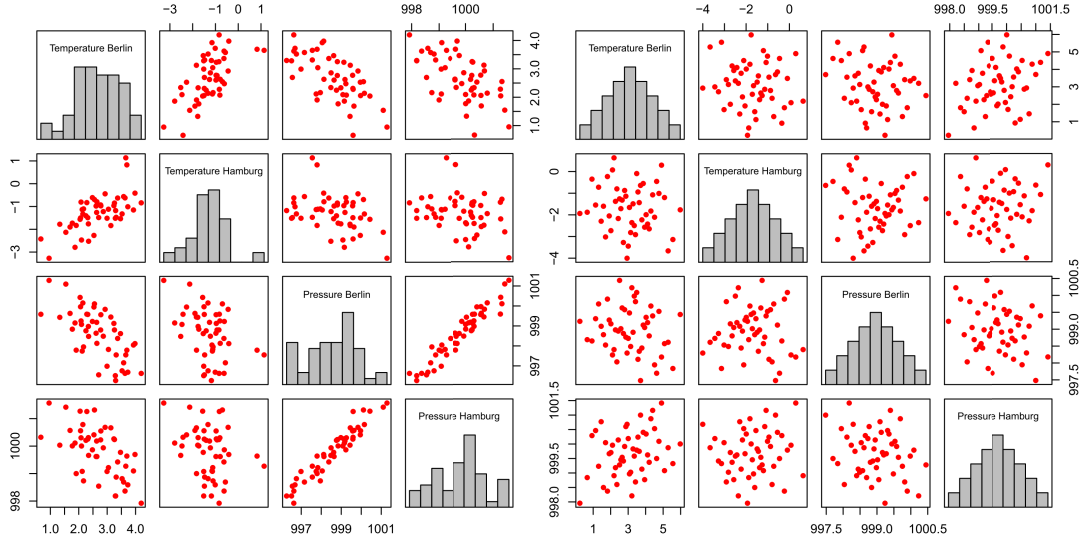
(a) ECMWF raw ensemble

(b) Individual BMA postprocessing



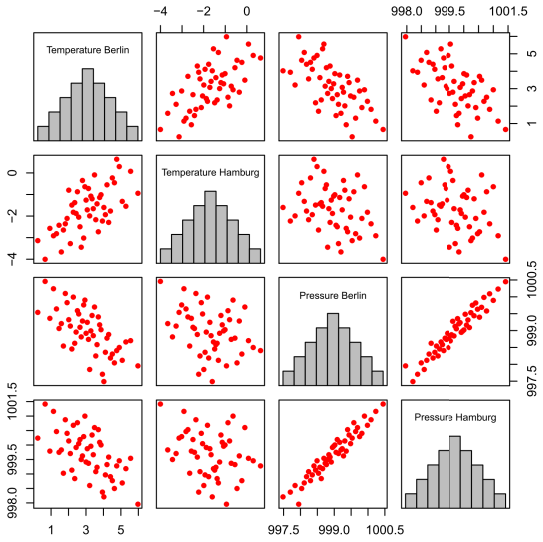
(c) ECC postprocessed ensemble

Figure 4.1: 24 hour ahead ensemble forecasts of temperature (in °C) and pressure (in hPa) at Berlin and Hamburg, valid 1:00 am on 8 November 2010. The 50-member ECMWF raw ensemble is shown in (a), an individually BMA postprocessed ensemble in (b) and the ECC postprocessed ensemble in (c). Similar plots valid for a different date are shown in Figure 5 in Schefzik et al. (2013).



(a) ECMWF raw ensemble

(b) Individual EMOS postprocessing



(c) ECC postprocessed ensemble

Figure 4.2: 24 hour ahead ensemble forecasts of temperature (in °C) and pressure (in hPa) at Berlin and Hamburg, valid 1:00 am on 8 November 2010. The 50-member ECMWF raw ensemble is shown in (a), an individually EMOS postprocessed ensemble in (b) and the ECC postprocessed ensemble in (c).

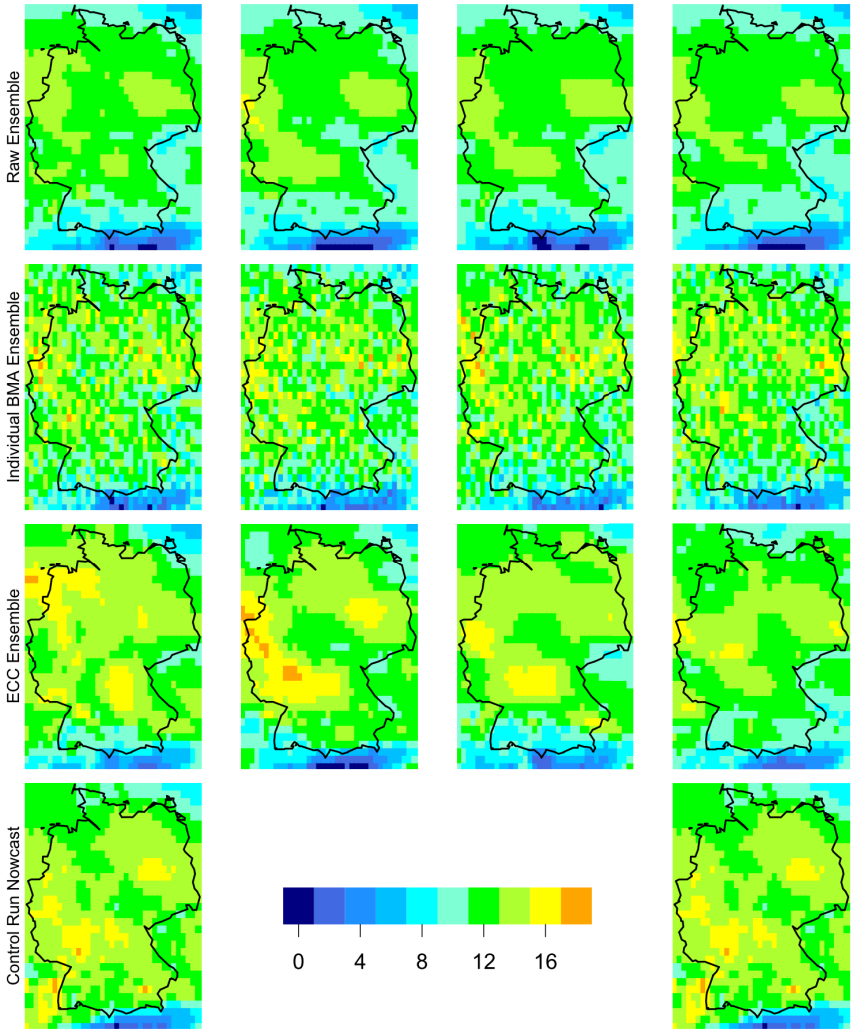


Figure 4.3: 24 hour ahead ensemble forecasts of temperature (in °C) on a grid over Germany, valid 2:00 am on 23 April 2011. First row: Four selected members of the ECMWF raw ensemble. Second row: Four members obtained by individual site-by-site BMA postprocessing, where for each grid point, a random number from the respective BMA postprocessed predictive distribution is drawn. Third row: The four members of the respective ECC ensemble with rank order structures inherited from the corresponding ECMWF raw ensemble members in the first row. Fourth row: The single-valued control run nowcast, shown both at left and at right. A similar overview plot valid for a different date is shown in Figure 11 in Schefzik et al. (2013).

where we consider three different forecast ensembles for temperature and pressure at Berlin and Hamburg, valid at 1:00 am on 8 November 2010. The scatterplot matrix in 4.1 (a) shows the 50-member ECMWF raw ensemble forecasts, which clearly reveal dependencies between the margins. For instance, there is a strong positive correlation between pressure in Berlin and pressure in Hamburg, as well as a positive correlation between temperature in Berlin and temperature in Hamburg, while a negative association can be observed between temperature in Berlin and pressure in Berlin. The scatterplot matrix in 4.1 (b) consists of samples of the individual predictive distributions obtained by univariate postprocessing via BMA, using a sliding training window of 30 days. While biases and dispersion errors in the margins are corrected for, the original dependence structures from the raw ensemble in 4.1 (a) get completely lost. Finally, the scatterplot matrix in 4.1 (c) shows the effect of the ECC reordering approach, in that the rank dependence pattern of the unprocessed raw ensemble is retained, with the margins remaining unchanged from 4.1 (b), as is indicated by the coincidence of the histograms of the margins on the diagonal of the scatterplot matrices 4.1 (b) and 4.1 (c).

The descriptions and comments just made hold analogously for Figure 4.2, which works exactly in the same setting as Figure 4.1, except that the univariate postprocessing is performed via EMOS instead of BMA in this case. Hence, the marginals are given by single normal distributions, as indicated by the histograms on the diagonals in 4.2 (b) and 4.2 (c), rather than mixtures of normal distributions.

Another example for the positive effect of ECC in spatial settings is given in Figure 4.3, dealing with ensemble temperature field forecasts valid at 2:00 am on 23 April 2011, made on a grid over Germany and adjacent areas comprising $33 \times 37 = 1221$ model grid points. The first row displays the temperature field predictions of four selected members from the ECMWF raw ensemble. In the second row, four realizations of individual univariate BMA postprocessing are shown, where for each grid point, a random number from the corresponding BMA predictive distribution is drawn. The third row shows those four members of the corresponding postprocessed ECC ensemble which adopt their dependence structure from the respective ECMWF raw ensemble members in the first row. The ground truth on the grid is given by the corresponding single-valued nowcast initialization of the ECMWF control run, as described in Section 1.3, and is displayed twice in the fourth row. While they appear to capture spatial patterns quite well, the ECMWF raw ensemble members exhibit an overall negative bias, in particular in the Alps region in the south and in Southwest Germany, respectively. The BMA postprocessing, using a training period of 20 days here, is able to correct this shortcoming, where the respective BMA parameters are estimated once and for all based on data of all grid points, thus avoiding inconsistencies between the univariate postprocessed predictive distributions themselves. However, the individual sampling leads to spatially noisy and inconsistent temperature fields, where the spatial pattern of the ECMWF raw ensemble member forecasts vanishes. This missing spatial structure is then restored by the corresponding ECC ensemble members, which both inherit the bias-corrected marginals from the individual BMA postprocessed forecasts and account for the dependencies within the raw ensemble.

For the same setting as in Figure 4.3, plots of all 50 members of the raw, individual BMA and ECC ensemble, respectively, are shown in Figures 4.4 to 4.6 as a supplement. The 50 single plots within each figure are arranged in a consistent manner with respect to the respective ensemble member indices, thus allowing for a direct comparison.

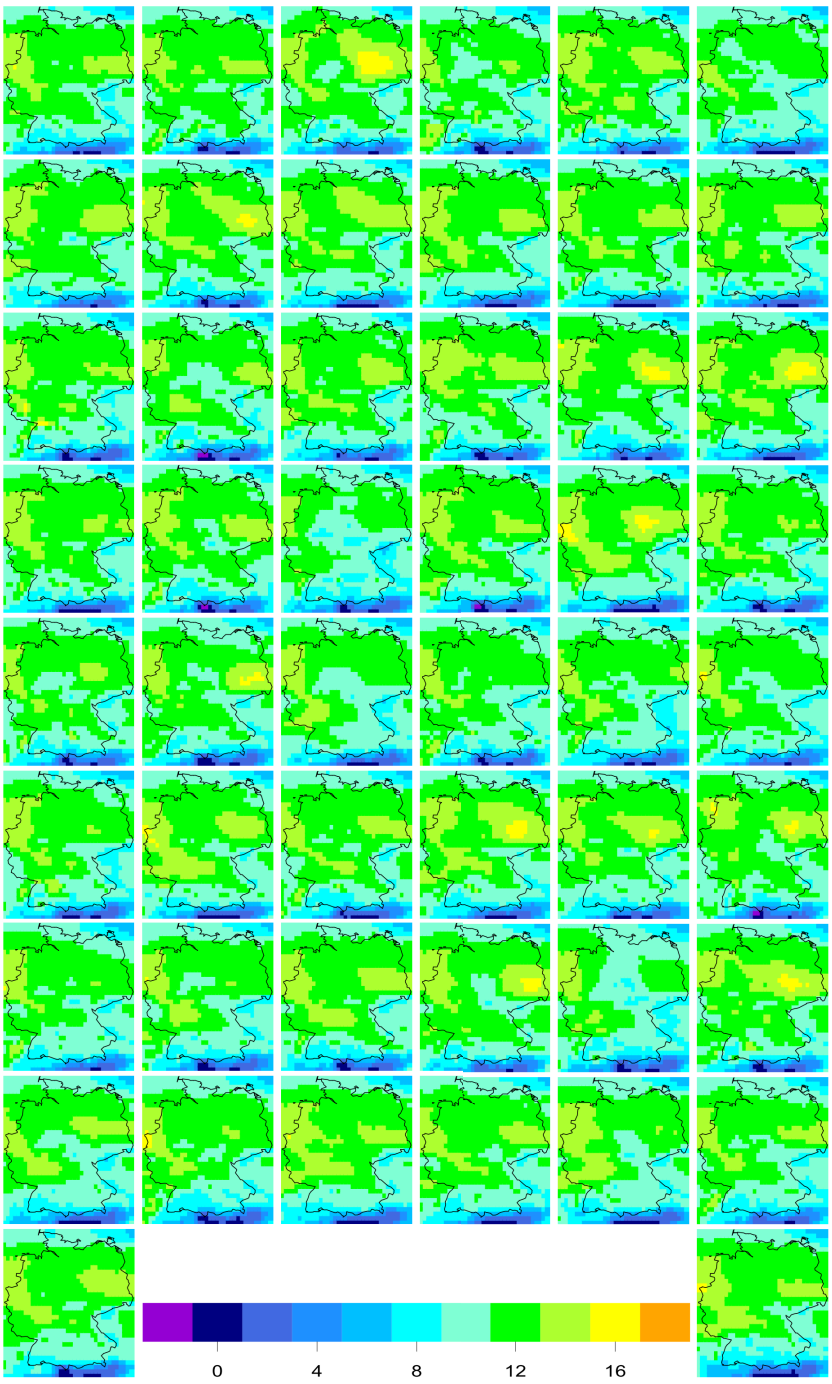


Figure 4.4: 24 hour ahead 50-member ECMWF raw ensemble forecast for temperature (in °C) on a grid over Germany, valid 2:00 am on 23 April 2011.

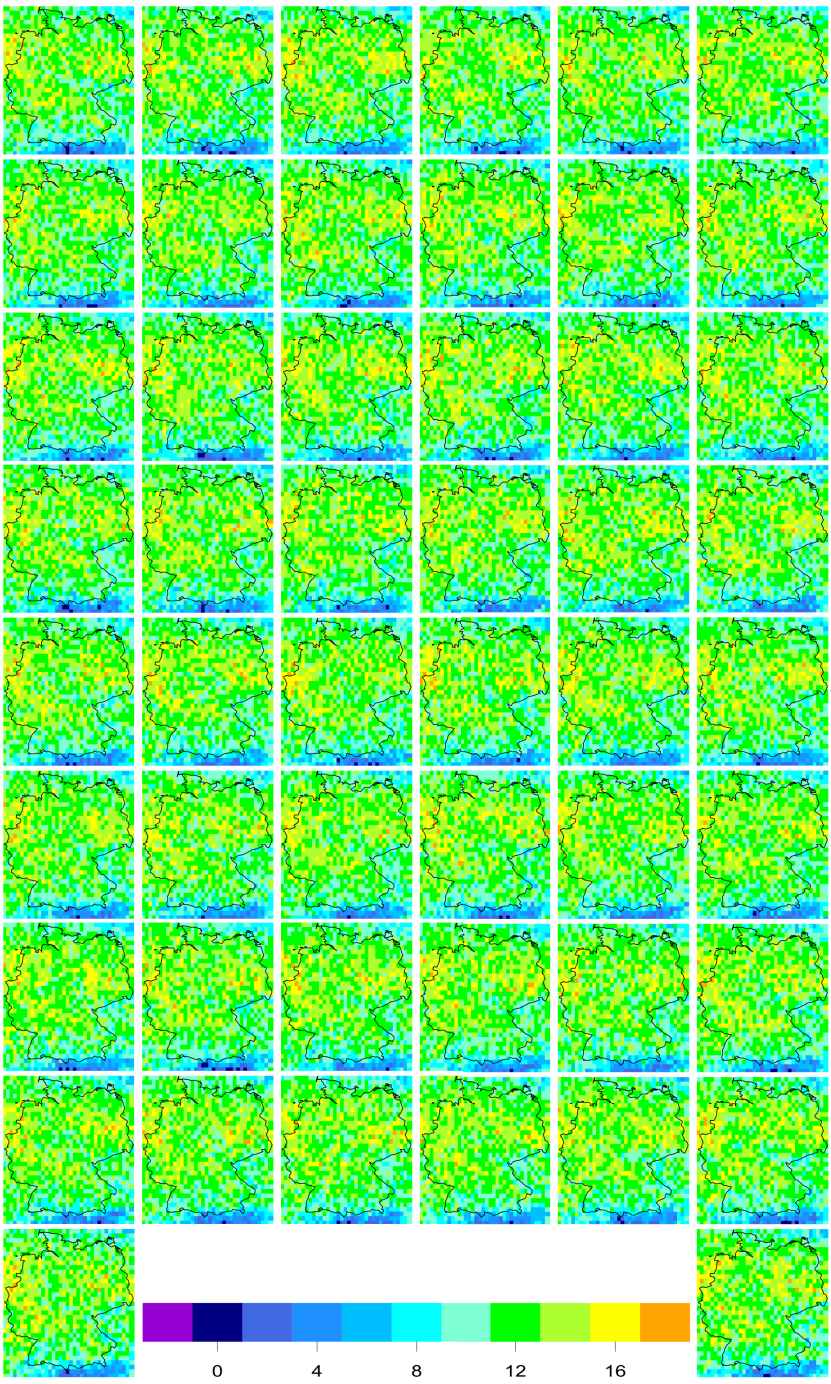


Figure 4.5: 24 hour ahead 50-member individual BMA ensemble forecast for temperature (in °C) on a grid over Germany, valid 2:00 am on 23 April 2011.

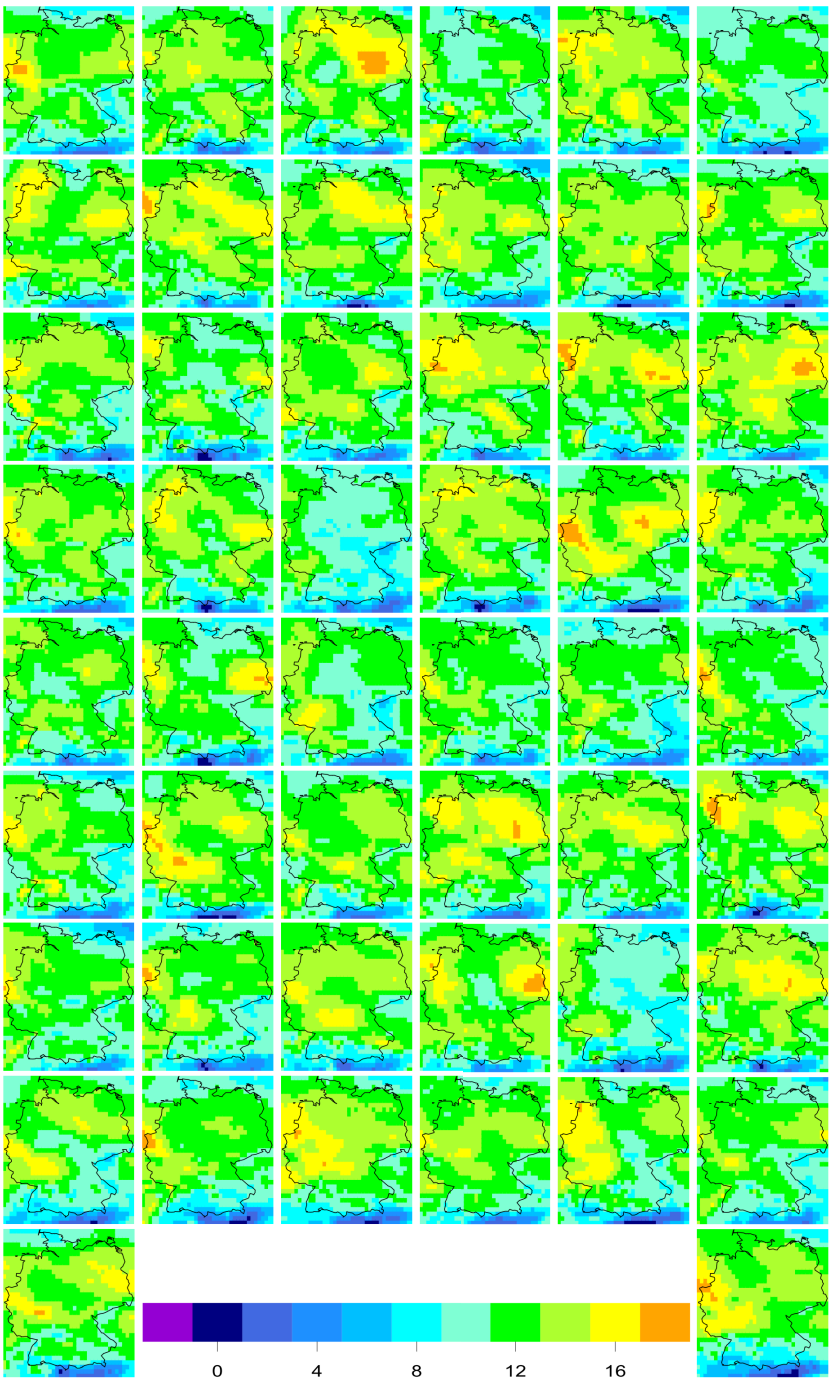


Figure 4.6: 24 hour ahead 50-member ECC ensemble forecast for temperature (in °C) on a grid over Germany, valid 2:00 am on 23 April 2011.

The ECC approach, which combines parametric univariate ensemble postprocessing and a non-parametric reordering notion, has a couple of benefits. It retains the marginal predictive distributions obtained by any univariate postprocessing method, while inheriting the rank dependence structure from the unprocessed raw NWP ensemble, thus capturing the flow dependence and retaining the bivariate Spearman rank correlation coefficients of the raw ensemble. Hence, ECC combines analytic, numerical and statistical modeling and embodies a universal concept which is likely to be broadly applicable in the context of uncertainty quantification, also in application areas apart from weather forecasting. Moreover, it is comparably easy to explain, understand and implement and can be applied in settings of any, potentially high, dimension, in which parametric methods are likely to fail. Once an NWP model has been run and univariate statistical postprocessing has been performed, the crucial ECC reordering step to account for multivariate dependencies requires negligible additional computational effort and essentially comes for free. Last but not least, ECC has a mathematical-theoretical backing, being embedded in the frame of discrete copulas, as will be discussed in Chapter 6.

ECC appears to be especially valuable if the raw ensemble is rather large, as is the case with the 50-member ECMWF ensemble considered in this thesis, such that much information can be drawn from it. In contrast, ECC might be less beneficial if the underlying raw ensemble size is comparably small, or if there are many ties among the marginal raw ensemble values, as might be the case for precipitation, with several ensemble members simultaneously predicting zero precipitation, for instance. In these situations, one cannot learn much about the dependence structure from the raw ensemble, and thus, applying ECC is expected to be not that powerful. Apart from this, ECC in the form presented before has further limitations. It cannot correct any inconsistencies between the postprocessed marginal distributions themselves. Moreover, ECC assumes a perfect model as far as the multivariate rank dependence structure across weather quantities, locations and prediction horizons is concerned. For NWP models, this assumption typically appears reasonable and can be verified empirically via diagnostic checks. However, it most likely does not hold each and every day, and hence, it might be more realistic to suppose that NWP models may exhibit errors in dependence patterns, which should be diagnosed and improved by using the dependence structures of the observations, which are not accounted for in the present ECC version. Thus, alternative parametric postprocessing methods are likely to outperform ECC in low-dimensional or highly structured settings, given that there is enough training data available to statistically correct dependence patterns. This is for instance witnessed by the case studies in Schuhen et al. (2012), in which the bivariate EMOS approach presented in Section 3.2.2 outperforms ECC in an $L = 2$-dimensional scenario considering u- and v-wind jointly at individual locations and for a fixed look-ahead time of 48 hours. In addition, ECC can only be applied to ensembles with exchangeable members, and the postprocessed ECC ensemble is constrained to have the same number of members as the original NWP ensemble, which is typically rather small. Part of these shortcomings will be addressed later in this thesis by proposing both suitable modifications of the current ECC approach and a related technique in Sections 4.4 and 4.5 of this chapter and in Chapter 8, respectively.

4.2 The quantization step

An essential property of the ECC postprocessing method is that its final output takes the form of an ensemble. Hence, an important part within ECC is the quantization or sampling step, in which each univariate marginal predictive CDF F_ℓ obtained by an ensemble postprocessing method is discretized, such that each F_ℓ is represented by a sample $\tilde{x}_1^\ell, \ldots, \tilde{x}_M^\ell$ from F_ℓ. While various procedures are possible to achieve this, we describe and investigate three different methods in this section here, which we call the sampling methods (R), (T) and (Q), as the samples are obtained by using Random draws, a Transformation approach and equally spaced Quantiles, respectively. In addition, the different sampling approaches are tested and compared in a case study.

4.2.1 The sampling methods (R), (T) and (Q)

The sampling methods hinted at above are designed as follows.

(a) The sampling method (R)

A simple option to obtain a sample $\tilde{x}_1^\ell, \ldots, \tilde{x}_M^\ell$ from a univariate predictive CDF F_ℓ is to just draw random samples from F_ℓ of the form

$$\tilde{x}_1^\ell := F_\ell^{-1}(u_1), \ldots, \tilde{x}_M^\ell := F_\ell^{-1}(u_M), \tag{R}$$

with $u_1, \ldots, u_M$ denoting independent standard uniform random variates, and F_ℓ^{-1} the (generalized) inverse of F_ℓ.
We refer to the corresponding ECC variant using this sampling method (R) as ECC-R.

(b) The sampling method (T)

Alternatively, a sample $\tilde{x}_1^\ell, \ldots, \tilde{x}_M^\ell$ from a univariate predictive CDF F_ℓ can be obtained via the following quantile mapping or transformation approach, which bases on the ensemble smoothing method of Wilks (2002). Let S_ℓ be a parametric, continuous CDF fitted to the raw ensemble margin $x_1^\ell, \ldots, x_M^\ell$. Then, we use those quantiles from F_ℓ as a sample $\tilde{x}_1^\ell, \ldots, \tilde{x}_M^\ell$ which correspond to the percentiles of the raw ensemble values in S_ℓ, that is,

$$\tilde{x}_1^\ell := F_\ell^{-1}(S_\ell(x_1^\ell)), \ldots, \tilde{x}_M^\ell := F_\ell^{-1}(S_\ell(x_M^\ell)). \tag{T}$$

We refer to the corresponding ECC variant using this sampling method (T) as ECC-T.
Due to its design, the sampling scheme (T) automatically retains the rank dependence structure of the raw ensemble, such that the reordering step 5 in the standard ECC implementation described in Section 4.1 becomes obsolete.

In the special case of S_ℓ and F_ℓ belonging to the same location-scale family, the transformation from the raw ensemble value x_m^ℓ to the postprocessed sample value $\tilde{x}_m^\ell$ for each fixed ℓ and for member $m \in \{1, \ldots, M\}$ becomes affine, which is manifested by the following lemma, where we omit the indices ℓ and m for convenience.

Lemma 4.1. Let S and F be two continuous CDFs that belong to the same location-scale family. Then, the transformation from x to $\tilde{x} := F^{-1}(S(x))$ is affine.

Proof. As the continuous CDFs S and F belong to the same location-scale family, there exists a continuous CDF G such that $S(x) = G((x-\mu)/\sigma)$ and $F(x) = G((x-\mu^*)/\sigma^*)$, where $\mu, \mu^* \in \mathbb{R}$ and $\sigma, \sigma^* \in \mathbb{R}_+$. By using this location-scale property (LSP) twice, we get

$$\begin{aligned}\tilde{x} &= F^{-1}(S(x)) \overset{\text{LSP}}{=} F^{-1}\left(G\left(\frac{x-\mu}{\sigma}\right)\right) \overset{\text{LSP}}{=} \left[G^{-1}\left(G\left(\frac{x-\mu}{\sigma}\right)\right)\right]\sigma^* + \mu^* \\ &= \frac{(x-\mu)}{\sigma}\,\sigma^* + \mu^* = \mu^* + \frac{\sigma^*}{\sigma}(x-\mu).\end{aligned}$$

Hence, the transformation from x to $\tilde{x}$ is affine and has the form $\tilde{x} = \mu^* + \frac{\sigma^*}{\sigma}(x-\mu)$. □

In the special setting of Lemma 4.1, ECC-T thus retains the raw ensemble's bivariate Pearson product moment correlation coefficients, in addition to preserving its bivariate Spearman rank correlation coefficients.

In many cases, such as for temperature, pressure, u- and v-wind, S_ℓ can be assumed to be normal, with mean equal to the empirical raw ensemble mean and variance equal to the empirical raw ensemble variance. When combining such a normal CDF S_ℓ with a univariate postprocessed normal CDF F_ℓ obtained by EMOS for the suitable weather variables, Lemma 4.1 can be applied, and we have a member-by-member postprocessing method based on a monotone transformation. Such approaches are further discussed in Chapter 5, and indeed many methods proposed in the extant literature turn out to be special implementations of the ECC-T type.

The ECC-T quantization scheme can principally be applied also in non-Gaussian settings. Exemplarily, we address precipitation in what follows, and describe how to choose F_ℓ and S_ℓ in scheme (T) then.

In the case of precipitation, the univariate predictive CDF F_ℓ can be modeled by the standard BMA and EMOS postprocessing approaches, as discussed in Section 3.1. We could for instance take F_ℓ to be the CDF of a mixture of Gamma distributions with point mass at zero, as obtained by the corresponding BMA approach of Sloughter et al. (2007). Alternatively, F_ℓ could be considered to be the CDF of a GEV distribution left-censored at zero, as in the respective EMOS approach of Scheuerer (2014). In contrast, the choice of a suitable CDF S_ℓ fitted to the raw ensemble margin is somewhat more involved. One possibility might be to let S_ℓ be the CDF of a Bernoulli-Gamma mixture with point mass at zero equal to the fraction of ensemble members equal to zero and mean and variance of the Gamma part equal to the empirical mean and variance, respectively, of the remaining members. However, the variance of the Gamma part may become zero, for instance if exactly one ensemble member has a strictly positive value or if all ensemble members apart from those predicting zero precipitation have exactly the same strictly positive value. In this case, one may then take a very small variance near, but not equal to, zero, for example. Alternatively, S_ℓ could be taken to be the CDF of a GEV distribution left-censored at zero with location parameter or mean equal to the ensemble mean and scale parameter equal to the ensemble mean difference (Scheuerer, 2014), with adaptations in the case of an ensemble mean difference of value zero as described above.

(c) The sampling method (Q)

Arguably, the most intuitive and natural way of sampling from a predictive CDF F_ℓ is to compute equidistant quantiles from F_ℓ and take these as a sample $\tilde{x}^\ell_1, \ldots, \tilde{x}^\ell_M$ (Hagedorn,

2010; Hagedorn et al., 2012). Here, we consider two options to do so. An obvious procedure is to take equally spaced quantiles of the form (Hamill, 2001; Bremnes, 2007)

$$\tilde{x}_1^\ell := F_\ell^{-1}\left(\frac{1}{M+1}\right), \ldots, \tilde{x}_M^\ell := F_\ell^{-1}\left(\frac{M}{M+1}\right). \tag{Q}$$

Alternatively, equidistant quantiles of the form (Graf and Luschgy, 2000; Bröcker, 2012)

$$\tilde{x}_1^\ell := F_\ell^{-1}\left(\frac{1/2}{M}\right), \tilde{x}_2^\ell := F_\ell^{-1}\left(\frac{3/2}{M}\right), \ldots, \tilde{x}_M^\ell := F_\ell^{-1}\left(\frac{M-1/2}{M}\right) \tag{Q*}$$

can be chosen.

Both equidistant sampling variants reveal advantages and disadvantages. As shown by Bröcker (2012), the equally spaced quantiles according to sampling method (Q*) are optimal in expectation if the predictive performance is assessed by the common and popular continuous ranked probability score (CRPS). This result by Bröcker (2012) to a certain extent also underlines the findings of Examples 4.17 and 5.5, respectively, in Graf and Luschgy (2000). Therein, it is stated that for a standard uniform distribution $\mathcal{U}([0,1])$, the uniquely determined so-called M-optimal set of centers is given by

$$\mathcal{S} := \left\{ \frac{m-1/2}{M} \,\middle|\, m = 1, \ldots, M \right\}.$$

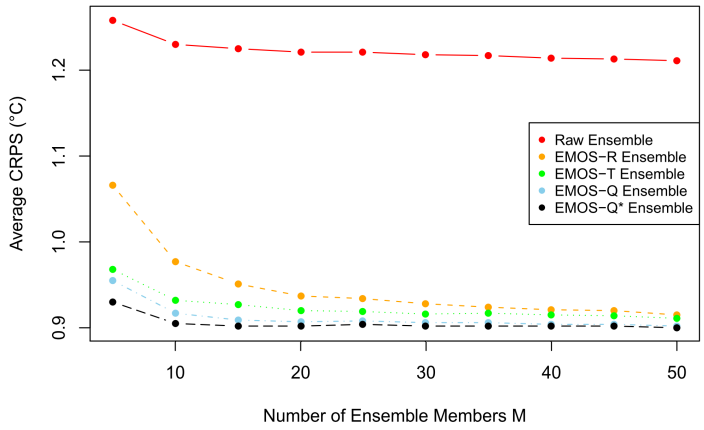
According to Graf and Luschgy (2000), finding $\mathcal{S}$ is equivalent to determining the M-optimal quantizers, which is the alternative optimality notion used by them, for $\mathcal{U}([0,1])$. Hence, the choice of the equally spaced $\frac{1/2}{M}$-, …, $\frac{(M-1/2)}{M}$-quantiles as samples from the predictive CDF F_ℓ can be regarded as optimal in the sense of Graf and Luschgy (2000) as discussed before.

While sampling method (Q*) is optimal in terms of the CRPS, it fails to maintain calibration to some extent. Specifically, Bröcker (2012) argues that the ensemble obtained by sampling method (Q*) cannot be expected to produce a flat rank histogram, which is however an intuitive criterion that is widely used as a diagnostic check for calibration. In contrast, the equidistant quantiles according to sampling method (Q) are not affected by this problem and maintain the calibration of univariate ensemble forecasts, although not being optimal with respect to a widely used score.

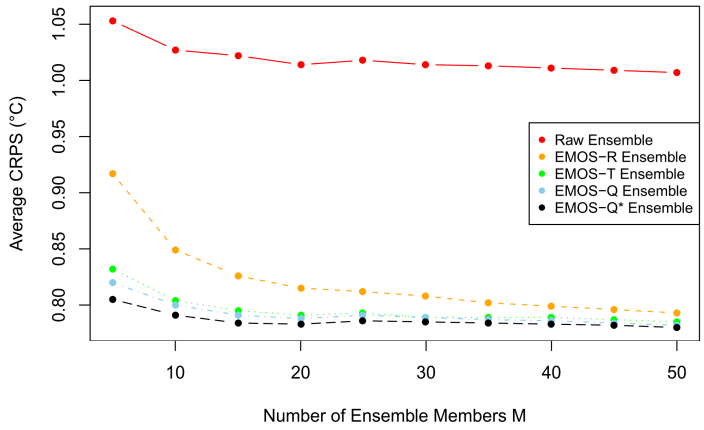
We have to make a trade-off here and follow Gneiting et al. (2007), arguing that the goal of probabilistic forecasting is maximizing the sharpness of a predictive distribution subject to calibration. Thus, we generally favor the equidistant sampling variant (Q) during the further course of this thesis. However, sampling method (Q*) will also be employed in the following case study, aiming at a comparison of the different sampling methods and a confirmation of the theoretical aspects discussed before.

4.2.2 Case study

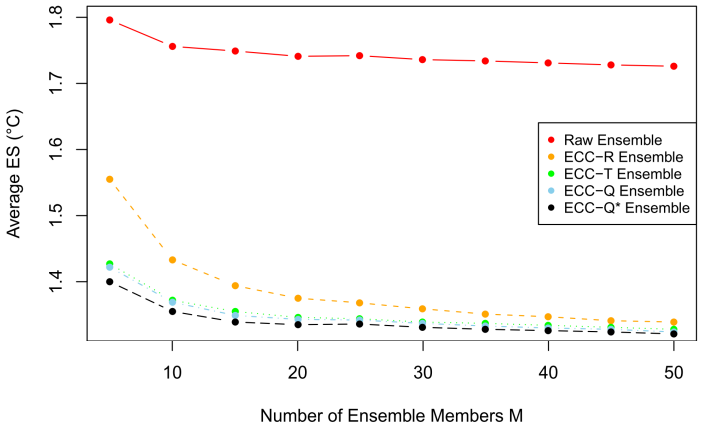
We now investigate the effect of the different sampling types on the predictive performance in a case study, dealing with 24 hour ahead forecasts of temperature, pressure, u- and v-wind at Berlin and Hamburg. Univariate postprocessing of the 50-member ECMWF raw ensemble forecasts is performed via EMOS, employing a sliding training period of 30 days. When



(a) CRPS Berlin



(b) CRPS Hamburg



(c) ES Berlin and Hamburg jointly

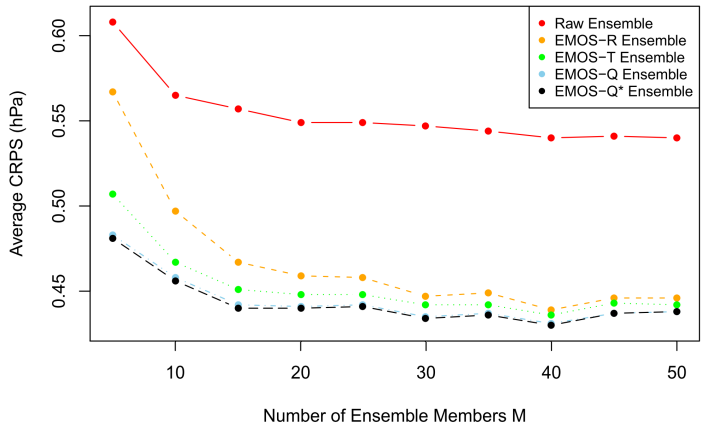
Figure 4.7: CRPS and ES, respectively, for 24 hour ahead temperature forecasts according to different quantization schemes at (a) Berlin, (b) Hamburg and (c) Berlin and Hamburg jointly, as a function of the ensemble size. The results are averaged over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS, and the results for sampling scheme (R) are averaged over 100 runs.

focussing on the univariate case, we sample 50 times from each marginal predictive EMOS CDF according to the schemes (R), (T), (Q) and (Q*), respectively, and call the resulting ensembles EMOS-R, EMOS-T, EMOS-Q and EMOS-Q*. As we only deal with weather quantities, for which the Gaussian EMOS model (3.4) is employed, we take the CDF S_ℓ in the sampling scheme (T) to be normal with mean equal to the empirical ensemble mean and variance equal to the empirical ensemble variance, as described before. We will also consider forecasts jointly at Berlin and Hamburg in our case study, where the weather quantity is fixed. In this bivariate setting, each marginal sample is additionally reordered with respect to the rank dependence structure of the raw ensemble according to the ECC notion, and the corresponding ensembles are referred to as ECC-R, ECC-T, ECC-Q and ECC-Q*, respectively. The reordering is obsolete for ECC-T, as the rank dependence pattern from the raw ensemble is retained by construction in this case. Intuitively, the predictive performance of the different EMOS ensembles might depend on the size M of the underlying raw ensemble. Thus, we assess our postprocessed ensembles for different ensemble sizes ranging from $M = 5$ to $M = 50$ in steps of 5 members. For the purposes of our case study, the corresponding M-member raw ensemble is constructed by just taking the first M members out of the 50-member ECMWF raw ensemble, where $M \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$ here. For evaluation, we consider the one-year test period from 1 May 2010 to 30 April 2011.

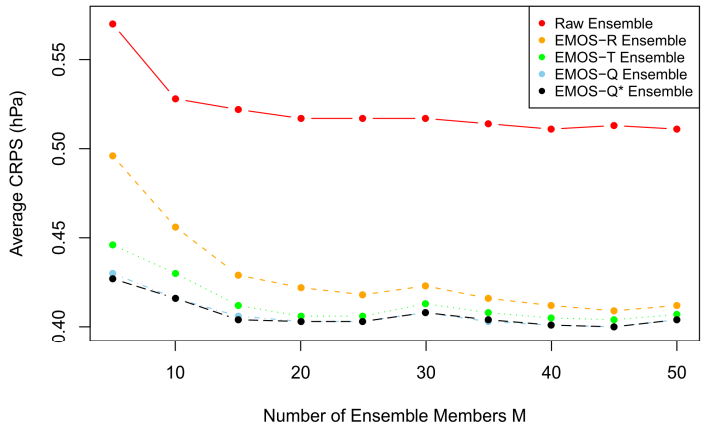
As overall performance measures, we employ the CRPS in the univariate case and the energy score (ES) in the multivariate case, respectively. The average results over the test period for the four weather quantities are shown in Figures 4.7 to 4.10, where panel (a) refers to the CRPS for Berlin only, panel (b) to the CRPS for Hamburg only, and panel (c) to the ES for Berlin and Hamburg jointly in each case. In the case of sampling scheme (R), the results are averaged over 100 runs.

Concerning the CRPS for the different weather quantities at Berlin and Hamburg individually, the EMOS-based ensembles clearly outperform the unprocessed raw ensemble, regardless of the sampling type. The EMOS-Q* and EMOS-Q ensembles based on equally spaced quantiles outperform the EMOS-T ensemble, which on the other hand outperforms the EMOS-R ensemble. Well in accordance with the optimality results of Bröcker (2012) discussed before, EMOS-Q* performs best throughout, although the differences to the performance of the EMOS-Q ensemble are partly very minor, typically in the case of large ensembles. Generally, the larger the ensemble size M is, the smaller the differences of the scores between the individual sampling types, with those based on equidistant quantiles nevertheless still performing best. Thus, the quantization type gets to some extent less important for rather large ensemble sizes, but not completely unimportant. In contrast, the sampling scheme appears to be much more important when dealing with small ensemble sizes, for which the differences in the predictive performance of the individual quantization schemes are more pronounced. In the context of the evaluation of ensembles, Fricker et al. (2013) discuss fairness aspects and argue that assessing random samples as in EMOS-R via the CRPS might be unfair and inappropriate to some extent. However, for consistency reasons and since sampling type (R) will not play a major role in the further course of the thesis, the CRPS is used as a scoring rule for all types of ensembles in this case study.

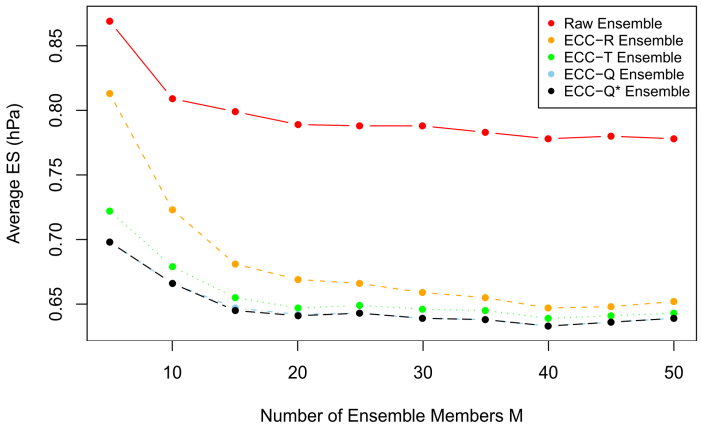
The univariate results generally extend to the multivariate case, in which we consider the predictive performances of the ensembles at Berlin and Hamburg jointly. In particular, ECC-Q* also performs best with respect to the ES, although no optimality result as in the



(a) CRPS Berlin



(b) CRPS Hamburg



(c) ES Berlin and Hamburg jointly

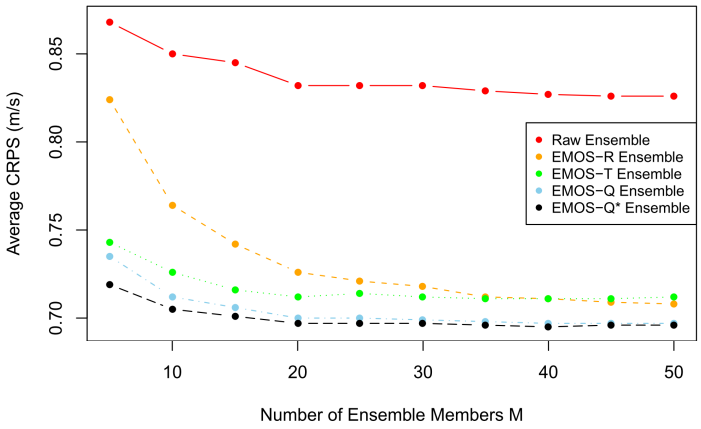
Figure 4.8: CRPS and ES, respectively, for 24 hour ahead pressure forecasts according to different quantization schemes at (a) Berlin, (b) Hamburg and (c) Berlin and Hamburg jointly, as a function of the ensemble size. The results are averaged over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS, and the results for sampling scheme (R) are averaged over 100 runs.

univariate case of the CRPS is known for the multivariate setting. The derivation of an optimal sampling scheme with respect to the ES for multivariate settings is an issue for future research.

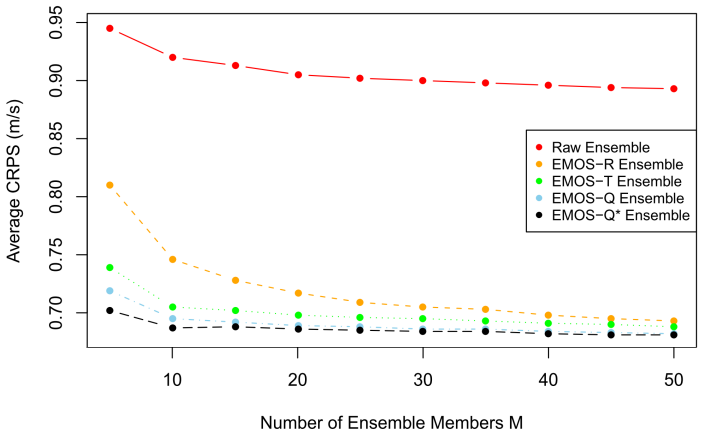
For the assessment of calibration only, we use the verification rank histogram in the univariate case and the multivariate rank histogram in the multivariate case. We exemplarily focus on pressure and u-wind here, where the ensemble sizes are restricted to $M \in \{5, 10, 15, 35, 50\}$ here. For pressure, the corresponding histograms for Berlin and Hamburg individually and for the two locations jointly are shown in Figures 4.11 to 4.13, respectively. Analogously, the histograms for u-wind can be found in Figures 4.14 to 4.16. In the univariate case of Berlin and Hamburg separately, all EMOS-based ensembles are, regardless of the employed sampling scheme, better calibrated than the raw ensemble, which reveals skewed (pressure) and U-shaped (u-wind) verification rank histograms, indicating bias and underdispersion, respectively. For small ensemble sizes, EMOS-R shows very slightly U-shaped verification rank histograms in the case of pressure only, while the histograms are partly inverse U-shaped for EMOS-T, indicating overdispersion. There appears to be too little data to fit S_ℓ appropriately in the case of small M. The EMOS-Q and EMOS-Q* ensembles reveal good calibration throughout. In this context, recall that according to Bröcker (2012), the EMOS-Q* ensemble cannot be expected to produce flat, uniform rank histograms. Specifically, Bröcker (2012) argues that in case of EMOS-Q*, the probability to attain the extreme ranks 1 and $M + 1$, respectively, is $1/(2M)$ and thus lower than the probability $1/M$ to attain one of the ranks $2, \ldots, M$. In our examples, the underpopulation of the ranks 1 and 6 in the case of $M = 5$ for the EMOS-Q* ensemble is in accordance with the theory by Bröcker (2012). For large ensemble sizes M, all sampling schemes employed to form the EMOS ensembles effectively perform equally well, and differences between the quantization methods vanish, which is well in line with the findings for the CRPS discussed before.

We stress again that our results should be rather viewed as a proof-of-concept than a comprehensive case study. The test period comprising 365 forecast cases might be rather short to check for uniformity of rank histograms, especially for large ensembles with $M = 35$ or $M = 50$, for instance, yielding 36 and 51 possible ranks, respectively, that the verifying observation can attain.

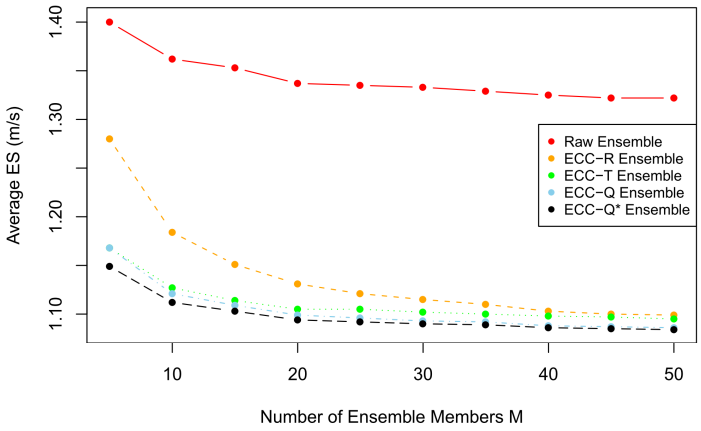
In summary, sampling scheme (Q) appears to be the most appropriate, and it is already widely used in meteorological and hydrological applications (Hamill, 2001; Bremnes, 2007). Even though optimal sampling tables for some special fixed distributions based on the alternative notion of M-optimal centers (Graf and Luschgy, 2000) are available (Max, 1960; Paez and Glisson, 1972; Lloyd, 1982), sampling scheme (Q) can be readily and conveniently applied to any type of distribution, hence allowing for a general implementation. In our case study, it performed better than the sampling schemes (R) and (T), respectively, and in contrast to method (T), there is no need to fit a CDF S_ℓ to the raw ensemble values, which can be demanding in non-Gaussian cases. Although the choice of quantization type (Q*) would be optimal with respect to the widely used CRPS, sampling scheme (Q) is only very slightly outperformed by (Q*) in terms of the CRPS, particularly if the ensemble size is rather large, as in the case of the 50-member ECMWF ensemble used in our case studies. We therefore set (Q) as the natural standard sampling method in the case studies to follow in this thesis, even though we also get back to quantization type (T) in a comparative case study in Chapter 5.



(a) CRPS Berlin

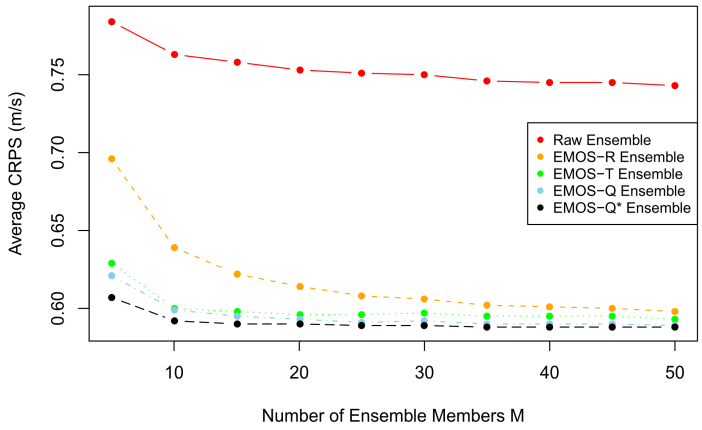


(b) CRPS Hamburg

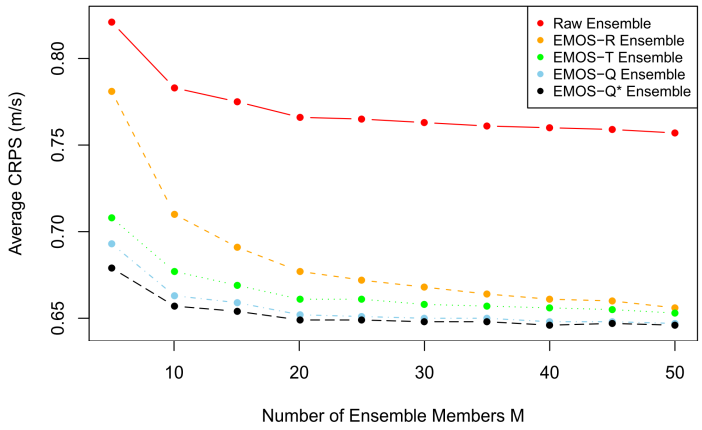


(c) ES Berlin and Hamburg jointly

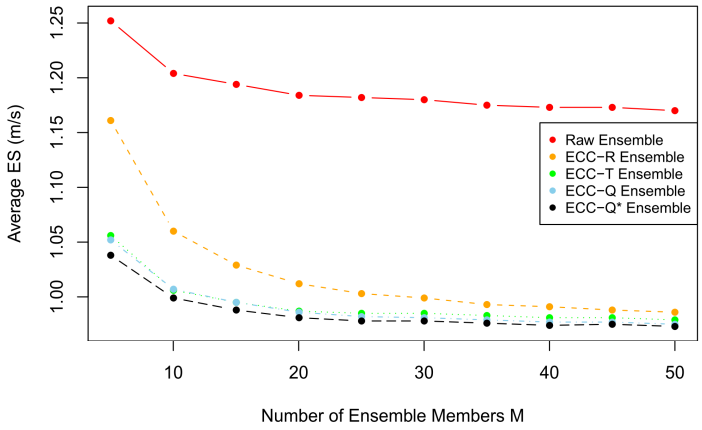
Figure 4.9: CRPS and ES, respectively, for 24 hour ahead u-wind forecasts according to different quantization schemes at (a) Berlin, (b) Hamburg and (c) Berlin and Hamburg jointly, as a function of the ensemble size. The results are averaged over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS, and the results for sampling scheme (R) are averaged over 100 runs.



(a) CRPS Berlin



(b) CRPS Hamburg



(c) ES Berlin and Hamburg jointly

Figure 4.10: CRPS and ES, respectively, for 24 hour ahead v-wind forecasts according to different quantization schemes at (a) Berlin, (b) Hamburg and (c) Berlin and Hamburg jointly, as a function of the ensemble size. The results are averaged over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS, and the results for sampling scheme (R) are averaged over 100 runs.

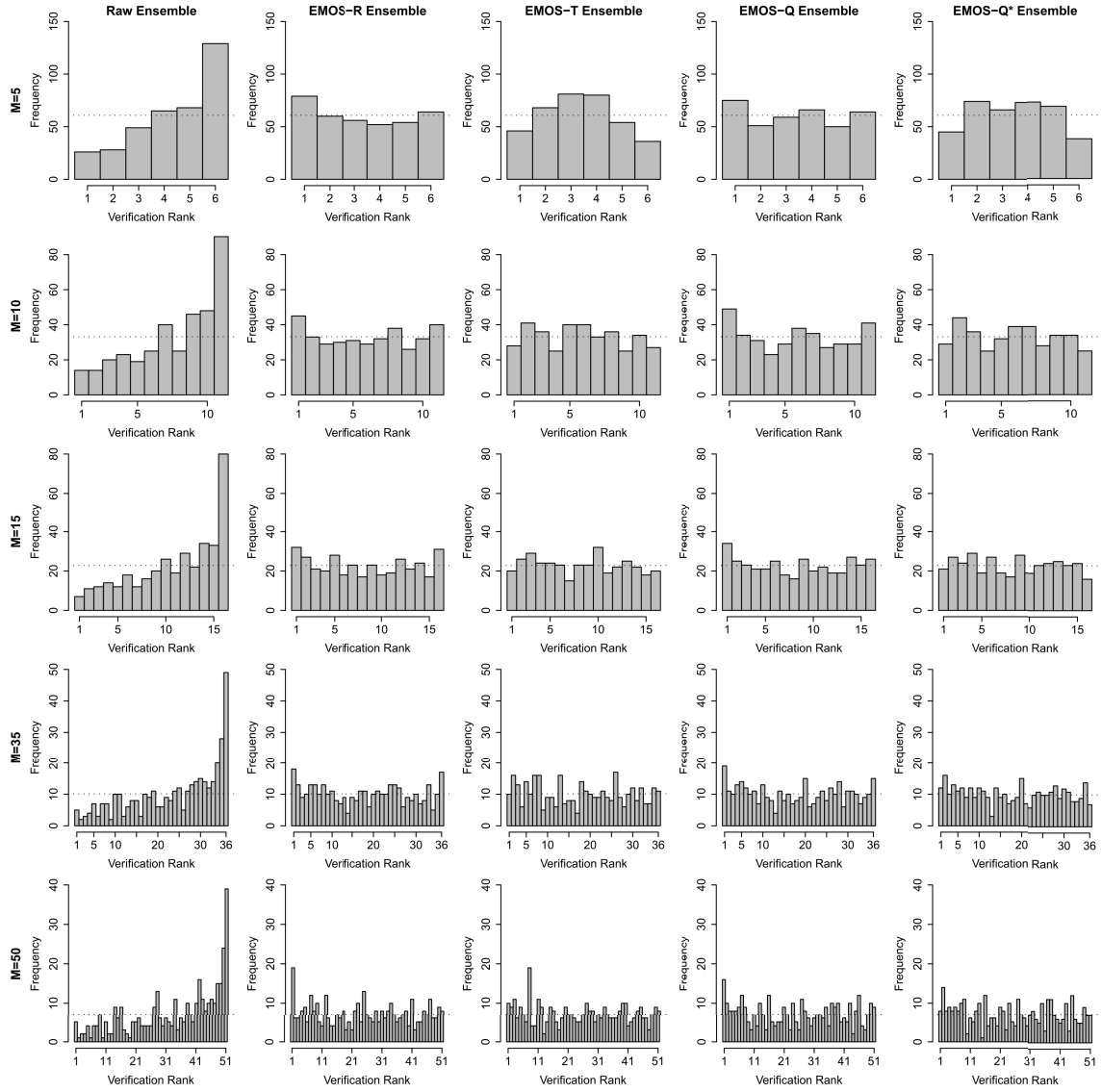


Figure 4.11: Verification rank histograms for 24 hour ahead pressure forecasts at Berlin according to different quantization schemes and for different ensemble sizes over the test period from 1 May 2010 to 30 April 2011.

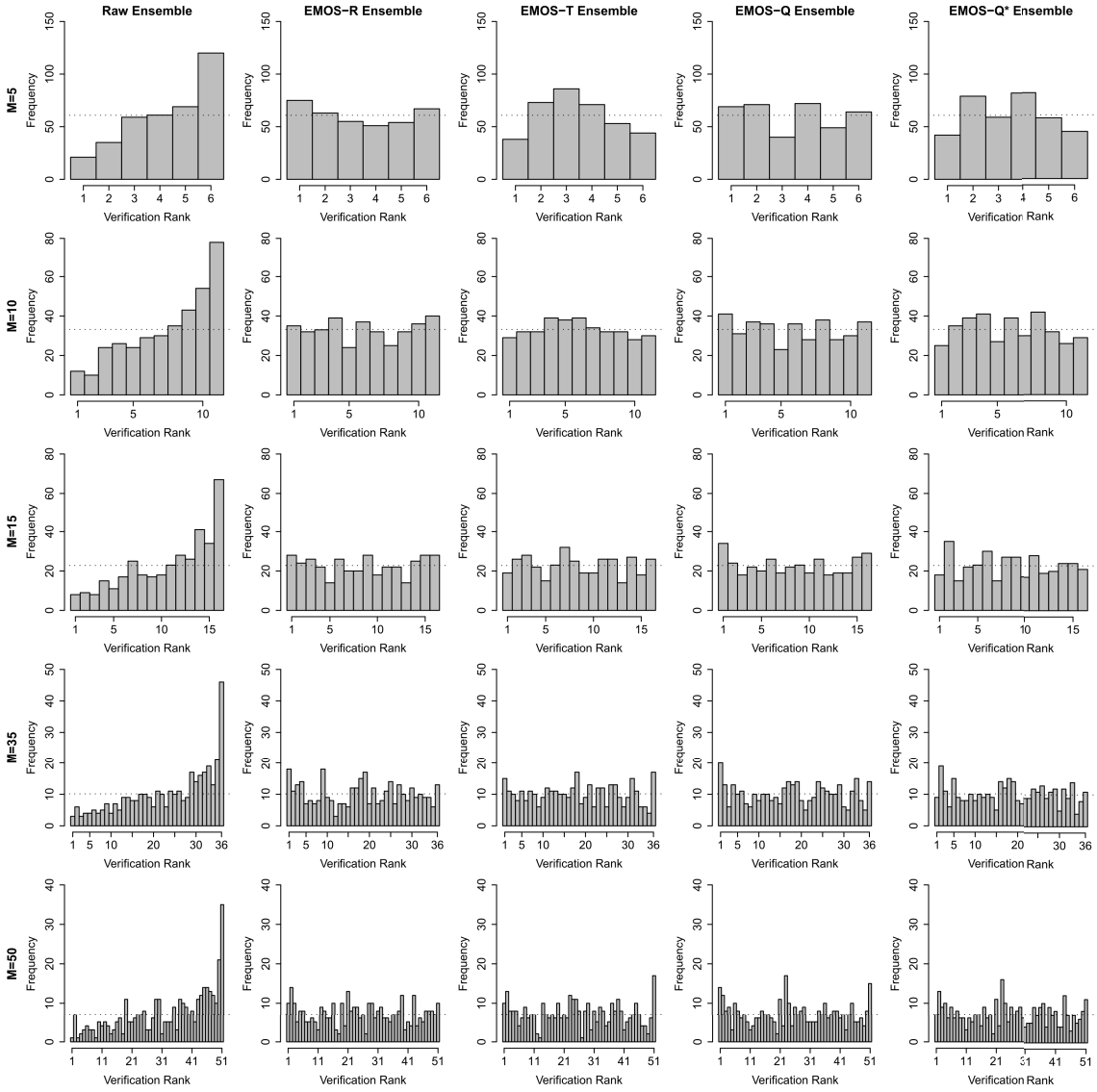


Figure 4.12: Verification rank histograms for 24 hour ahead pressure forecasts at Hamburg according to different quantization schemes and for different ensemble sizes over the test period from 1 May 2010 to 30 April 2011.

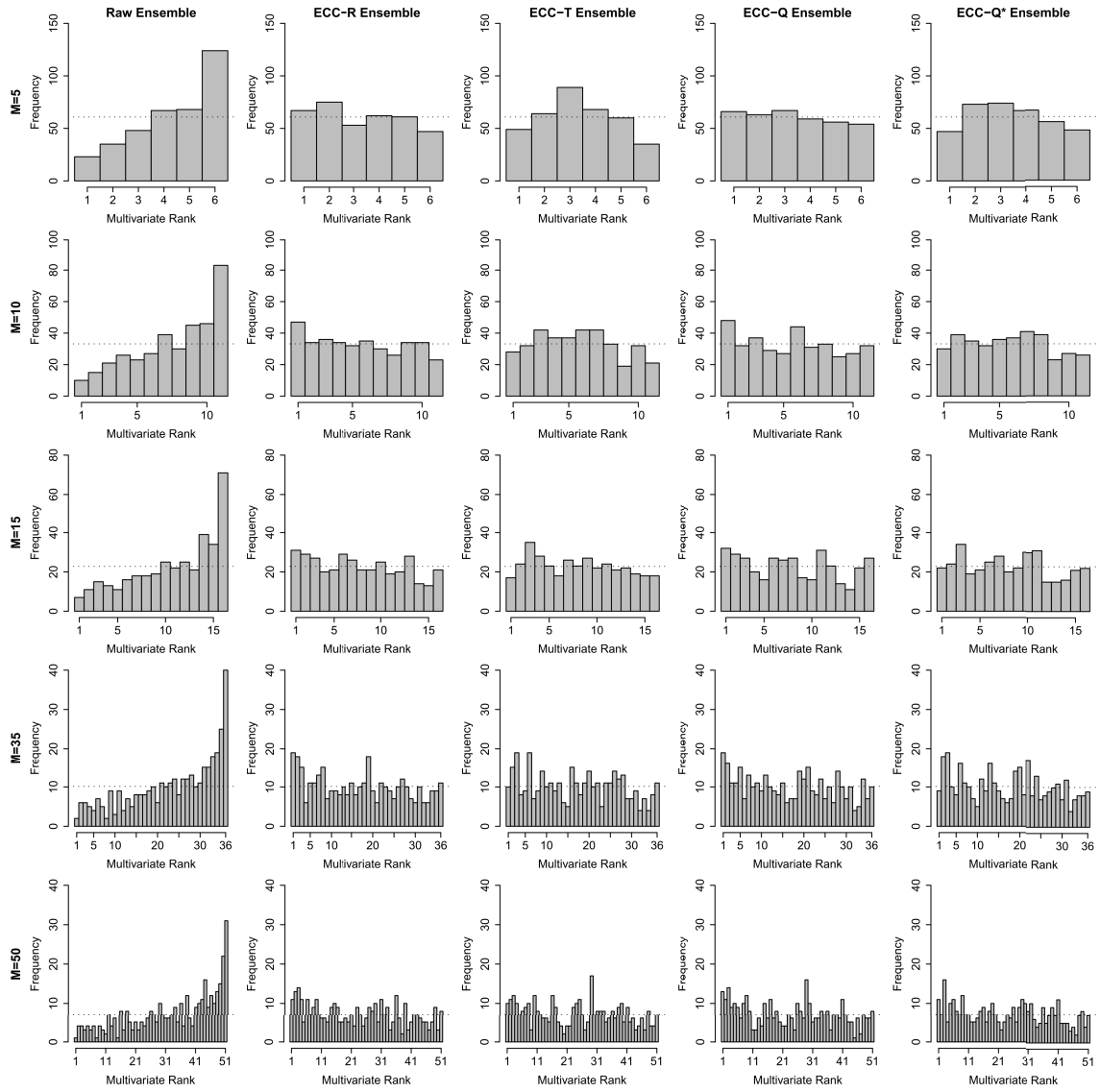


Figure 4.13: Multivariate rank histograms for 24 hour ahead pressure forecasts at Berlin and Hamburg jointly according to different quantization schemes and for different ensemble sizes over the test period from 1 May 2010 to 30 April 2011.

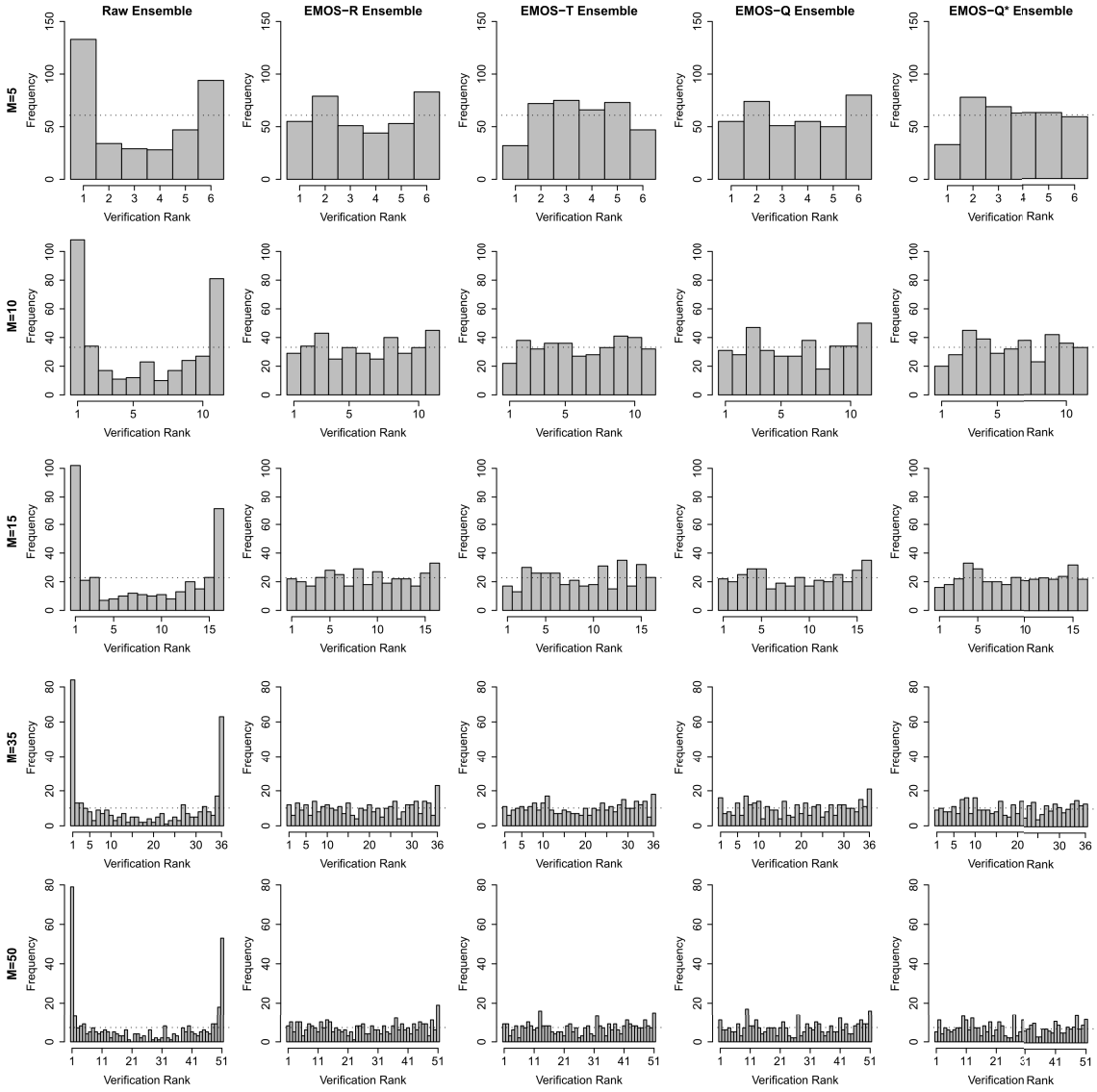


Figure 4.14: Verification rank histograms for 24 hour ahead u-wind forecasts at Berlin according to different quantization schemes and for different ensemble sizes over the test period from 1 May 2010 to 30 April 2011.

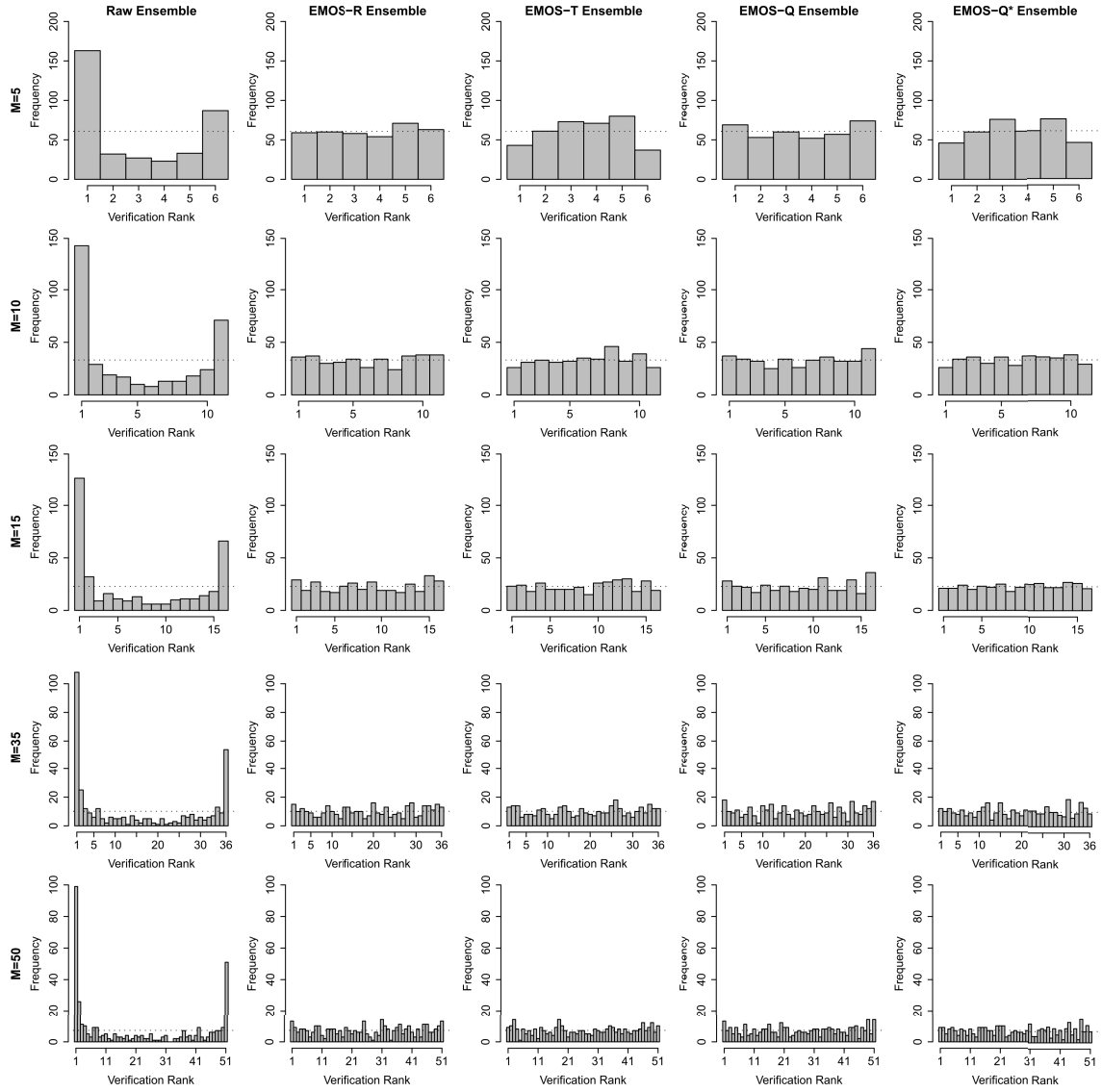


Figure 4.15: Verification rank histograms for 24 hour ahead u-wind forecasts at Hamburg according to different quantization schemes and for different ensemble sizes over the test period from 1 May 2010 to 30 April 2011.

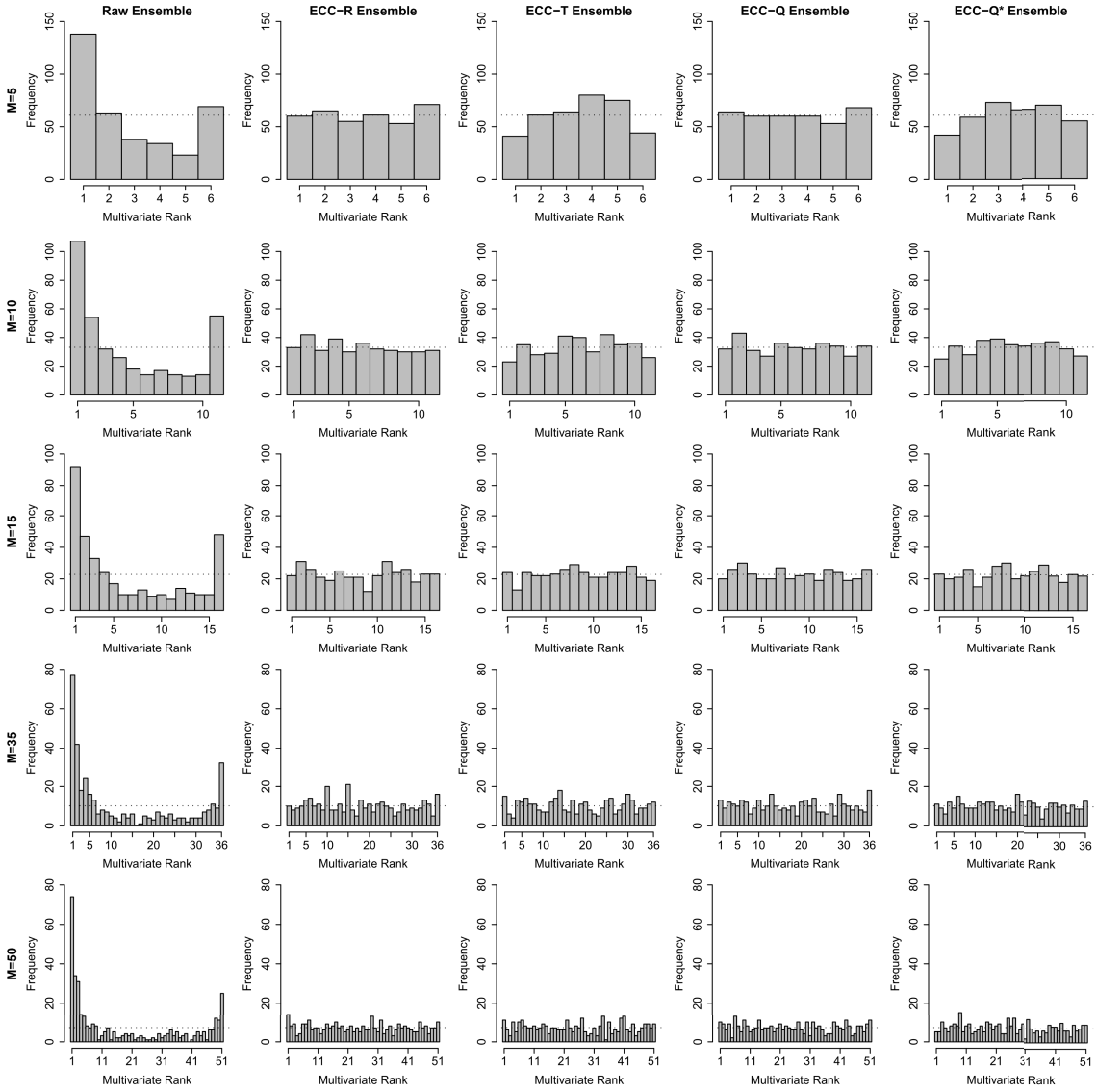


Figure 4.16: Multivariate rank histograms for 24 hour ahead u-wind forecasts at Berlin and Hamburg jointly according to different quantization schemes and for different ensemble sizes over the test period from 1 May 2010 to 30 April 2011.

4.3 The predictive performance of ECC: Case studies

In this section, we evaluate the predictive performance of the ECC ensemble in a comprehensive real-data case study, which is meant to be a proof-of-concept here, and compare it to that of both the raw and related postprocessed ensembles. To this end, we consider several multivariate settings dealing with spatial, inter-variable and temporal aspects, or combinations thereof.

4.3.1 Implementation and reference ensembles

We first describe the implementation of ECC in our case study, the data we employ and the reference ensembles ECC is compared to.

The raw data are provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble consisting of $M = 50$ members, as introduced in Section 1.3. For each margin $\ell := (i, j, k)$ corresponding to weather variable $i \in \{1, \dots, I\}$, location $j \in \{1, \dots, J\}$ and look-ahead time $k \in \{1, \dots, K\}$, the raw ensemble forecast $x_1^\ell, \dots, x_{50}^\ell$ is univariately postprocessed using either BMA or EMOS with a sliding training window of 30 days, as presented in Sections 3.1.1 and 3.1.2 and implemented in the R packages `ensembleBMA` and `ensembleMOS`, respectively (R Core Team, 2013), leading to a predictive CDF, F_ℓ.

From a spatial point of view, we concentrate on the observation sites of Berlin, Hamburg and Frankfurt and on grid-based test regions over Germany and adjacent areas. For the stations, we have real verifying observations, and the corresponding raw ensemble forecasts for each location are obtained by bilinear interpolation of the grid-based raw ensemble forecasts. In contrast, we also consider grid-based test areas consisting of several points of the grid, on which the ECMWF raw ensemble forecasts are issued. As real observations are not available at all grid points, we employ the corresponding 0 hour ahead nowcasts of the ECMWF control run as the ground truth in this case, as discussed in Section 1.3. Moreover, we confine ourselves to 24, 48, 72 and 96 hour ahead forecasts, and concerning the weather quantities, we focus on temperature, pressure, u- and v-wind, while precipitation is left out here. As already hinted at in Section 3.1.3, univariate BMA and EMOS in some situations fail to outperform the unprocessed raw ensemble for precipitation. This deficiency is spread to the corresponding multivariate techniques relying on such poor univariate forecasts. Hence, failures in the multivariate case are very likely grounded on shortcomings concerning the univariate case, which is why we skip results for precipitation in the following case study, which focuses on multivariate predictive performance.

After the univariate postprocessing, we then apply sampling scheme (Q) and generate the $M = 50$ equally spaced quantiles $\tilde{x}_1^\ell := F_\ell^{-1}\left(\frac{1}{51}\right), \dots, \tilde{x}_{50}^\ell := F_\ell^{-1}\left(\frac{50}{51}\right)$ from each marginal predictive CDF F_ℓ. For each fixed ℓ, these quantiles can be aggregated in different ways to obtain a final postprocessed ensemble $\hat{x}_1^\ell, \dots, \hat{x}_{50}^\ell$. In our case study, we examine the following ensembles.

- The increasingly ordered quantiles (IOQ) ensemble is given by

$$\hat{x}_1^\ell := \tilde{x}_1^\ell, \dots, \hat{x}_{50}^\ell := \tilde{x}_{50}^\ell.$$

  That is, the equidistant quantiles stay increasingly ordered as they are.

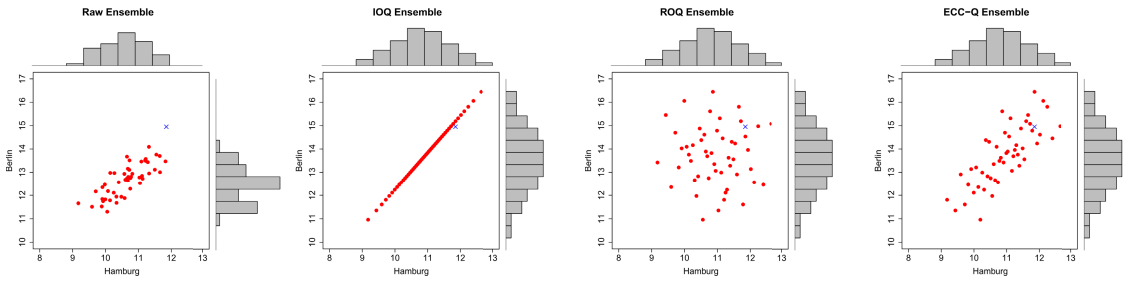


Figure 4.17: 24 hour ahead ensemble forecasts of temperature (in °C) at Berlin and Hamburg, valid 2:00 am on 27 June 2010. The ensemble forecasts are indicated by the red dots, while the blue cross represents the verifying observation. The respective margins are illustrated by the histograms at the top and to the right of each scatterplot. Univariate postprocessing is performed via BMA.

- The randomly ordered quantiles (ROQ) ensemble is given by

$$\hat{x}^\ell_1 := \tilde{x}^\ell_{\lambda_\ell(1)}, \ldots, \hat{x}^\ell_{50} := \tilde{x}^\ell_{\lambda_\ell(50)},$$

  with a random permutation λ_ℓ.
  That is, the equidistant quantiles are just randomly reordered.

- The ECC-Q ensemble as in Section 4.1 is given by

$$\hat{x}^\ell_1 := \tilde{x}^\ell_{(\sigma_\ell(1))}, \ldots, \hat{x}^\ell_{50} := \tilde{x}^\ell_{(\sigma_\ell(50))},$$

  where the permutation σ_ℓ is defined by the raw ensemble forecast $x^\ell_1, \ldots, x^\ell_{50}$ via $\sigma_\ell(m) := \text{rank}(x^\ell_m)$ for $m \in \{1, \ldots, 50\}$, with ties resolved at random.
  That is, the equidistant quantiles are reordered with respect to the rank dependence structure of the raw ensemble.

Along with the raw ensemble, the three postprocessed ensembles are illustrated in Figure 4.17, showing 24 hour ahead temperature forecasts at Berlin and Hamburg, valid at 2:00 am on 27 June 2010, where univariate postprocessing is performed via BMA. The ensemble forecasts are indicated by the red dots, while the blue cross represents the verifying observation.

In contrast to the raw ensemble, all three postprocessed ensembles, for which biases and dispersion errors have been corrected, exhibit the same margins, as evidenced by the histograms alongside of the scatterplots. However, they differ drastically in their multivariate dependence structures. The IOQ ensemble provides a maximal possible positive correlation, which is (often close to) linear when the sample stems from (mixtures of) Gaussian distributions, as is the case for the weather variables considered here. Although the IOQ ensemble appears to be rather unrealistic in most situations, we nevertheless examine this scheme in what follows, as it contributes to a better interpretation and comprehension of the results and illustrates properties of the multivariate calibration verification methods of Thorarinsdottir et al. (2014). In contrast, the ROQ ensemble assumes that no correlation structure is evident. It essentially corresponds to the individually postprocessed ensembles illustrated in Figures 4.1, 4.2, 4.3 and 4.5, respectively, in which any spatial dependencies get lost. Finally, the ECC-Q ensemble inherits its multivariate rank dependence structure from the raw ensemble, such that the scatterplots of the raw and the ECC-Q ensemble show the same rank correlation pattern.

Table 4.1: Average ES for 24 hour ahead ensemble forecasts of temperature (temp), pressure (press), u- and v-wind, respectively, at Berlin (Ber), Hamburg (Ham) and Frankfurt (Fra) jointly ($L = 3$) and at Berlin (Ber) and Hamburg (Ham) jointly ($L = 2$), respectively. Univariate postprocessing is performed via BMA on the one hand and EMOS on the other hand. The results are averaged over the test period from 1 May 2010 to 30 April 2011. For the ROQ ensemble, the scores are averaged over 100 runs.

		Ber/Ham/Fra jointly				Ber/Ham jointly			
		Temp (°C)	Press (hPa)	u-Wind (m/s)	v-Wind (m/s)	Temp (°C)	Press (hPa)	u-Wind (m/s)	v-Wind (m/s)
	Raw Ensemble	2.278	1.005	1.740	1.839	1.726	0.778	1.322	1.170
BMA	IOQ Ensemble	1.817	0.833	1.405	1.644	1.375	0.634	1.126	1.012
	ROQ Ensemble	1.726	0.827	1.328	1.569	1.335	0.640	1.087	0.969
	ECC-Q Ensemble	1.724	0.811	1.331	1.568	1.329	0.625	1.088	0.969
EMOS	IOQ Ensemble	1.811	0.850	1.407	1.650	1.368	0.648	1.125	1.017
	ROQ Ensemble	1.721	0.841	1.330	1.577	1.329	0.650	1.084	0.974
	ECC-Q Ensemble	1.720	0.829	1.334	1.578	1.324	0.639	1.086	0.975

4.3.2 Spatial aspects

We start our case study with the assessment of the predictive performance of our reference ensembles in modeling spatial dependence structures only. That is, we work with a single, fixed prediction horizon, which is 24 hours here, and look at the weather variables temperature, pressure, u- and v-wind separately, while locations are considered jointly. The locations themselves are of a twofold shape: In the first part of our study, we focus on the sites of Berlin, Hamburg and Frankfurt, for which verifying observations are available. In the second part, we consider grid-based forecast areas over Germany and its surroundings, with the corresponding 0 hour ahead nowcasts of the ECMWF control run serving as the ground truth, as described before.

(a) Berlin, Hamburg and Frankfurt

The 24 hour ahead ensemble forecasts for the individual weather quantities are considered jointly at Berlin (13.58° longitude east, 52.52° latitude north), Hamburg (10.00° longitude east, 53.55° latitude north) and Frankfurt am Main (8.68° longitude east, 50.12° latitude north) in an $L = 3$-dimensional setting, and jointly at Berlin and Hamburg in an $L = 2$-dimensional setting. The three locations are marked on the map in Figure 4.22 (a), where the linear distance from Frankfurt to either Berlin or Hamburg is on the order of 400 kilometers, while the linear distance between Berlin and Hamburg is about 250 kilometers. Forecast errors for pressure reveal rather strong long range dependence patterns. This holds to some lesser extent for temperature, while the dependencies for both u-and v-wind are on a moderate level, as wind patterns typically vary at much smaller spatial scales.

At this stage, we consistently employ either BMA or EMOS for univariate postprocessing at the indivaidual observation sites for comparative reasons in these scenarios, using a rolling training period of 30 days. The one-year test period comprises the days from 1 May 2010 to 30 April 2011.

The overall predictive performance of the ensembles is evaluated via the energy score (ES)

Table 4.2: Average EE for 24 hour ahead ensemble forecasts of temperature (temp), pressure (press), u- and v-wind, respectively, at Berlin (Ber), Hamburg (Ham) and Frankfurt (Fra) jointly ($L = 3$) and at Berlin (Ber) and Hamburg (Ham) jointly ($L = 2$), respectively. Univariate postprocessing is performed via BMA on the one hand and EMOS on the other hand. The results are averaged over the test period from 1 May 2010 to 30 April 2011. For the ROQ ensemble, the scores are averaged over 100 runs.

		Ber/Ham/Fra jointly				Ber/Ham jointly			
		Temp (°C)	Press (hPa)	u-Wind (m/s)	v-Wind (m/s)	Temp (°C)	Press (hPa)	u-Wind (m/s)	v-Wind (m/s)
	Raw Ensemble	2.84	1.38	2.19	2.27	2.15	1.07	1.67	1.49
BMA	IOQ Ensemble	2.40	1.12	1.84	2.19	1.86	0.87	1.51	1.35
	ROQ Ensemble	2.40	1.13	1.84	2.19	1.86	0.87	1.52	1.35
	ECC-Q Ensemble	2.40	1.12	1.84	2.19	1.85	0.87	1.52	1.35
EMOS	IOQ Ensemble	2.38	1.11	1.83	2.19	1.83	0.86	1.51	1.35
	ROQ Ensemble	2.38	1.12	1.84	2.19	1.83	0.86	1.51	1.35
	ECC-Q Ensemble	2.38	1.11	1.84	2.19	1.83	0.86	1.51	1.35

in Table 4.1. Moreover, we also show results for the Euclidean error (EE) in Table 4.2, with the multivariate or spatial median serving as a point forecast based on the respective ensemble, as described in Section 2.3. Last but not least, the sharpness of the forecast ensembles is measured by the determinant sharpness (DS) in Table 4.3. For the ROQ ensemble, all numerical scores are averaged over 100 runs. Calibration is assessed via the multivariate, band depth and average rank histograms, respectively, where we exemplarily show the results for each weather variable for the $L = 3$-dimensional situation of Berlin, Hamburg and Frankfurt jointly in Figures 4.18 to 4.21, based on univariate postprocessing via EMOS. The corresponding histograms for the $L = 2$-dimensional setting of Berlin and Hamburg jointly are at large qualitatively very similar to those for $L = 3$ and are therefore not shown here.

In terms of the ES, all postprocessed ensembles outperform the unprocessed raw ensemble for both scenarios. In all cases, the IOQ ensemble performs worse than the ECC-Q ensemble, and except for pressure in the $L = 2$-dimensional setting, the IOQ ensemble also performs worse than the ROQ ensemble. For temperature and pressure, the ECC-Q ensemble outperforms the ROQ ensemble, while in the case of u- and v-wind, the ROQ ensemble mostly outperforms the ECC-Q ensemble. However, the differences between the ROQ and the ECC-Q ensemble with respect to the ES are generally rather minor, except for pressure. Meanwhile, it is known that the discrimination ability of the ES sometimes can be rather low (Pinson and Tastu, 2013), and finding remedies for this is a topic of current research (Scheuerer and Hamill, 2014). Our results in the above case may also be subject to this problem, which should also be kept in mind for the subsequent case studies.

With respect to the EE, all postprocessed ensembles also outperform the raw ensemble. However, all three postprocessed ensembles basically more or less have the same EE in each respective setting. This is likely due to the fact that they can largely be expected to return similar multivariate median values, as they have the same marginal distributions. Similar findings have also been obtained in the case studies by Möller et al. (2013).

Concerning the DS, both the ROQ and the ECC-Q ensemble typically reveal a higher DS

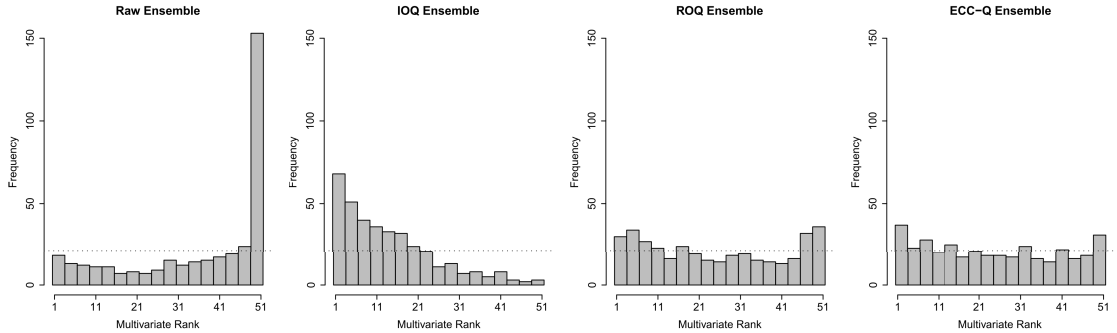
Table 4.3: Average DS for 24 hour ahead ensemble forecasts of temperature (temp), pressure (press), u- and v-wind, respectively, at Berlin (Ber), Hamburg (Ham) and Frankfurt (Fra) jointly ($L = 3$) and at Berlin (Ber) and Hamburg (Ham) jointly ($L = 2$), respectively. Univariate postprocessing is performed via BMA on the one hand and EMOS on the other hand. The results are averaged over the test period from 1 May 2010 to 30 April 2011. For the ROQ ensemble, the scores are averaged over 100 runs.

		Ber/Ham/Fra jointly				Ber/Ham jointly			
		Temp	Press	u-Wind	v-Wind	Temp.	Press	u-Wind	v-Wind
	Raw Ensemble	0.65	0.49	0.56	0.50	0.65	0.51	0.54	0.52
BMA	IOQ Ensemble	0.07	0.13	0.03	0.03	0.18	0.21	0.11	0.09
	ROQ Ensemble	1.22	0.73	0.99	1.10	1.20	0.74	1.06	0.92
	ECC-Q Ensemble	1.18	0.51	0.93	1.06	1.16	0.53	0.99	0.89
EMOS	IOQ Ensemble	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ROQ Ensemble	1.19	0.65	0.95	1.08	1.16	0.66	1.03	0.89
	ECC-Q Ensemble	1.15	0.46	0.89	1.04	1.13	0.47	0.96	0.87

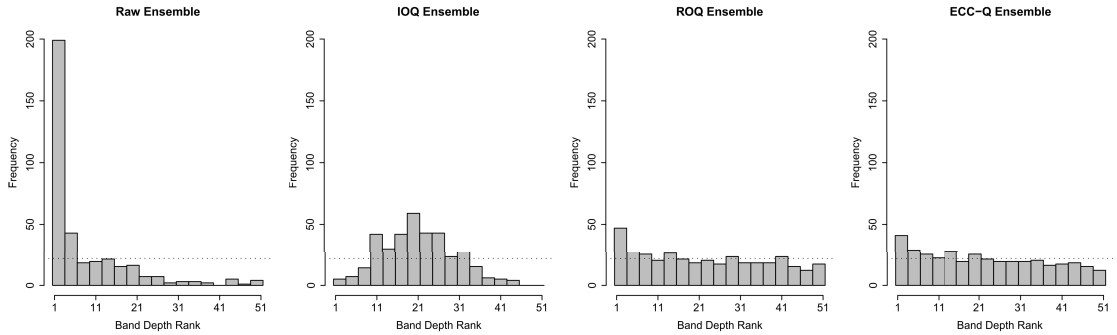
and therefore a lower sharpness than the raw ensemble. This is not that surprising, as raw ensembles typically are already sharp, however at the expense of calibration, in that they tend to be underdispersive with too little spread, and thus call for statistical postprocessing. The only exception can be observed for pressure, where the ECC-Q ensemble is sharper than the raw ensemble when univariate postprocessing is performed via BMA. However, the ECC-Q ensemble is in general sharper than the ROQ ensemble. For the IOQ ensemble, the situation is somewhat involved. This ensemble yields by far the lowest DS values and thus provides the sharpest forecasts when univariate postprocessing is done via BMA. However, the DS values somehow appear much too small to be realistic. This is due to the potentially inappropriately high correlation structure assumed by the IOQ ensemble, which is reflected in the covariance matrix Σ used for the calculation of the DS, having a small determinant. In the case of univariate postprocessing via EMOS, Σ is singular with determinant zero. This is because EMOS yields a single fitted normal distribution in our examples, and the IOQ ensemble then assumes an again higher, namely exactly linear, correlation than it does in case of working with a BMA postprocessed predictive distribution based on mixtures of normal distributions.

In terms of calibration only, the raw ensembles for each weather quantity exhibit either skewed or U-shaped rank histograms and are thus not calibrated. The IOQ ensembles show skewed multivariate rank histograms, indicating a bias, and inverse U-shaped band depth and average rank histograms, respectively, additionally pointing at an overestimation of the correlation structure. The ROQ ensemble is notably uncalibrated in the case of pressure. In terms of the other weather variables, its calibration appears to be better, but not totally satisfactory and basically inferior to that of the ECC-Q ensemble. Against the background that a "perfect" calibration is hard to achieve in our setting with a rather small test period, the ECC-Q ensembles generally perform fairly well and are all in all calibrated by far the best for each weather variable.

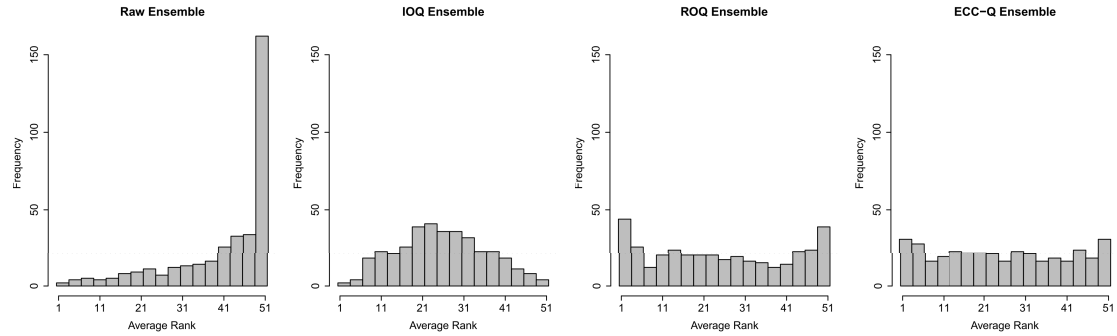
Putting all results together, the ECC-Q ensemble performs best. While all postprocessed ensembles basically perform equally well in terms of the EE, ECC-Q generally shows by far the best calibration and moreover the most reasonable DS, thus being in accordance with



(a) Multivariate rank histograms

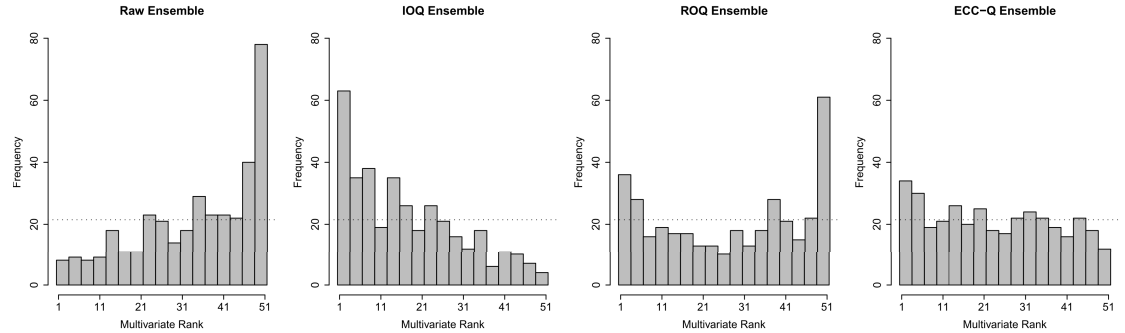


(b) Band depth rank histograms

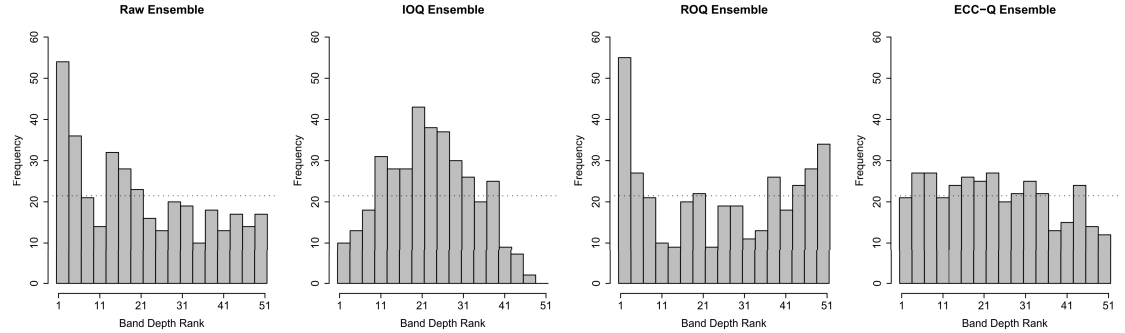


(c) Average rank histograms

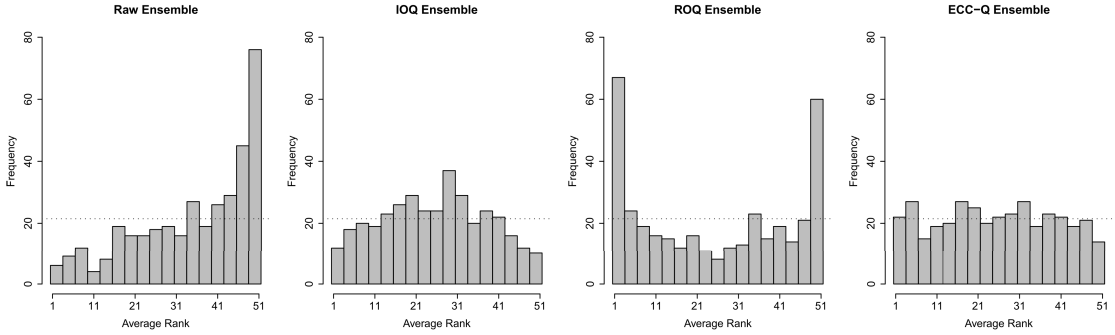
Figure 4.18: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.



(a) Multivariate rank histograms

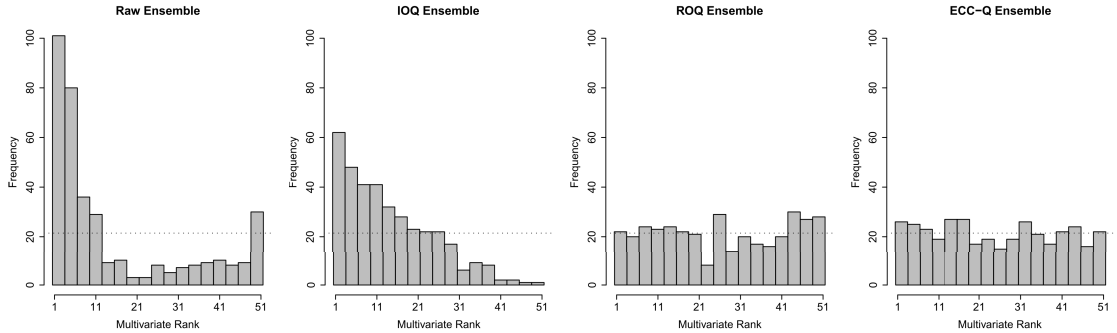


(b) Band depth rank histograms

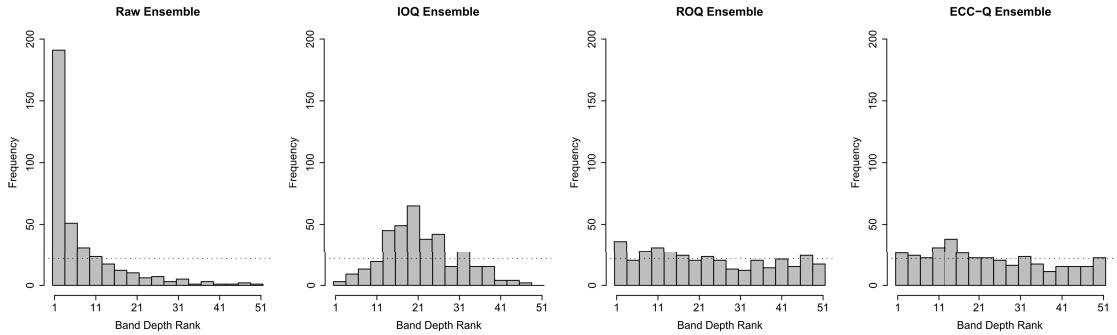


(c) Average rank histograms

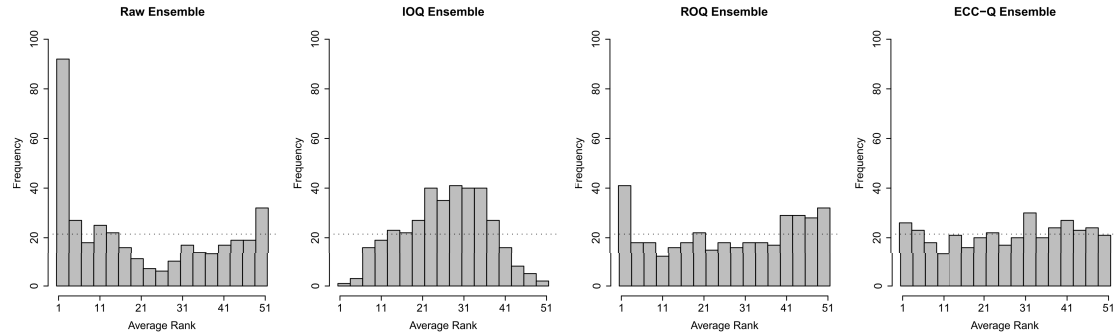
Figure 4.19: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead pressure forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.



(a) Multivariate rank histograms

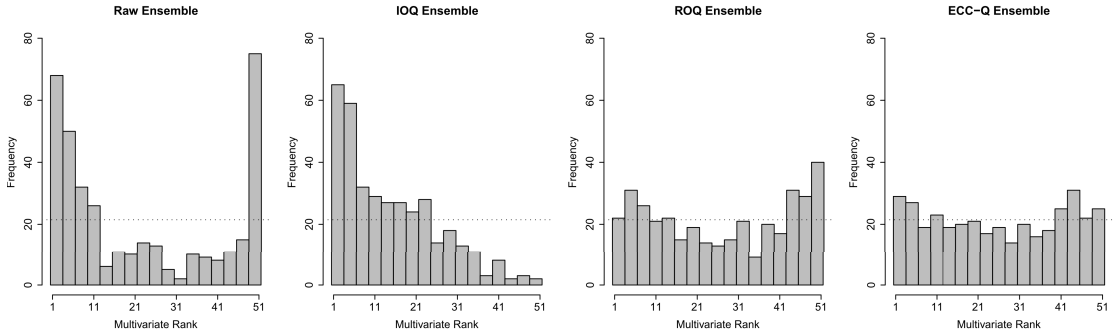


(b) Band depth rank histograms

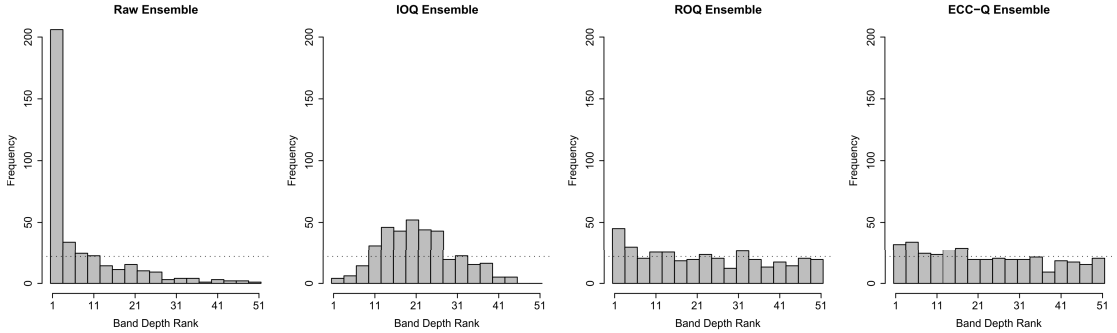


(c) Average rank histograms

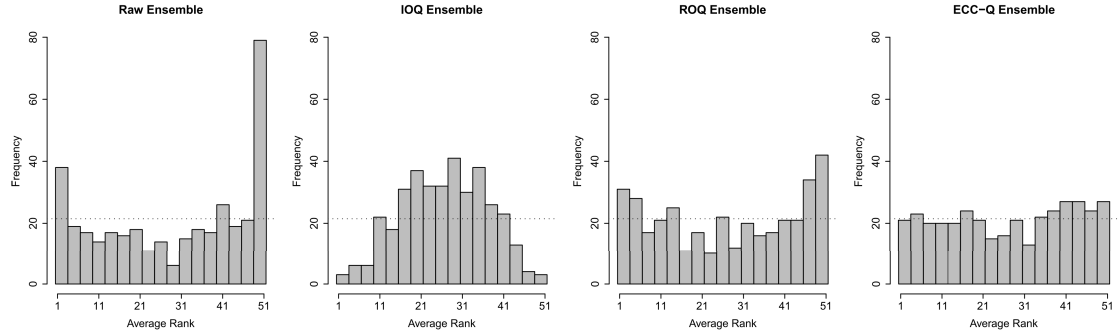
Figure 4.20: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead u-wind forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.



(a) Multivariate rank histograms



(b) Band depth rank histograms



(c) Average rank histograms

Figure 4.21: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead v-wind forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.

the goal of probabilistic forecasting, namely maximizing the sharpness of the prediction subject to calibration (Gneiting et al., 2007).

The performance of the postprocessed ensembles to some extent also depends on the underlying correlation structure in the forecast and observation data. The IOQ ensemble does an acceptable job and outperforms at least the ROQ ensemble if there are very strong dependencies, as is indeed the case for pressure at Berlin and Hamburg jointly. Generally, this cannot be expected, and we are typically confronted with less pronounced correlation structures. Thus, the ROQ ensemble mostly outperforms the IOQ ensemble, and the less pronounced the dependence structure the more the differences between ROQ and ECC-Q ensemble appear to vanish, aided by a potential shortcoming of the ES in discrimination ability. However, the ECC-Q ensemble appears to be appropriate for any type of underlying dependence structure, regardless of whether strong or weak, in particular if it is a priori unknown.

In our initial multivariate investigations, we obtain similar results for the corresponding reference ensembles, regardless of whether we employ BMA or EMOS in the univariate postprocessing step. This confirms the findings in Section 3.1.3 for purely univariate settings. As it is computationally faster, we confine ourselves to EMOS for univariate postprocessing in the rest of the case studies, although BMA might sometimes yield slightly better results than EMOS.

(b) Grid-based test regions over Germany and its surroundings

We now consider not only two or three locations simultaneously as before, but test areas over Germany and its surroundings. Each area comprises a different number L of points, corresponding to the dimension of the treated setting, on the ECMWF grid, which provides forecasts at a resolution of 0.25°, both for longitude and latitude. In total, we focus on four different contiguous regions I, II, II and IV, respectively, and one set V consisting of scattered grid points. The composition and characteristics of our test areas are summarized in Table 4.4, and the regions are marked on the maps in Figures 4.22 (b) and (c), respectively. Again, we deal with 24 hour ahead forecasts for pressure, temperature, u- and v-wind, respectively, for the one-year test period from 1 May 2010 to 30 April 2011. For the four contiguous test areas, the spatial correlations are stronger than in the case of the three individual stations considered before. Especially for pressure, they are very high, but also for the other weather quantities pronounced dependencies can be observed. Not unexpectedly, the smaller test areas show a somewhat stronger spatial correlation than the larger ones. In the case of the scattered test set, the dependencies are less pronounced.

Univariate postprocessing is performed exclusively via EMOS, employing a sliding training period of 30 days. For evaluation, we use the same verification tools as before. Results in terms of the ES are shown in Table 4.5. Calibration checks via the usual rank histograms are exemplarily shown for all weather variables in the case of test area II in Figures 4.23 to 4.26.

In contrast to before, we do not explicitly show the results for the EE and the DS, respectively, in this grid-based setting here. These are, however, very similar to what we saw when considering Berlin, Hamburg and Frankfurt.

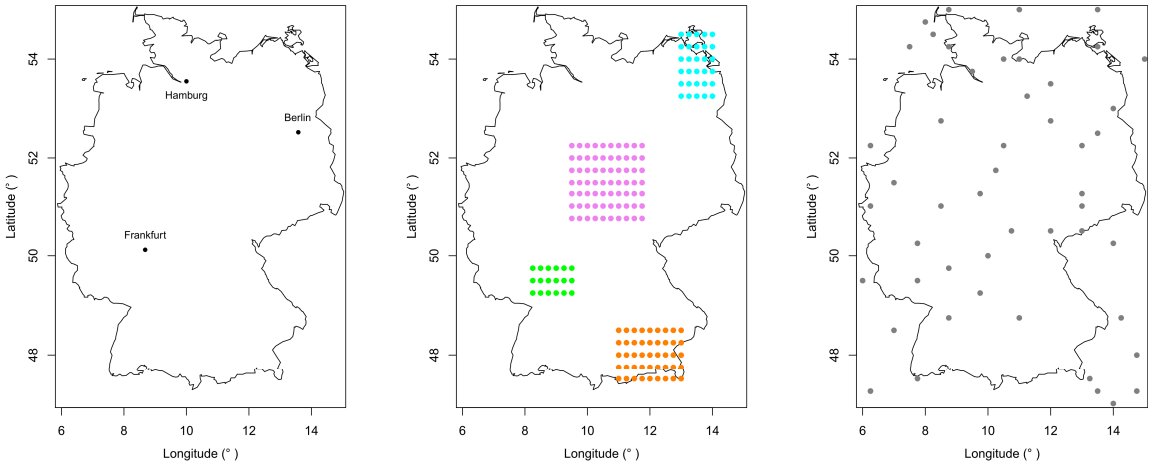
Table 4.4: Test areas over Germany and surrounding regions. Compare Figure 4.22.

Area index	Characteristic regions	Longitude	Latitude	Number L of grid points	Color in Figure 4.22
I	Central Germany, Harz Mountains, North Hesse, Lower Saxony	9.50°–11.75°	50.75°–52.25°	70	Violet
II	Alps, Upper Bavaria	11.00°–13.00°	47.50°–48.50°	45	Orange
III	Baltic Sea, Mecklenburg-Western Pomerania	13.00°–14.00°	53.25°–54.50°	30	Cyan
IV	Rhine-Neckar Area, Forest of Odes	8.25°–9.50°	49.25°–49.75°	18	Green
V	Scattered test set	Various	Various	50	Gray

Specifically, all three postprocessed ensembles outperform the raw ensemble in terms of the EE, except occasionally for u- or v-wind, and they yield similar EE values, which can be expected, as explained before. In the few cases in which the postprocessed ensembles do not outperform the raw ensemble, this is thus very likely due to shortcomings of the EMOS postprocessing already existing in the univariate case with respect to the mean absolute error, with the univariate medians of all the postprocessed ensembles being the same.

Concerning the DS, the corresponding values get smaller with an increasing number L of grid points. For the larger test areas with $L = 50$ and $L = 70$, respectively, the computation of the DS is useless, as it yields effectively a value of zero for all ensembles, due to the high dimensionality of the settings. Again, as we use EMOS postprocessing here, the DS values of the IOQ ensemble are reported as zero throughout anyway, with the same explanation as in Section 4.3.2 (a). In the case of the smaller test areas, for which reasonable DS values can be computed, the raw ensemble is, as expected and explained before, sharper than the ROQ and ECC-Q ensemble, however at the expense of calibration, as we will see later. On the other hand, the ECC-Q ensemble is always much sharper than the ROQ ensemble.

Regarding the ES in the case of temperature and pressure, all postprocessed ensembles outperform the raw ensemble in each test area, with the ECC-Q ensemble performing best throughout. For the four contiguous test areas with pronounced spatial correlations, the IOQ ensemble outperforms the ROQ ensemble, whereas this holds vice versa for the scattered test area with low spatial dependencies. In the case of pressure, the ES values of the IOQ and the ECC-Q ensemble are closer together than for temperature, as the dependencies are again higher. Turning to u- and v-wind, respectively, the ECC-Q ensemble also always outperforms the raw ensemble and performs best among all postprocessed ensembles, except for the scattered test area V, where we have the lowest correlation overall, with the ROQ ensemble performing equally well. Although they are undoubtably existent, the improvements from the unprocessed raw ensemble to the postprocessed ECC-Q ensemble are not that high as for temperature or pressure. For u- and v- wind, respectively, the IOQ ensemble does not perform well and partly even does not manage to outperform the raw ensemble. As the underlying correlations in the case of u- and v-wind, respectively, are mainly at a moderate level, the ROQ ensemble typically outperforms the IOQ ensemble, but occasionally even the ROQ ensemble is outperformed by the raw ensemble. In contrast, the ECC-Q ensemble always outperforms the raw ensemble, and thus, the specific design of the postprocessed

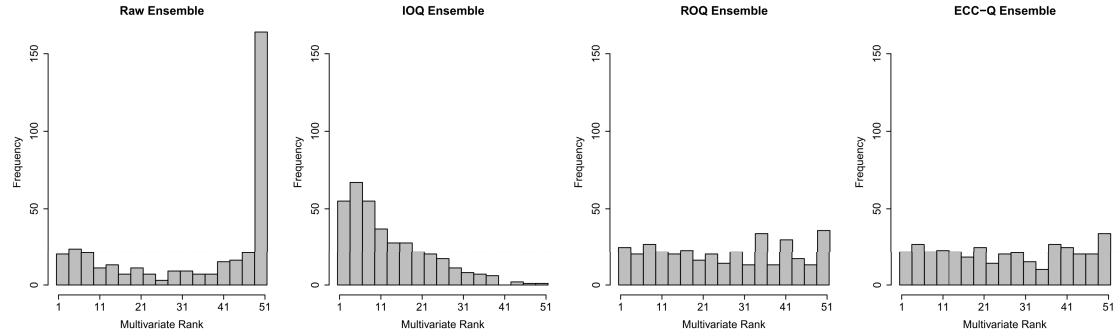


(a) (b) (c)

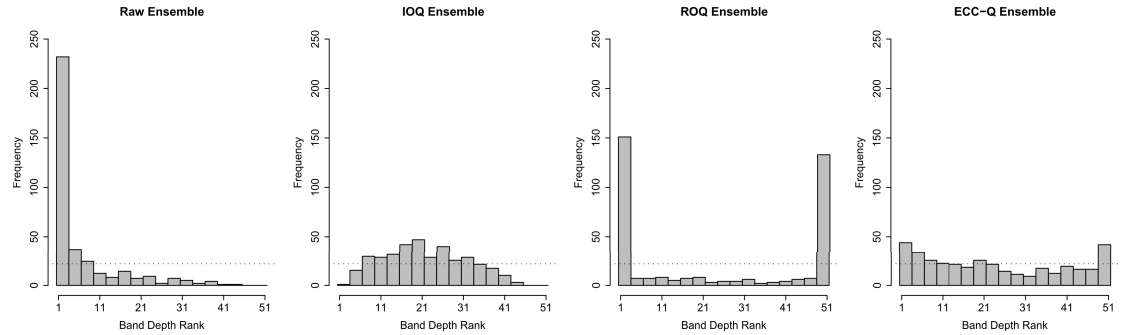
Figure 4.22: Test locations and areas over Germany and surrounding regions, including (a) the three observation sites Berlin, Hamburg and Frankfurt, (b) the four contiguous test regions and (c) the scattered test set. Compare Table 4.4.

Table 4.5: Average ES for 24 hour ahead ensemble forecasts of temperature, pressure, u- and v-wind separately over the test areas I to V according to Table 4.4 and Figure 4.22, respectively. Univariate postprocessing is performed via EMOS, and the scores for the ROQ ensemble are averaged over 100 runs. The results are averaged over the test period from 1 May 2010 to 30 April 2011.

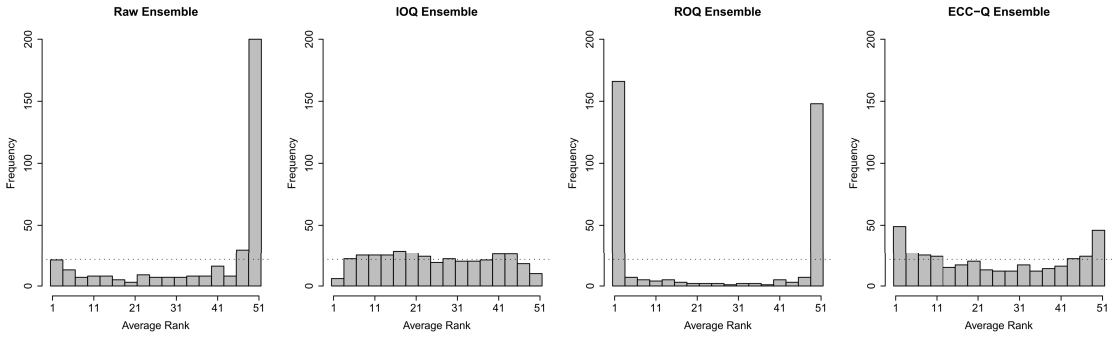
		Area I	Area II	Area III	Area IV	Area V
Temperature (°C)	Raw Ensemble	9.17	14.27	4.44	5.23	8.70
	IOQ Ensemble	8.02	10.71	3.89	4.36	7.53
	ROQ Ensemble	8.11	10.98	3.91	4.50	7.12
	ECC-Q Ensemble	7.84	10.53	3.77	4.29	7.09
Pressure (hPa)	Raw Ensemble	4.80	4.66	2.87	2.39	4.40
	IOQ Ensemble	4.10	3.99	2.39	2.07	3.92
	ROQ Ensemble	4.46	4.20	2.65	2.29	3.88
	ECC-Q Ensemble	4.08	3.94	2.38	2.07	3.78
u-Wind (m/s)	Raw Ensemble	5.12	4.60	3.50	2.45	4.71
	IOQ Ensemble	5.21	4.58	3.51	2.48	5.00
	ROQ Ensemble	5.06	4.47	3.53	2.45	4.60
	ECC-Q Ensemble	4.98	4.37	3.41	2.41	4.60
v-Wind (m/s)	Raw Ensemble	4.74	4.81	3.01	2.43	4.50
	IOQ Ensemble	4.86	4.52	3.11	2.40	4.76
	ROQ Ensemble	4.74	4.40	3.11	2.39	4.38
	ECC-Q Ensemble	4.64	4.30	3.00	2.33	4.38



(a) Multivariate rank histograms

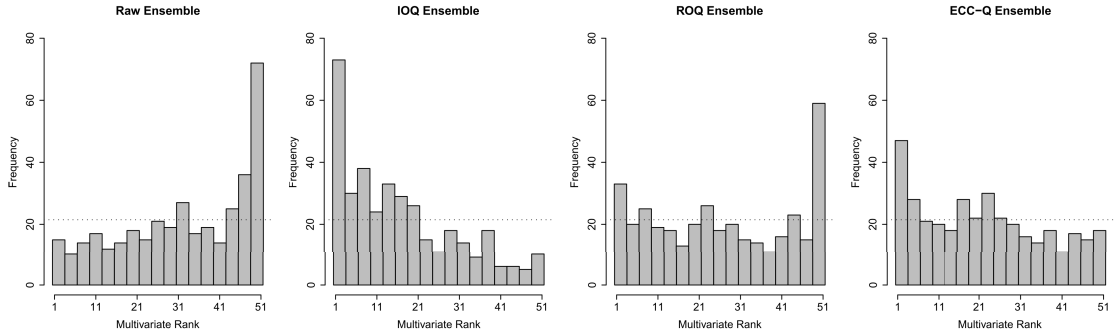


(b) Band depth rank histograms

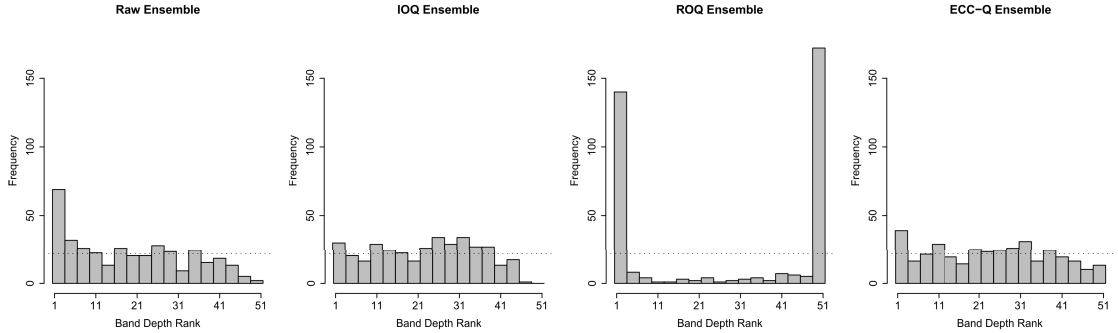


(c) Average rank histograms

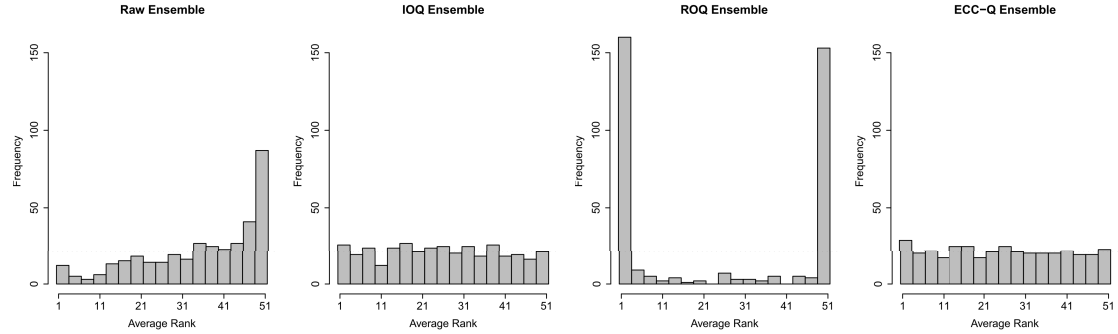
Figure 4.23: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature forecasts for test area II, that is, $L = 45$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.



(a) Multivariate rank histograms

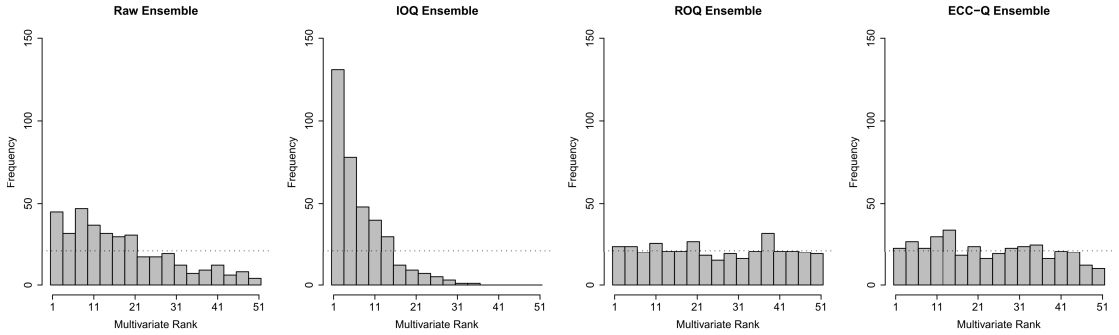


(b) Band depth rank histograms

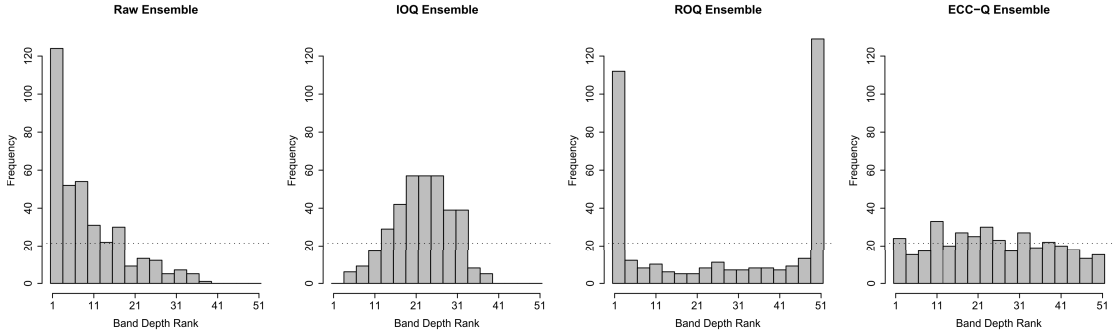


(c) Average rank histograms

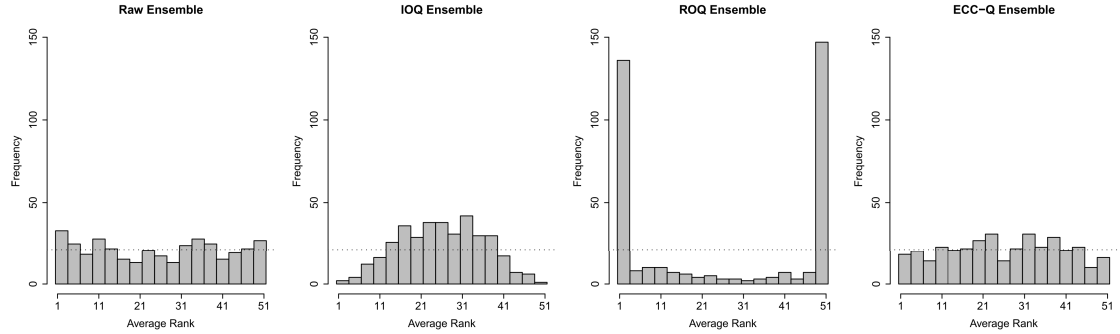
Figure 4.24: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead pressure forecasts for test area II, that is, $L = 45$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.



(a) Multivariate rank histograms

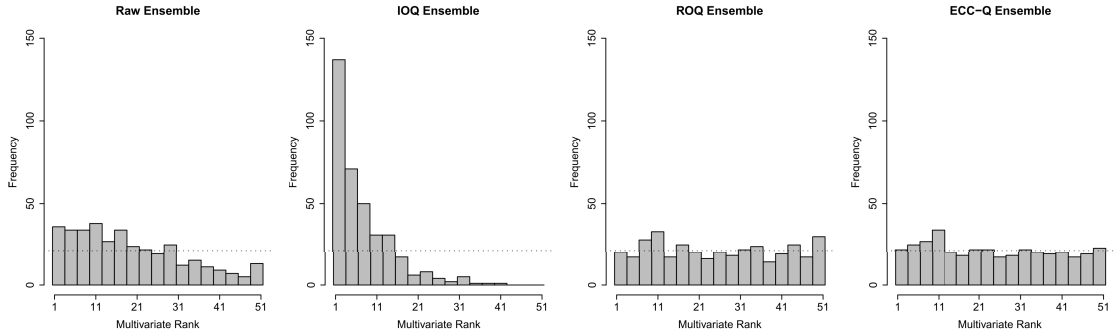


(b) Band depth rank histograms

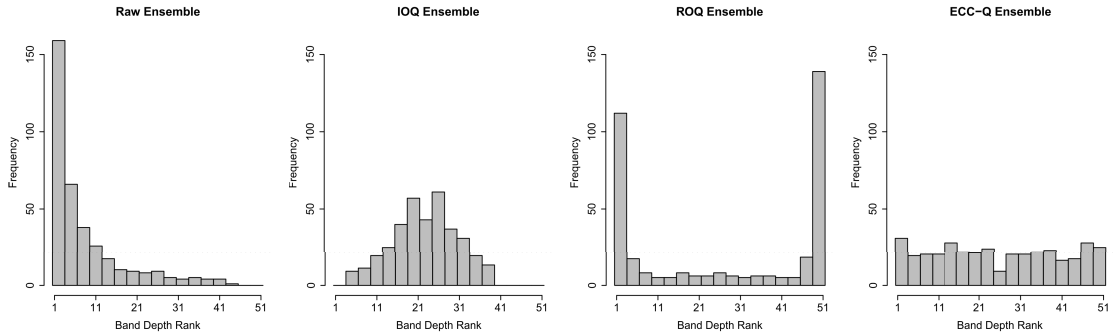


(c) Average rank histograms

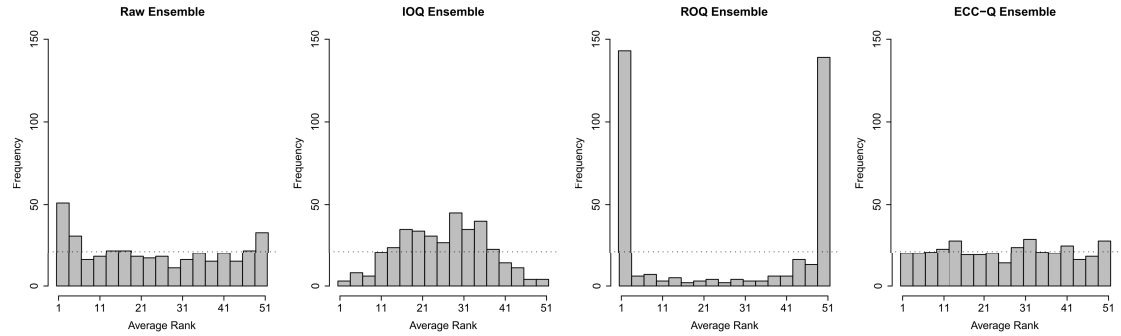
Figure 4.25: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead u-wind forecasts for test area II, that is, $L = 45$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.



(a) Multivariate rank histograms



(b) Band depth rank histograms



(c) Average rank histograms

Figure 4.26: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead v-wind forecasts for test area II, that is, $L = 45$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.

ensembles appears to be of still greater importance if the positive effects of postprocessing are not that pronounced.

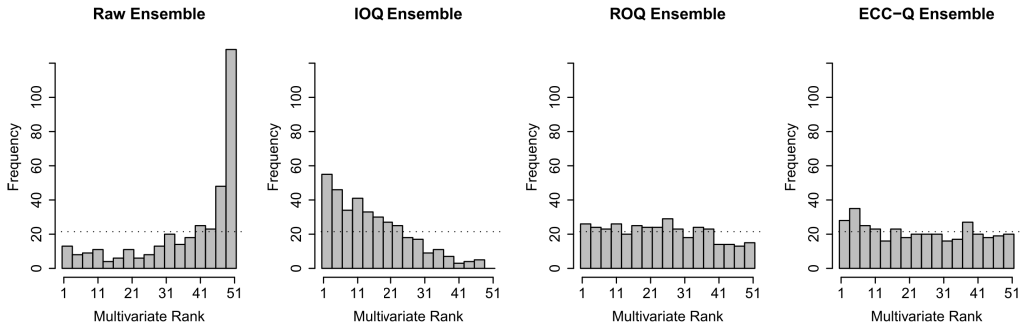
Relating to calibration in the representative example of test area II, the multivariate rank histograms detect the miscalibration of the raw and IOQ ensembles. However, they do not show differences between the ROQ and ECC-Q ensembles, except in case of pressure, where the ECC-Q ensemble is calibrated best. This changes drastically when considering the band depth and average rank histograms, respectively, in which the miscalibration of the ROQ ensemble is clearly shown by a U-shape. In addition, the band depth rank histograms confirm the miscalibration of the raw and the IOQ ensembles. The same holds for the average rank histograms, except for the IOQ ensemble in the case of pressure and the raw ensemble in the case of u-wind, which appear to be calibrated quite well with respect to this verification tool. For both the band depth and average rank histograms, the ECC-Q ensemble reveals the best calibration overall, even though the IOQ ensemble in the case of pressure performs similarly well and the ECC-Q ensemble in the case of temperature reveals slightly U-shaped histograms. Putting the results of all three types of rank histograms together, the ECC-Q ensemble exhibits the best calibration.

In a nutshell, the ECC-Q ensemble performs best in the grid-based settings. It has the lowest ES and fulfills the ideal of maximized sharpness subject to calibration (Gneiting et al., 2007), being well calibrated and best among our ensembles. If the underlying correlation structure is high, as in the case of temperature and pressure over the contiguous test areas, the IOQ ensemble performs better than the ROQ ensemble. This holds vice versa if the dependencies are moderate or low, as for u- and v-wind, respectively, or for the scattered test set.

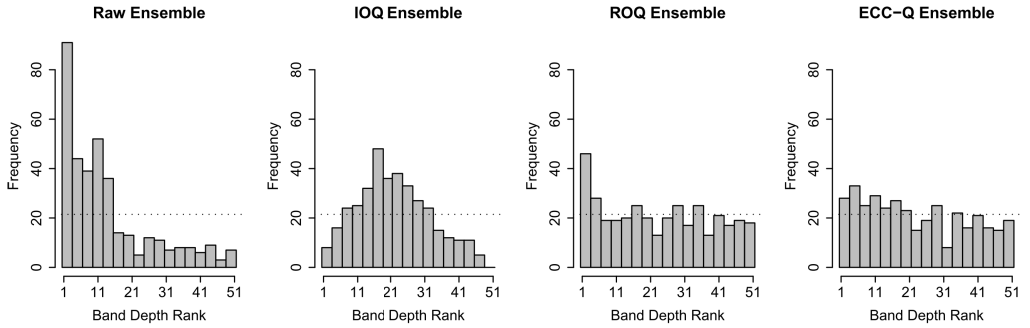
4.3.3 Inter-variable aspects

After having assessed predictive performance in terms of spatial aspects, we now investigate inter-variable settings, in which weather quantities are considered jointly. To this end, we again deal with 24 hour ahead forecasts only and get back to the three observation sites Berlin, Hamburg and Frankfurt, which are now addressed individually. We examine joint temperature and pressure forecasts on the one hand, and joint u- and v-wind forecasts on the other hand, thus being confronted with $L = 2$-dimensional settings. In contrast to some of the spatial correlations, the inter-variable dependencies turn out to be rather low in our considered data set. The most pronounced pattern can be observed between temperature and pressure, which occasionally exhibit a negative correlation, as is illustrated in Figures 4.1 and 4.2, but not each and every day.

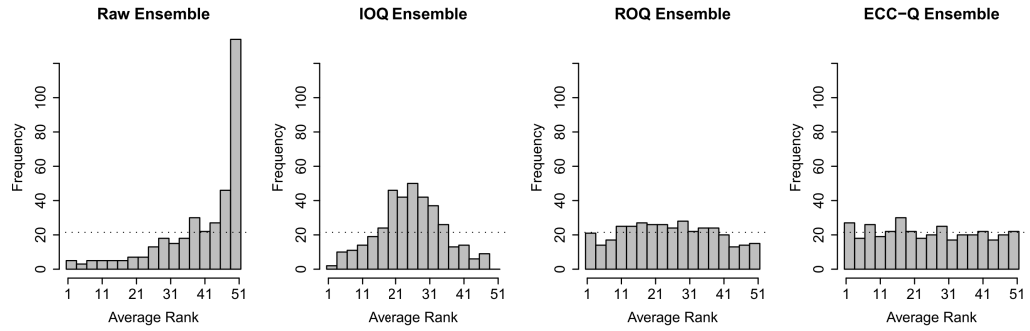
Univariate postprocessing is done via EMOS, using a sliding training period of 30 days, while the test period ranges from 1 May 2010 to 30 April 2011. The overall predictive performance is measured by the ES in Table 4.6, where as usual, the results for the ROQ ensembles are averaged over 100 runs. Note that for the assessment here, the corresponding forecasts and observations have been standardized, to take account of the different magnitudes and units. Specifically, each forecast x_m is transformed to the normalized forecast x'_m given by $x'_m := (x_m - \mu_y)/s_y$ for each ensemble member m and each verification day, with μ_y and s_y denoting the empirical mean and the empirical standard deviation, respectively, of the observations over the one-year test period from 1 May 2010 to 30 April 2011. Analogously, we proceed with the observations. Thus, the ES values do not have a unit in this



(a) Multivariate rank histograms

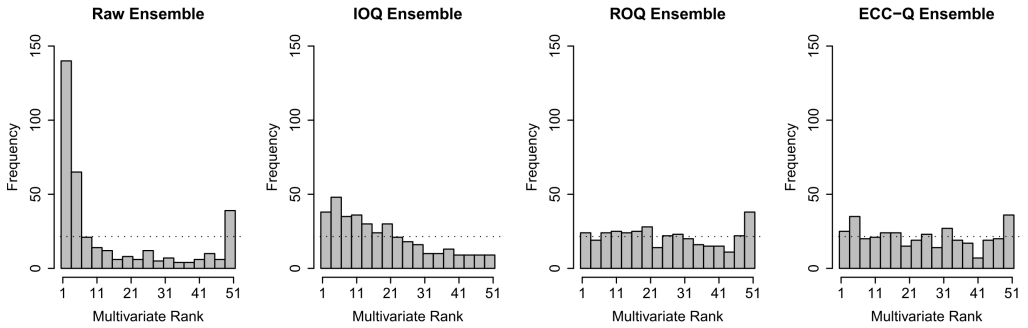


(b) Band depth rank histograms

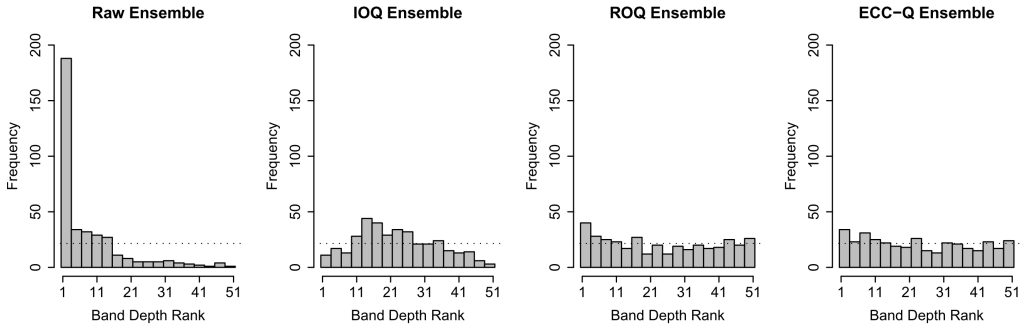


(c) Average rank histograms

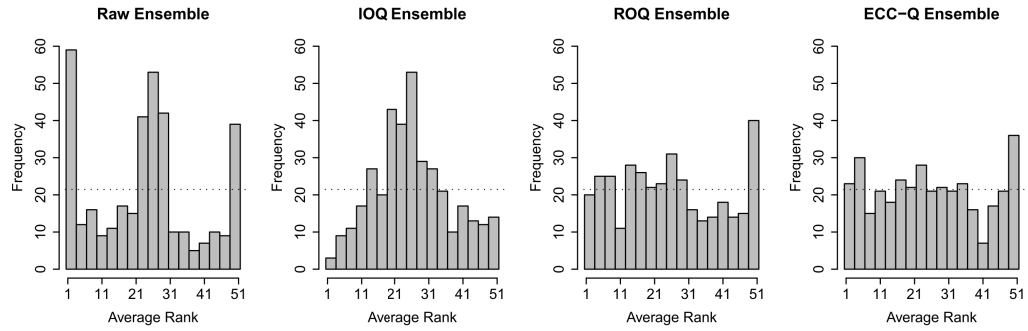
Figure 4.27: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead joint temperature and pressure forecasts at Berlin, that is, $L = 2$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.



(a) Multivariate rank histograms



(b) Band depth rank histograms



(c) Average rank histograms

Figure 4.28: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead joint u- and v-wind forecasts at Frankfurt, that is, $L = 2$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.

Table 4.6: Average ES for 24 hour ahead standardized joint forecasts of temperature and pressure and joint forecasts of u- and v-wind, respectively, at the individual stations of Berlin, Hamburg and Frankfurt. The results are averaged over the test period from 1 May 2010 to 30 April 2011, and the scores for the ROQ ensembles are averaged over 100 runs.

	Temperature/Pressure jointly			u-/v-Wind jointly		
	Berlin	Hamburg	Frankfurt	Berlin	Hamburg	Frankfurt
Raw Ensemble	0.176	0.166	0.201	0.543	0.504	0.787
IOQ Ensemble	0.140	0.136	0.155	0.468	0.438	0.616
ROQ Ensemble	0.133	0.130	0.147	0.443	0.414	0.588
ECC-Q Ensemble	0.133	0.130	0.147	0.447	0.418	0.590

case. Calibration is assessed via the usual rank histograms, where we exemplarily show the results for temperature and pressure jointly at Berlin in Figure 4.27 and for u- and v-wind jointly at Frankfurt in Figure 4.28.

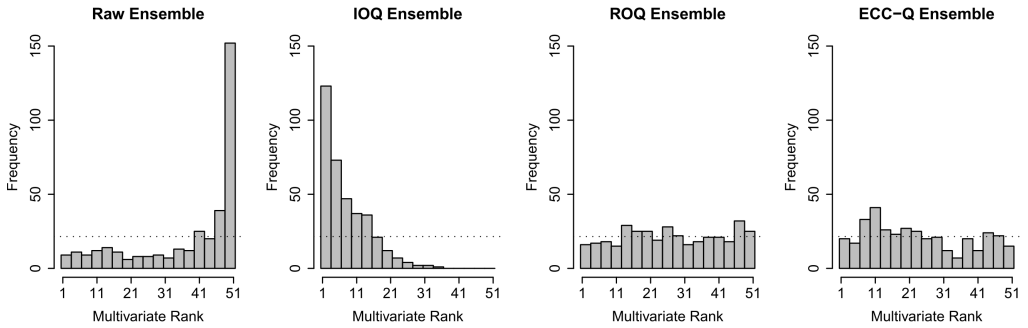
With respect to the ES, all postprocessed ensembles outperform the unprocessed raw ensemble, while the ROQ and ECC-Q ensembles generally perform better than the IOQ ensemble. When considering temperature and pressure jointly, dealing with a moderate correlation structure, the ROQ and ECC-Q ensemble perform equally well in terms of the ES. However, in the case of joint u- and v-wind, revealing a rather low dependence pattern, the ROQ ensemble outperforms the ECC-Q ensemble.

Regarding calibration only, both the raw and the IOQ ensemble turn out to be uncalibrated, with the IOQ ensemble overestimating the correlation structure. For the joint temperature and pressure setting, the ECC-Q ensemble appears to be slightly better calibrated than the ROQ ensemble, which can be seen best in the band depth rank histogram. In contrast, in the case of the joint u- and v-wind scenario, differences between the ROQ and ECC-Q ensemble in terms of calibration are hardly visible, with ECC-Q maybe performing very slightly better.

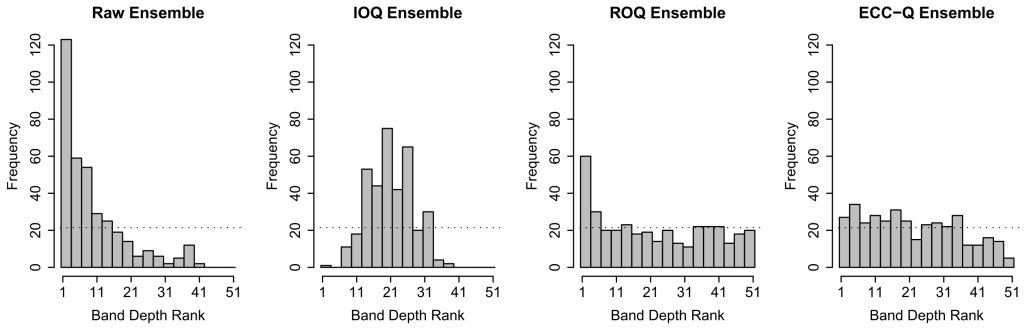
4.3.4 Joint spatial and inter-variable aspects

Now, both spatial and inter-variable aspects are considered jointly, while the prediction horizon of 24 hours still remains fixed. Specifically, we investigate three different settings. First, we consider joint temperature and pressure forecasts at Berlin and Hamburg simultaneously, thus an $L = 4$-dimensional scenario. Second, we look at the $L = 6$-dimensional setting including joint temperature and pressure forecasts at Berlin, Hamburg and Frankfurt together. Third, we combine all weather quantities and locations and hence consider joint temperature, pressure, u- and v-wind forecasts at Berlin, Hamburg and Frankfurt simultaneously in an $L = 12$-dimensional situation. The correlations structures for spatial and inter-variable aspects, respectively, discussed in Sections 4.3.2 and 4.3.3 extend to the combined settings we deal with here. That is, the $L = 4$-dimensional scenario shows the strongest dependencies, while for the $L = 12$-dimensional example, correlation is rather low.

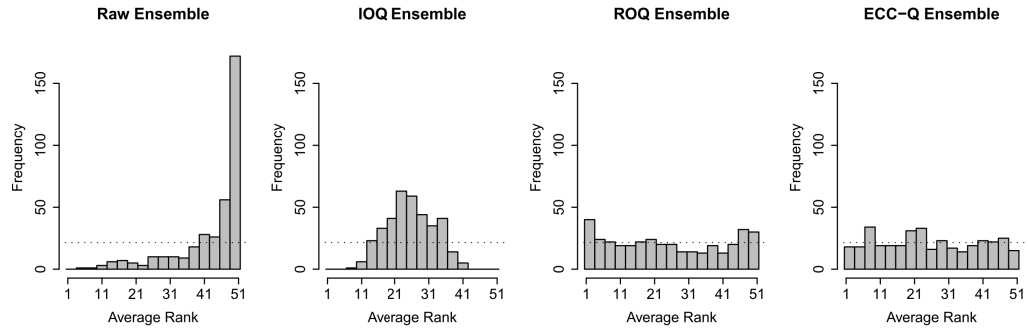
Univariate postprocessing is performed via EMOS, employing a rolling training period of 30 days, while the test period ranges from 1 May 2010 to 30 April 2011. The overall predictive performance is measured by the ES in Table 4.7, where as usual, the results for the ROQ



(a) Multivariate rank histograms

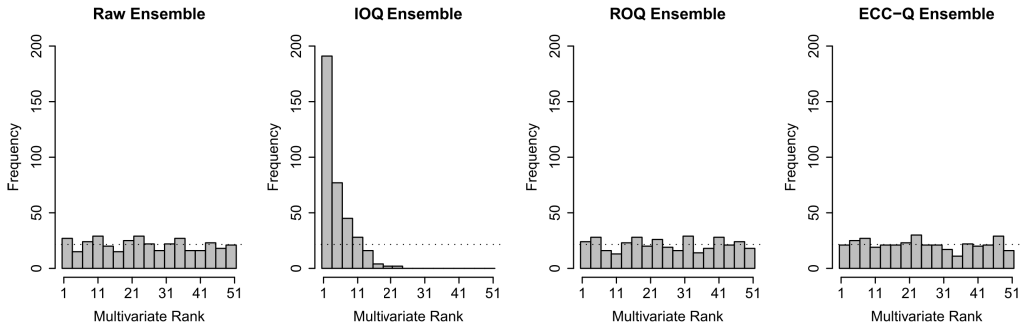


(b) Band depth rank histograms

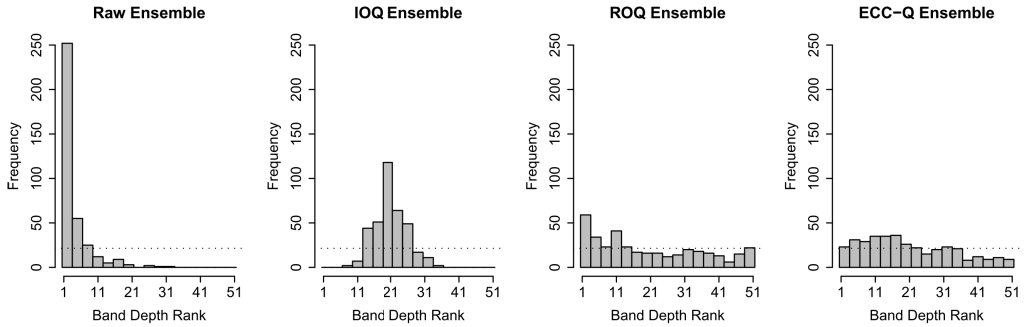


(c) Average rank histograms

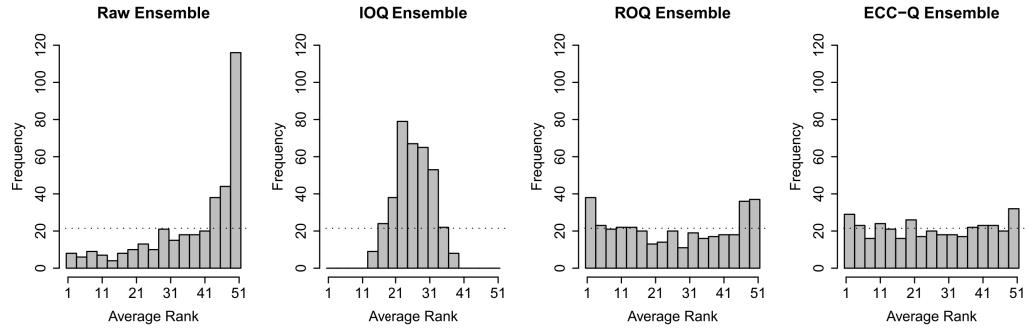
Figure 4.29: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature and pressure forecasts jointly at Berlin, Hamburg and Frankfurt, that is, $L = 6$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.



(a) Multivariate rank histograms



(b) Band depth rank histograms



(c) Average rank histograms

Figure 4.30: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature, pressure, u- and v-wind forecasts jointly at Berlin, Hamburg and Frankfurt, that is, $L = 12$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.

Table 4.7: Average ES for different combinations of 24 hour ahead standardized joint forecasts of temperature (temp), pressure (press), u- and v-wind at Berlin (Ber), Hamburg (Ham) and Frankfurt (Fra). The results are averaged over the test period from 1 May 2010 to 30 April 2011, and those for the ROQ ensembles are averaged over 100 runs.

	Ber/Ham jointly Temp/Press jointly	Ber/Ham/Fra jointly Temp/Press jointly	Ber/Ham/Fra jointly Temp/Press/u-/v-Wind jointly
Raw Ensemble	0.256	0.342	1.211
IOQ Ensemble	0.212	0.282	1.037
ROQ Ensemble	0.199	0.262	0.960
ECC-Q Ensemble	0.199	0.262	0.965

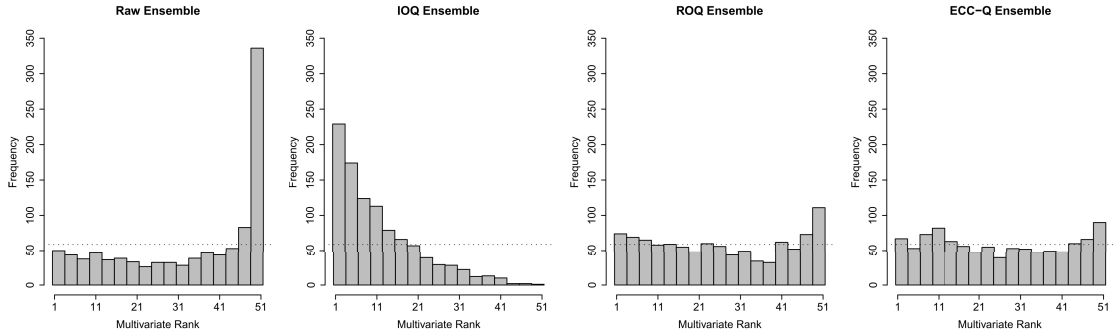
ensembles are averaged over 100 runs. Moreover, the scores are standardized to account for the different magnitudes and units of the corresponding weather variables, where the normalization of the involved forecasts and observations is achieved as described in Section 4.3.3. Calibration is checked via the usual rank histograms, where we restrict our attention to the $L = 6$- and $L = 12$-dimensional scenarios in Figures 4.29 and 4.30, respectively.

In terms of the ES, all postprocessed ensembles outperform the raw ensemble in each scenario, and the ROQ and ECC-Q ensembles generally outperform the IOQ ensemble. In the $L = 4$- and $L = 6$-dimensional settings considering temperature and pressure jointly, the ROQ and the ECC-Q ensembles show equally good performances, while in the $L = 12$-dimensional setting additionally including u- and v-wind, the ROQ ensemble performs a bit better than the ECC-Q ensemble. All in all, this is consistent with the results in Sections 4.3.2 and 4.3.3.

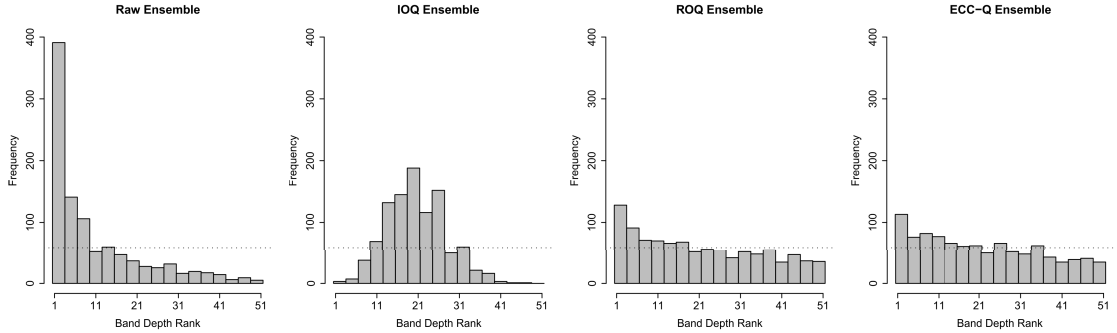
Concerning calibration in the $L = 6$-dimensional setting, the raw and IOQ ensemble turn out to be uncalibrated, with the latter showing strongly inverse U-shaped band depth and average rank histograms, respectively, pointing at a strong overestimation of the correlation structure. The ECC-Q ensemble is calibrated best overall, while the ROQ ensemble appears to be a bit underdispersed, as indicated by the corresponding band depth and average rank histograms. At large, the results for the $L = 6$-dimensional scenario apply analogously in the $L = 12$-dimensional setting. Interestingly, the multivariate rank histogram fails to detect the miscalibration of the raw ensemble, which becomes obvious when looking at the band depth and average rank histograms. Moreover, the multivariate rank histogram does not distinguish between the ROQ and the ECC-Q ensemble and suggests calibration in both cases, as it wrongly does for the raw ensemble. This is also fixed by the band depth and average rank histograms, which reveal that the ROQ ensemble is uncalibrated, whereas the ECC-Q ensemble is calibrated fairly well and best among the ensembles. These shortcomings of the multivariate rank histogram are likely due to the involvement of four different weather quantities and also the comparably high dimensionality of $L = 12$ in this setting, thus confirming the corresponding findings in Thorarinsdottir et al. (2014). In contrast, the multivariate rank histogram in the $L = 12$-dimensional setting manages at least to detect the miscalibration of the IOQ ensemble, as the other two rank histograms also do.

4.3.5 Temporal aspects

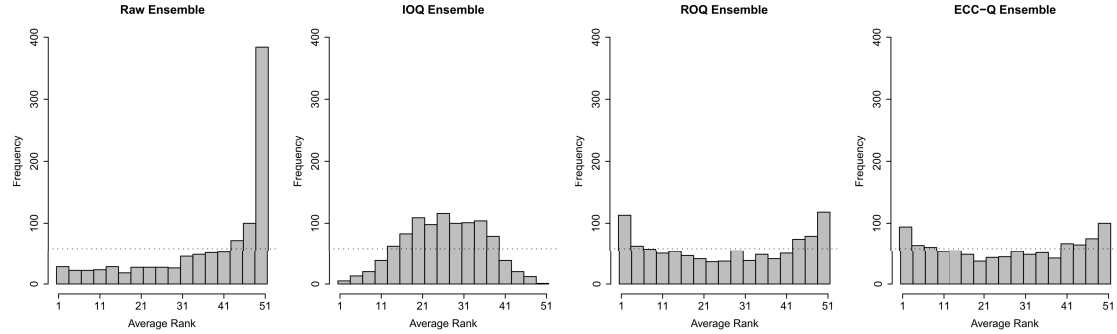
Finally, we investigate the predictive performance of ECC and the reference ensembles with respect to temporal dependence structures. To this end, we consider joint 24, 48, 72 and



(a) Multivariate rank histograms



(b) Band depth rank histograms



(c) Average rank histograms

Figure 4.31: (a) Multivariate, (b) band depth and (c) average rank histograms for joint 24, 48, 72 and 96 hour ahead temperature forecasts at Hamburg, that is, $L = 4$, over the 1000 initialization days from 1 April 2010 to 25 December 2012. Univariate postprocessing is performed via EMOS.

Table 4.8: Average ES (in °C) and DSS for joint 24, 48, 72 and 96 hour ahead forecasts of temperature at Berlin, Hamburg and Frankfurt, respectively. The results are averaged over 1000 initialization days from 1 April 2010 to 25 December 2012. Note that the calculation of the DSS for the IOQ ensemble fails due to the computational singularity of the involved matrix, that is, one cannot calculate the inverse and determinant. Hence, the corresponding results are reported as "not available (NA)". The results for the ROQ ensembles are averaged over 100 runs.

	ES			DSS		
	Berlin	Hamburg	Frankfurt	Berlin	Hamburg	Frankfurt
Raw Ensemble	2.601	2.379	2.815	19.96	18.55	24.57
IOQ Ensemble	2.354	2.273	2.540	NA	NA	NA
ROQ Ensemble	2.228	2.137	2.396	9.29	8.54	9.66
ECC-Q Ensemble	2.224	2.134	2.394	9.20	8.44	9.66

96 hour ahead temperature forecasts initialized at the same date, for each of the locations Berlin, Hamburg and Frankfurt separately, thus dealing with $L = 4$-dimensional settings. Univariate postprocessing is done via EMOS, using a sliding training period of 30 days. This time, our test period comprises the 1000 initialization dates from 1 April 2010 to 25 December 2012. The results for the scores as overall performance measures can be found in Table 4.8. In addition to the ES, we also show the values for the Dawid-Sebastiani score (DSS) computed according to (2.10) with empirical mean vector and empirical covariance matrix, respectively, an alternative score that will also be used in Chapter 5. For both scores, the results for the ROQ ensemble are as usual averaged over 100 runs.

Concerning the ES, all postprocessed ensembles outperform the unprocessed raw ensemble, with the ECC-Q ensemble performing best. Among the postprocessed ensembles, both the ECC-Q and the ROQ ensemble outperform the IOQ ensemble, whereas the differences between the ES of the ROQ and ECC-Q ensembles are rather minor, with ECC-Q yielding the best ES.

With respect to the DSS, the ROQ and ECC-Q ensembles also outperform the raw ensemble. Note that the calculation of the DSS for the IOQ ensemble fails due to the singularity of the covariance matrix, with the same explanation as for the calculation of the DS in Section 4.3.2. This prevents us from computing the required inverses. For Frankfurt, the ROQ and the ECC-Q ensembles yield the same DSS, whereas in the case of Berlin and Hamburg, ECC-Q outperforms the ROQ ensemble, in fact more clearly than with respect to the ES. Hence, the DSS is in these situations more sensitive and distinguishes better between the two ensembles than the ES.

To check calibration, we exemplarily focus on Hamburg in Figure 4.31. The corresponding histograms for the other two sites Berlin and Frankfurt look very similar. All three histogram types show that the raw and the IOQ ensemble are uncalibrated. The inverse U-shape of the band depth and average rank histograms, respectively, in the case of the IOQ ensemble again indicates an overestimation of the correlation structure. The ECC-Q ensemble performs best in terms of calibration. With respect to the multivariate rank histogram, the ECC-Q ensemble appears to be fairly well calibrated, while with respect to the band depth and average rank histogram, its calibration is not perfect, but still best.

While the standard EMOS implementation has been used in this case study here, constraints can be put on the EMOS parameters, such that they vary smoothly across look-ahead times, ensuring the temporal coherence of the postprocessed marginal predictive distributions.

4.3.6 Conclusions

We now summarize and discuss the main results of the case studies in this section.

The ECC-Q ensemble mostly achieves the goal of probabilistic forecasting according to Gneiting et al. (2007), namely maximizing the sharpness of the forecast subject to calibration. While the raw and the IOQ ensemble are typically sharp, they lack calibration, except perhaps for pressure. The ROQ ensemble is at least partly uncalibrated and less sharp than the ECC-Q ensemble in general.

The underlying correlation structures of both the ensemble forecasts and the observations have an impact on the performances of the IOQ, ROQ and ECC-Q ensembles. The IOQ ensemble performs very well if there is an extremely strong, close to linear, dependence. In this case, the IOQ ensemble might even slightly outperform the ECC-Q ensemble. The less correlation structure is existent, the worse is the overall performance of the IOQ ensemble, which then may fail drastically, in that it is occasionally outperformed by the unprocessed raw ensemble, while the ROQ and ECC-Q ensembles are not. In contrast, the ROQ ensemble performs well in situations, in which the correlation is weak, here for instance for (joint) u- and v-wind at Berlin, Hamburg and Frankfurt, individually and together. However, the ROQ ensemble is clearly the worst among the three postprocessed ensembles when being confronted with strong dependence structures. The ECC-Q ensemble, being based on the rank correlation pattern of the raw ensemble, can be thought of as in between the IOQ ensemble, which assumes maximal dependence, on the one hand, and the ROQ ensemble, which assumes no dependence, on the other hand. The ECC-Q ensemble generally performs well and best in by far the majority of the situations considered in our case study. It might be occasionally slightly outperformed by the IOQ and ROQ ensembles in cases of extremely high and weak correlation, respectively. However, the ECC-Q ensemble performs not at all much worse than the IOQ or ROQ ensemble then, but remarkably better than the corresponding other reference ensemble. Hence, the ECC-Q ensemble provides a reasonable, solid and reliable tool for multivariate ensemble postprocessing, in particular if the underlying correlation structure is a priori unknown.

With that said, the employment of ECC-Q turned out to be most beneficial in our case studies in the settings dealing with purely spatial or temporal dependence structures, and to some lesser extent in the situations involving multiple weather variables.

The results of our case studies confirm the findings in Thorarinsdottir et al. (2014) about the interpretation of band depth and average rank histograms, respectively. In particular, it has been illustrated that an over- and underestimation of the underlying correlation structure, due to the specific design of the postprocessed ensembles, results in an inverse U-shaped and U-shaped band depth and average rank histogram, respectively. Moreover, we have seen that the multivariate rank histogram tends to be not able to distinguish between the ensembles, especially in rather high dimensions. In particular, the uniformity of a multivariate rank histogram does not necessarily mean that the corresponding ensemble is calibrated. Hence, it is generally advisable to check calibration via multiple verification tools, as is also

suggested by Thorarinsdottir et al. (2014).

The ECC-Q ensemble is often clearly better calibrated than the ROQ ensemble in empirical checks via rank histograms, even for short test periods. In addition, the ECC-Q ensemble is generally sharper than the ROQ ensemble. However, these benefits appear to be not always reflected in the corresponding values of the ES as an overall performance measure, which occasionally reveals weaknesses in distinguishing between the ROQ and ECC-Q ensemble. This effect also emerges in a recent case study of Scheuerer and Hamill (2014). It is a highly topical issue to examine and possibly improve the discrimination ability of the ES (Pinson and Tastu, 2013), and there is much room for future work into this direction (Scheuerer and Hamill, 2014).

4.4 ECC variants when the desired ensemble size after post-processing exceeds that of the raw ensemble

In the ECC method presented in Section 4.1, which is referred to as the standard ECC approach in this section here, the postprocessed ECC ensemble is constrained to consist of exactly the same number M of members as the unprocessed raw ensemble. Sometimes, it might however be required to produce a postprocessed ensemble of a size differing from that of the raw ensemble, while preserving dependence structures as in standard ECC.

If a reduction of the ensemble size after postprocessing is desired, we could just take $N < M$ out of the M exchangeable raw ensemble members and then apply standard ECC to the N-member ensemble. The corresponding N members can be chosen either randomly or according to a specific scheme. For instance, one could generally take the first N of the M ensemble members.

However, the raw ensemble size often is quite small such that it is typically of interest to increase the size of the ensemble rather than to reduce it. In the following, we propose a modification of the standard ECC technique, called extended ECC, to achieve this and test it in a case study. Moreover, we briefly discuss two alternative approaches based on recycling and the use of so-called lagged ensembles, respectively.

4.4.1 The extended ECC approach

To address the challenge of creating an ECC ensemble whose size exceeds that of the unprocessed raw ensemble, we present the following modification of the standard ECC method, valid for ensembles consisting of exchangeable members and named extended ECC in what follows.
Let $M \in \mathbb{N}$ denote the number of members in the raw ensemble, and let $N \in \mathbb{N}$, $N \geq M$, be the requested number of members in the postprocessed ECC ensemble. As in the standard ECC method, we assume all M raw ensemble members to be exchangeable. Then, $N = r \cdot M + P$, where $r \in \mathbb{N}$ and $P \in \mathbb{N}_0$, $0 \leq P < M$, are uniquely determined non-negative integers. Obviously, standard ECC with $N = M$ can be interpreted as a special case of extended ECC with $r = 1$ and $P = 0$. For $N > M$, we now consider the following two scenarios and describe how to modify standard ECC in the respective cases.

(A) The requested extended ECC ensemble size is an integer multiple of the raw ensemble size, that is, $r \geq 2$ and $P = 0$.

(B) The requested extended ECC ensemble size is not an integer multiple of the raw ensemble size, that is, $r \geq 1$ and $P \neq 0$.

Scenario (A)

Suppose that $N = r \cdot M$, $r \geq 2$. As in the standard ECC approach, we are given the raw ensemble forecast $x_1^\ell, \ldots, x_M^\ell$ for each weather variable $i \in \{1, ..., I\}$, each location $j \in \{1, ..., J\}$ and each prediction horizon $k \in \{1, ..., K\}$ individually, summarized in the multi-index $\ell := (i, j, k)$. This defines a permutation $\sigma_\ell(m) := \text{rank}(x_m^\ell)$ for $m \in \{1, \ldots, M\}$, with ties resolved at random. For each ℓ, we employ state-of-the-art univariate ensemble postprocessing methods, such as BMA or EMOS, to obtain a predictive CDF F_ℓ. Then, we generate a sample $\tilde{x}_1^\ell, \ldots, \tilde{x}_N^\ell$ from each CDF F_ℓ not of size M as in standard ECC, but of size N. Since applying quantization scheme (T) appears to be inappropriate due to $N > M$, and we prefer scheme (Q) to scheme (R) based on the considerations in Section 4.2, this sampling is performed according to scheme (Q) by taking the equally spaced $\frac{1}{N+1}-, \ldots, \frac{N}{N+1}-$quantiles of F_ℓ, that is,

$$\tilde{x}_1^\ell := F_\ell^{-1}\left(\frac{1}{N+1}\right), \ldots, \tilde{x}_N^\ell := F_\ell^{-1}\left(\frac{N}{N+1}\right).$$

As we have $N = r \cdot M$, $r \geq 2$, we can divide our N-sized sample

$$\tilde{\boldsymbol{x}}^\ell := (\tilde{x}_1^\ell, \ldots, \tilde{x}_N^\ell)$$

into r subsamples

$$\tilde{\boldsymbol{x}}^{\ell,[1]} := (\tilde{x}_1^{\ell,[1]}, \ldots, \tilde{x}_M^{\ell,[1]}), \ldots, \tilde{\boldsymbol{x}}^{\ell,[r]} := (\tilde{x}_1^{\ell,[r]}, \ldots, \tilde{x}_M^{\ell,[r]}),$$

each consisting of M elements. We propose the following two approaches (i) and (ii), respectively, to achieve this.

(i) The sample $\tilde{\boldsymbol{x}}^\ell$ is randomly divided into r subsamples, each consisting of M elements. That is, if λ_ℓ denotes a random permutation of $\{1, \ldots, N\}$, the r subsamples are given by

$$\begin{aligned}
\tilde{\boldsymbol{x}}^{\ell,[1]} &:= (\tilde{x}_1^{\ell,[1]}, \ldots, \tilde{x}_M^{\ell,[1]}) := (\tilde{x}_{\lambda_\ell(1)}^\ell, \ldots, \tilde{x}_{\lambda_\ell(M)}^\ell), \\
\tilde{\boldsymbol{x}}^{\ell,[2]} &:= (\tilde{x}_1^{\ell,[2]}, \ldots, \tilde{x}_M^{\ell,[2]}) := (\tilde{x}_{\lambda_\ell(M+1)}^\ell, \ldots, \tilde{x}_{\lambda_\ell(2M)}^\ell), \\
&\vdots \\
\tilde{\boldsymbol{x}}^{\ell,[r]} &:= (\tilde{x}_1^{\ell,[r]}, \ldots, \tilde{x}_M^{\ell,[r]}) := (\tilde{x}_{\lambda_\ell((r-1)M+1)}^\ell, \ldots, \tilde{x}_{\lambda_\ell(rM)}^\ell),
\end{aligned}$$

with $N = r \cdot M$.

(ii) We partition the sample $\tilde{\boldsymbol{x}}^\ell$ into r subsamples such that each of these consists of M equidistant quantiles. This can be performed as follows:

- The first subsample $\tilde{\boldsymbol{x}}^{\ell,[1]}$ comprises the $\frac{r\nu+1}{N+1}$-quantiles of F_ℓ, where $\nu \in \{0, \ldots, M-1\}$ runs through all non-negative integers from 0 to $M-1$, that is,

$$\tilde{\boldsymbol{x}}^{\ell,[1]} := (\tilde{x}_1^{\ell,[1]}, \tilde{x}_2^{\ell,[1]}, \tilde{x}_3^{\ell,[1]} \ldots, \tilde{x}_M^{\ell,[1]}) := (\tilde{x}_1^\ell, \tilde{x}_{r+1}^\ell, \tilde{x}_{2r+1}^\ell, \ldots, \tilde{x}_{r(M-1)+1}^\ell).$$

- The second subsample $\tilde{\boldsymbol{x}}^{\ell,[2]}$ comprises the $\frac{r\nu+2}{N+1}$-quantiles of F_ℓ, with $\nu \in \{0, \ldots, M-1\}$, that is,

$$\tilde{\boldsymbol{x}}^{\ell,[2]} := (\tilde{x}_1^{\ell,[2]}, \tilde{x}_2^{\ell,[2]}, \tilde{x}_3^{\ell,[2]} \ldots, \tilde{x}_M^{\ell,[2]}) := (\tilde{x}_2^{\ell}, \tilde{x}_{r+2}^{\ell}, \tilde{x}_{2r+2}^{\ell}, \ldots, \tilde{x}_{r(M-1)+2}^{\ell}).$$

- $\cdots$
- The r-th subsample $\tilde{\boldsymbol{x}}^{\ell,[r]}$ comprises the $\frac{r\nu+r}{N+1}$-quantiles of F_ℓ, with $\nu \in \{0, \ldots, M-1\}$, that is,

$$\tilde{\boldsymbol{x}}^{\ell,[r]} := (\tilde{x}_1^{\ell,[r]}, \tilde{x}_2^{\ell,[r]}, \tilde{x}_3^{\ell,[r]} \ldots, \tilde{x}_M^{\ell,[r]}) := (\tilde{x}_r^{\ell}, \tilde{x}_{2r}^{\ell}, \tilde{x}_{3r}^{\ell}, \ldots, \tilde{x}_{Mr}^{\ell}),$$

  where $N = r \cdot M$.

Since each of the r subsamples obtained by variant (i) or (ii), respectively, consists of M elements, we can interpret each subsample as a postprocessed ensemble of size M and apply the standard ECC-Q approach to each. That is, each of the r postprocessed M-member ensembles is rearranged separately with respect to the corresponding ranks determined by the raw ensemble. For each ℓ, we thus get r different standard ECC-Q ensembles $\hat{\boldsymbol{x}}^{\ell,[1]}, \hat{\boldsymbol{x}}^{\ell,[2]}, \ldots, \hat{\boldsymbol{x}}^{\ell,[r]}$, consisting of M members each, via

$$\begin{aligned}
\hat{\boldsymbol{x}}^{\ell,[1]} &:= (\hat{x}_1^{\ell,[1]}, \ldots, \hat{x}_M^{\ell,[1]}) := (\tilde{x}_{(\sigma_\ell(1))}^{\ell,[1]}, \tilde{x}_{(\sigma_\ell(2))}^{\ell,[1]}, \ldots, \tilde{x}_{(\sigma_\ell(M))}^{\ell,[1]}),\\
\hat{\boldsymbol{x}}^{\ell,[2]} &:= (\hat{x}_1^{\ell,[2]}, \ldots, \hat{x}_M^{\ell,[2]}) := (\tilde{x}_{(\sigma_\ell(1))}^{\ell,[2]}, \tilde{x}_{(\sigma_\ell(2))}^{\ell,[2]}, \ldots, \tilde{x}_{(\sigma_\ell(M))}^{\ell,[2]}),\\
&\vdots\\
\hat{\boldsymbol{x}}^{\ell,[r]} &:= (\hat{x}_1^{\ell,[r]}, \ldots, \hat{x}_M^{\ell,[r]}) := (\tilde{x}_{(\sigma_\ell(1))}^{\ell,[r]}, \tilde{x}_{(\sigma_\ell(2))}^{\ell,[r]}, \ldots, \tilde{x}_{(\sigma_\ell(M))}^{\ell,[r]}).
\end{aligned}$$

These can then be aggregated in an extended ECC-Q ensemble $\hat{\boldsymbol{x}}^\ell$ that comprises $N = r \cdot M$ members via

$$\hat{\boldsymbol{x}}^\ell := (\hat{x}_1^\ell, \ldots, \hat{x}_N^\ell) := (\hat{\boldsymbol{x}}^{\ell,[1]}, \hat{\boldsymbol{x}}^{\ell,[2]}, \ldots, \hat{\boldsymbol{x}}^{\ell,[r]}).$$

Thereby, the order of the r standard ECC-Q ensembles in the aggregation can be in principle chosen arbitrarily, but has to be consistent for each ℓ. We employ arguably the most natural aggregation order above.

The variants (i) and (ii) are referred to as extended ECC-Q random and extended ECC-Q equidistant, respectively, in what follows.

As an illustrative example, we consider an ensemble with $M = 10$ members based on forecasts provided by the ECMWF. In fact, we just take the first 10 members of the 50-member ECMWF ensemble as the underlying raw ensemble of size 10 here. We want to increase the ensemble size such that the postprocessed ensemble has size $N = 30 = 3 \cdot 10$. In the exemplary scatterplots in Figure 4.32, we focus on forecasts of temperature at Hamburg and Berlin, valid 1:00 am on 24 November 2010. Univariate postprocessing is performed by EMOS, using a training period of 30 days. We show the plots for

- the raw ensemble consisting of $M = 10$ members,
- the standard ECC-Q ensemble consisting of $M = 10$ members,

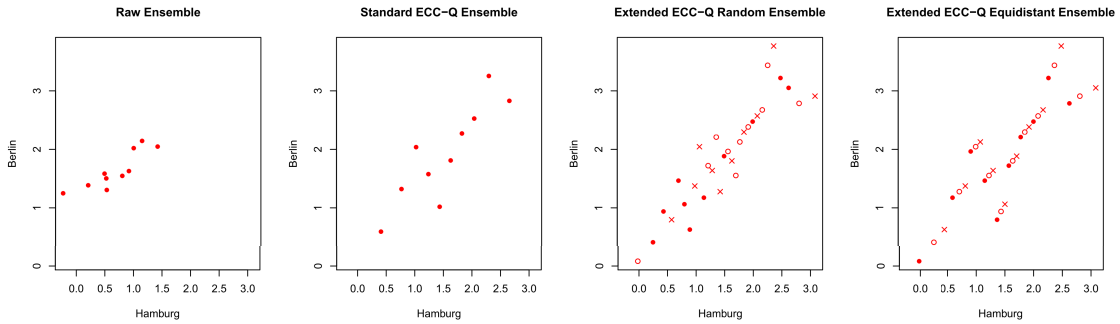


Figure 4.32: Scenario (A): Scatterplots for 24 hour ahead temperature forecasts (in °C) at Hamburg and Berlin, valid 1:00 am on 24 November 2010, of a 10-member raw ensemble, the corresponding 10-member standard ECC-Q ensemble and the 30-member extended ECC-Q random and extended ECC-Q equidistant ensembles, respectively, consisting of 3 subsamples of 10 members each, indicated by the red dots, circles and crosses, respectively.

- the extended ECC-Q random ensemble consisting of $N = 30$ members, in which the $\frac{1}{31}$-, ..., $\frac{30}{31}$-quantiles of the corresponding predictive EMOS CDF are randomly divided into $r = 3$ subsamples of size $M = 10$, indicated by red dots, circles and crosses, respectively, with each conserving the rank dependence structure of the raw ensemble, and

- the extended ECC-Q equidistant ensemble consisting of $N = 30$ members, in which the $\frac{1}{31}$-, ..., $\frac{30}{31}$-quantiles of the corresponding predictive CDF are divided into $r = 3$ subsamples of size $M = 10$, indicated by red dots, circles and crosses, respectively, with each conserving the rank dependence structure of the raw ensemble, being composed of the following members:

  - The first subsample comprises the $\frac{n}{31}$-quantiles form the corresponding predictive CDF, where $n \in \{1, 4, 7, 10, 13, 16, 19, 22, 25, 28\}$.
  - The second subsample consists of the $\frac{n}{31}$-quantiles, where $n \in \{2, 5, 8, 11, 14, 17, 20, 23, 26, 29\}$.
  - The third subsample is built-up by the $\frac{n}{31}$-quantiles, where $n \in \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$.

We recall that in both extended ECC-Q approaches, each subsample has the same rank dependence structure as the raw ensemble and the single standard ECC-Q ensemble.

Scenario (B)

We now focus on the case, in which the requested extended ensemble size after postprocessing is not an integer multiple of the raw ensemble size, that is, $N = r \cdot M + P$, where $r \in \mathbb{N}$ and $P \in \mathbb{N}$, $1 \leq P < M$. As before, we are given the raw ensemble forecast $x_1^\ell, \ldots, x_M^\ell$ for each weather variable $i \in \{1, ..., I\}$, each location $j \in \{1, ..., J\}$ and each prediction horizon $k \in \{1, ..., K\}$ individually, summarized in the multi-index $\ell := (i, j, k)$. This defines a permutation $\sigma_\ell(m) := \text{rank}(x_m^\ell)$ for $m \in \{1, \ldots, M\}$, with ties resolved at random. For each ℓ, we employ state-of-the-art univariate ensemble postprocessing methods, such as BMA or EMOS, to obtain a predictive CDF F_ℓ. For each ℓ, we draw a sample $\tilde{x}_1^\ell, \ldots, \tilde{x}_N^\ell$

of size N from F_ℓ by taking the equally spaced $\frac{1}{N+1}$-, ..., $\frac{N}{N+1}$- quantiles of F_ℓ, that is,

$$\tilde{x}_1^\ell := F_\ell^{-1}\left(\frac{1}{N+1}\right), \ldots, \tilde{x}_N^\ell := F_\ell^{-1}\left(\frac{N}{N+1}\right).$$

Contrary to Scenario (A), we now divide our N-sized sample into $r+1$ subsamples, where r of them consist of M members and one of them comprises P members. Since a partition of the $\frac{1}{N+1}$-, ..., $\frac{N}{N+1}$-quantiles into $r+1$ subsamples, such that each subsample has equidistant quantiles as in the extended ECC-Q equidistant method of Scenario (A), is obviously not possible in Scenario (B) here, we basically proceed as in the extended ECC-Q random method of Scenario (A) by randomly dividing the sample

$$\tilde{\boldsymbol{x}}^\ell := (\tilde{x}_1^\ell, \ldots, \tilde{x}_N^\ell)$$

into $r+1$ subsamples

$$\tilde{\boldsymbol{x}}^{\ell,[1]} := (\tilde{x}_1^{\ell,[1]}, \ldots, \tilde{x}_M^{\ell,[1]}), \ldots, \tilde{\boldsymbol{x}}^{\ell,[r]} := (\tilde{x}_1^{\ell,[r]}, \ldots, \tilde{x}_M^{\ell,[r]}), \tilde{\boldsymbol{x}}^{\ell,[r+1]} := (\tilde{x}_1^{\ell,[r+1]}, \ldots, \tilde{x}_P^{\ell,[r+1]}).$$

That is, if λ_ℓ is a random permutation of $\{1, \ldots, N\}$, the $r+1$ subsamples are given by

$$\begin{aligned}
\tilde{\boldsymbol{x}}^{\ell,[1]} &:= (\tilde{x}_1^{\ell,[1]}, \ldots, \tilde{x}_M^{\ell,[1]}) := (\tilde{x}_{\lambda_\ell(1)}^\ell, \ldots, \tilde{x}_{\lambda_\ell(M)}^\ell),\\
\tilde{\boldsymbol{x}}^{\ell,[2]} &:= (\tilde{x}_1^{\ell,[2]}, \ldots, \tilde{x}_M^{\ell,[2]}) := (\tilde{x}_{\lambda_\ell(M+1)}^\ell, \ldots, \tilde{x}_{\lambda_\ell(2M)}^\ell),\\
&\vdots\\
\tilde{\boldsymbol{x}}^{\ell,[r]} &:= (\tilde{x}_1^{\ell,[r]}, \ldots, \tilde{x}_M^{\ell,[r]}) := (\tilde{x}_{\lambda_\ell((r-1)M+1)}^\ell, \ldots, \tilde{x}_{\lambda_\ell(rM)}^\ell),\\
\tilde{\boldsymbol{x}}^{\ell,[r+1]} &:= (\tilde{x}_1^{\ell,[r+1]}, \ldots, \tilde{x}_P^{\ell,[r+1]}) := (\tilde{x}_{\lambda_\ell(rM+1)}^\ell, \ldots, \tilde{x}_{\lambda_\ell(rM+P)}^\ell),
\end{aligned}$$

with $N = r \cdot M + P$.
With regard to the r subsamples $\tilde{\boldsymbol{x}}^{\ell,[1]}, \ldots, \tilde{\boldsymbol{x}}^{\ell,[r]}$ consisting of M members each, we exactly proceed as in Scenario (A) and obtain r different standard ECC-Q ensembles $\hat{\boldsymbol{x}}^{\ell,[1]}, \hat{\boldsymbol{x}}^{\ell,[2]}, \ldots,$ $\hat{\boldsymbol{x}}^{\ell,[r]}$ via

$$\begin{aligned}
\hat{\boldsymbol{x}}^{\ell,[1]} &:= (\hat{x}_1^{\ell,[1]}, \ldots, \hat{x}_M^{\ell,[1]}) := (\tilde{x}_{(\sigma_\ell(1))}^{\ell,[1]}, \tilde{x}_{(\sigma_\ell(2))}^{\ell,[1]}, \ldots, \tilde{x}_{(\sigma_\ell(M))}^{\ell,[1]}),\\
\hat{\boldsymbol{x}}^{\ell,[2]} &:= (\hat{x}_1^{\ell,[2]}, \ldots, \hat{x}_M^{\ell,[2]}) := (\tilde{x}_{(\sigma_\ell(1))}^{\ell,[2]}, \tilde{x}_{(\sigma_\ell(2))}^{\ell,[2]}, \ldots, \tilde{x}_{(\sigma_\ell(M))}^{\ell,[2]}),\\
&\vdots\\
\hat{\boldsymbol{x}}^{\ell,[r]} &:= (\hat{x}_1^{\ell,[r]}, \ldots, \hat{x}_M^{\ell,[r]}) := (\tilde{x}_{(\sigma_\ell(1))}^{\ell,[r]}, \tilde{x}_{(\sigma_\ell(2))}^{\ell,[r]}, \ldots, \tilde{x}_{(\sigma_\ell(M))}^{\ell,[r]}).
\end{aligned}$$

For the single subsample $\hat{\boldsymbol{x}}^{\ell,[r+1]}$ comprising P members, we consider the first $P < M$ raw ensemble members and their corresponding forecast $x_1^\ell, \ldots, x_P^\ell$ defining a permutation ρ_ℓ of $\{1, \ldots, P\}$ via $\rho_\ell(\mu) := \text{rank}(x_\mu^\ell)$ for $\mu \in \{1, \ldots, P\}$, with ties resolved at random. The subsample $\hat{\boldsymbol{x}}^{\ell,[r+1]}$ comprising P members is then reordered according to the permutation ρ_ℓ to get the ensemble $\hat{\boldsymbol{x}}^{\ell,[r+1]}$ via

$$\hat{\boldsymbol{x}}^{\ell,[r+1]} := (\hat{x}_1^{\ell,[r+1]}, \ldots, \hat{x}_P^{\ell,[r+1]}) := (\tilde{x}_{(\rho_\ell(1))}^{\ell,[r+1]}, \tilde{x}_{(\rho_\ell(2))}^{\ell,[r+1]}, \ldots, \tilde{x}_{(\rho_\ell(P))}^{\ell,[r+1]}).$$

Hence, we try to kind of project the basic dependence structure within the M members onto the remaining $P < M$ members, such that it is roughly retained in the smaller sample. In principle, the P out of the M raw ensemble member indices may also be chosen according

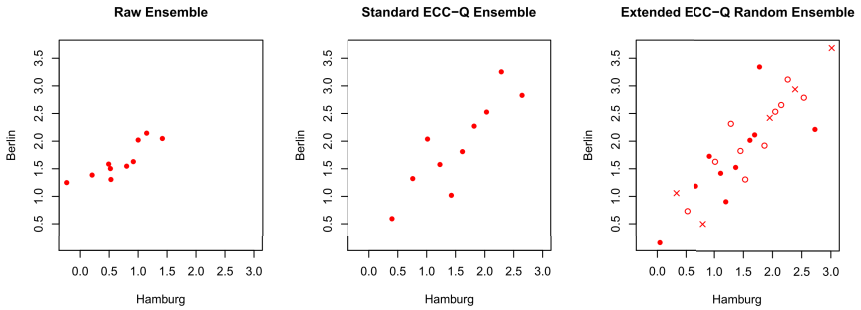


Figure 4.33: Scenario (B): Scatterplots for 24 hour ahead temperature forecasts (in °C) at Hamburg and Berlin, valid 1:00 am on 24 November 2010, of a 10-member raw ensemble, the corresponding 10-member standard ECC-Q ensemble and the 25-member extended ECC-Q random ensemble, which consists of 2 subsamples of 10 members each, indicated by the red dots and circles, respectively, and a single subsample of 5 members, indicated by the red crosses.

to other criteria, rather than just taking the first P as performed here. Finally, the $r+1$ ensembles $\hat{\boldsymbol{x}}^{\ell,[1]}, \ldots, \hat{\boldsymbol{x}}^{\ell,[r]}, \hat{\boldsymbol{x}}^{\ell,[r+1]}$ are aggregated in an extended ECC-Q random ensemble $\hat{\boldsymbol{x}}^{\ell}$ comprising $N = r \cdot M + P$ members via

$$\hat{\boldsymbol{x}}^{\ell} := (\hat{x}^{\ell}_1, \ldots, \hat{x}^{\ell}_N) := (\hat{\boldsymbol{x}}^{\ell,[1]}, \ldots, \hat{\boldsymbol{x}}^{\ell,[r]}, \hat{\boldsymbol{x}}^{\ell,[r+1]}).$$

Again, the order of the $r+1$ ensembles in the aggregation can be chosen arbitrarily, but has to be consistent for each ℓ, where we arguably use the most natural aggregation order above.

To illustrate the approach for Scenario (B) in Figure 4.33, we again consider the setting of the previous example. This time, the original $M = 10$-member ensemble shall be extended to a size of $N = 25 = 2 \cdot 10 + 5$ members. The 10-member raw and standard ECC-Q ensembles are shown in the left and mid-panel in Figure 4.33, respectively. Besides, the extended ECC-Q random ensemble consisting of $N = 25$ members is exhibited in the right panel, where the $r = 2$ subsamples of size $M = 10$ are indicated by red dots and circles, respectively, and the subsample of size $P = 5$ by red crosses. In the extended ECC-Q random ensemble, the 2 subsamples consisting of 10 members each exhibit the same rank dependence structure as both the raw and the standard ECC-Q ensemble, and the single 5-member subsample covers a part of that at the very least.

4.4.2 Case study

We now assess the extended ECC-Q ensembles as proposed before and compare their predictive performances to those of both the raw and the standard ECC-Q ensemble. For this purpose, we consider ECMWF 24 hour ahead temperature forecasts at Berlin and Hamburg jointly, for the two-year test period from 1 January 2011 to 31 December 2012. Univariate postprocessing is done via EMOS throughout, using a training period of 30 days. In what follows, we analyze the three cases, in which we start with raw ensembles consisting of $M = 10$, 30 and 50 members, by just taking the first 10, first 30 and all members, respectively, of the regular ECMWF ensemble. Each initial M-member raw ensemble is then extended to several test sizes N by employing the corresponding techniques for Scenario (A) or Scenario (B).

Figures 4.34 to 4.36 show the corresponding average energy score (ES), which is employed as

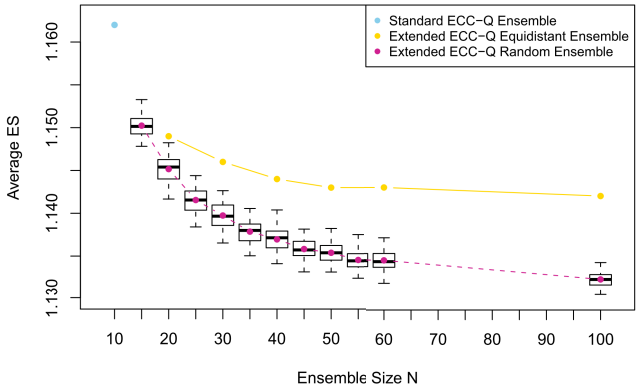


Figure 4.34: Average ES (in °C) for 24 hour ahead temperature forecasts at Berlin and Hamburg jointly over the test period from 1 January 2011 to 31 December 2012. Based on an $M = 10$-member raw ensemble, the scores for the corresponding standard ECC-Q, the extended ECC-Q random and the extended ECC-Q equidistant ensemble, respectively, are shown as a function of the desired extended ensemble size $N \geq M$. The results for the extended ECC-Q random ensemble are averaged over 100 runs and complemented by the corresponding standard boxplots. For comparison: The raw ensemble has an average energy score of 1.756.

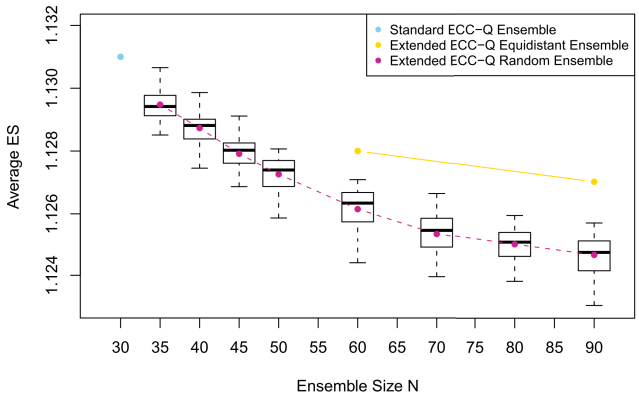


Figure 4.35: Average ES (in °C) for 24 hour ahead temperature forecasts at Berlin and Hamburg jointly over the test period from 1 January 2011 to 31 December 2012. Based on an $M = 30$-member raw ensemble, the scores for the corresponding standard ECC-Q, the extended ECC-Q random and the extended ECC-Q equidistant ensemble, respectively, are shown as a function of the desired extended ensemble size $N \geq M$. The results for the extended ECC-Q random ensemble are averaged over 100 runs and complemented by the corresponding standard boxplots. For comparison: The raw ensemble has an average energy score of 1.736.

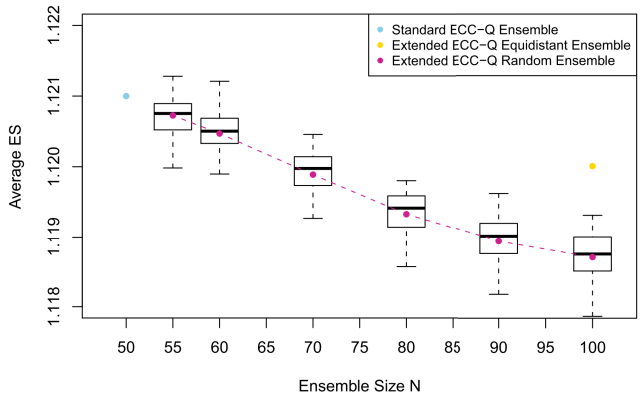


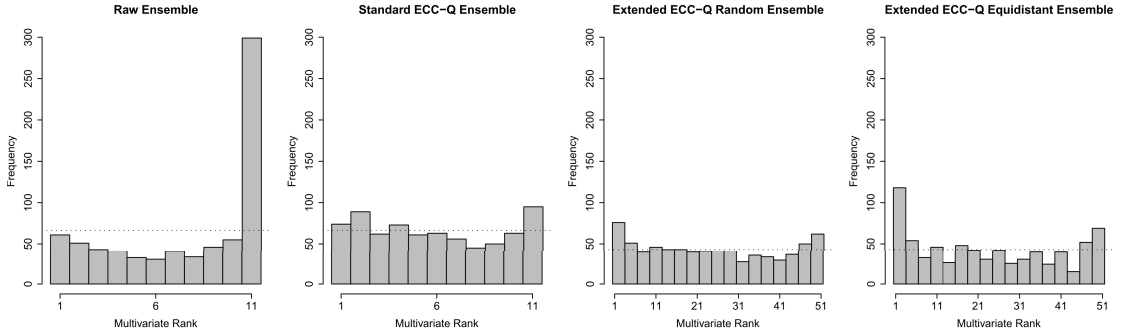
Figure 4.36: Average ES (in °C) for 24 hour ahead temperature forecasts at Berlin and Hamburg jointly over the test period from 1 January 2011 to 31 December 2012. Based on an $M = 50$-member raw ensemble, the scores for the corresponding standard ECC-Q, the extended ECC-Q random and the extended ECC-Q equidistant ensemble, respectively, are shown as a function of the desired extended ensemble size $N \geq M$. The results for the extended ECC-Q random ensemble are averaged over 100 runs and complemented by the corresponding standard boxplots. For comparison: The raw ensemble has an average energy score of 1.726.

an overall performance measure here, as a function of the desired ensemble size N after postprocessing over our test period. For scaling reasons, the scores for the raw ensembles are not explicitly shown in the plots, but mentioned in the corresponding captions for comparison. In the case of Scenario (A), the results both for the extended ECC-Q random and extended ECC-Q equidistant ensemble, respectively, are given. The values obtained when using the random partitioning are averaged over 100 runs and complemented by the respective standard boxplots, both for Scenario (A) and Scenario (B). In this context, the term "standard boxplot" refers to the classical box-and-whisker plot as introduced by Tukey (1977), where outliers are not explicitly shown here. To check calibration only, the multivariate, band depth and average rank histograms, respectively, for our test period are shown in Figure 4.37 for the exemplary situation of a raw ensemble of size $M = 10$, which is extended to a $N = 50$-member ensemble according to Scenario (A). Similarly, histograms for the situation in which a raw ensemble is extended from $M = 10$ to $N = 35$ members according to Scenario (B) can be found in Figure 4.38.

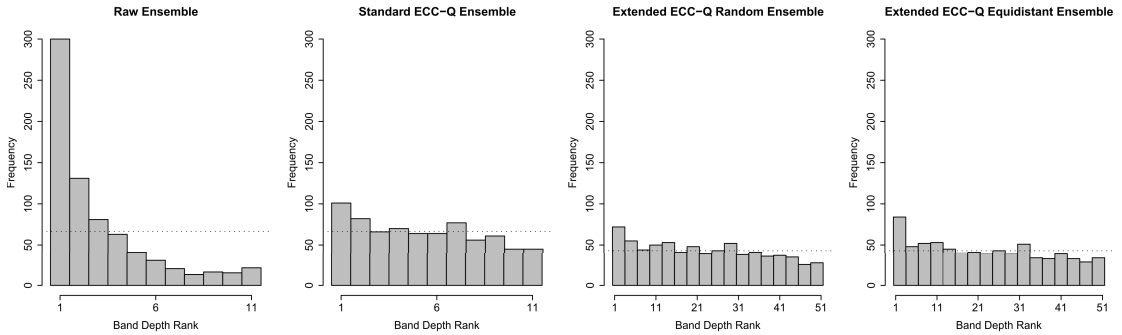
In all the cases, both the standard ECC-Q and the extended ECC-Q ensembles clearly outperform the raw ensemble. Hence, the basic goal of statistical ensemble postprocessing is achieved.

The extended ECC-Q ensembles essentially always outperform the standard ECC-Q ensemble in terms of the ES. If the initial raw ensemble is already sufficiently large, here for $M = 30$ or $M = 50$, the improvements are rather small, whereas if the raw ensemble is comparably small, here for $M = 10$, the improvements are considerable. Hence, if an ensemble is sufficiently large with already sufficient amount of information, extended ECC-Q methods become rather unimportant, such that it effectively makes no difference whether we employ standard ECC-Q or extended ECC-Q. However, they are very valuable for small ensembles in order to provide more information.

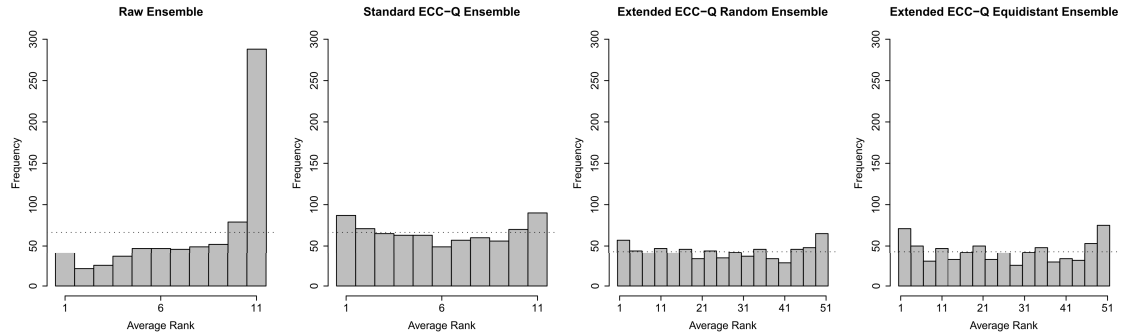
As far as extended ECC-Q for Scenario (A) is concerned, extended ECC-Q random per-



(a) Multivariate rank histograms

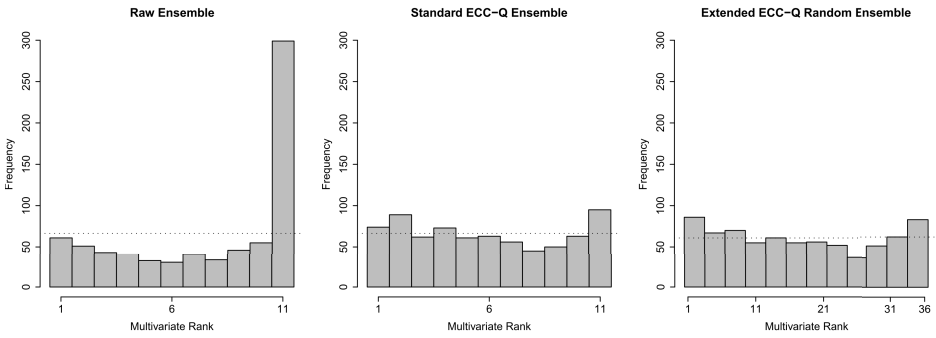


(b) Band depth rank histograms

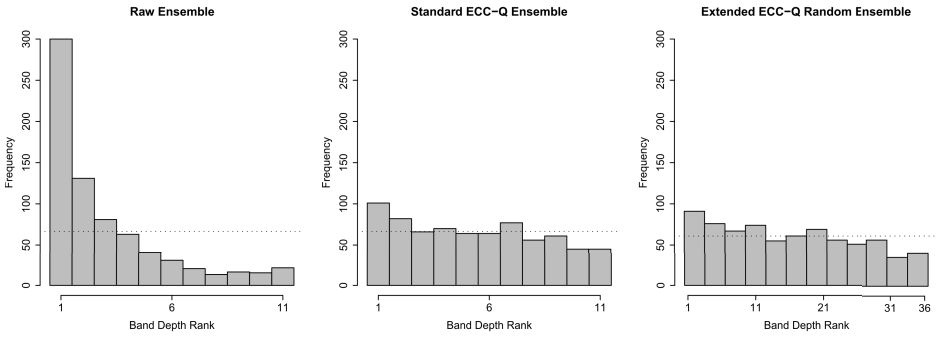


(c) Average rank histograms

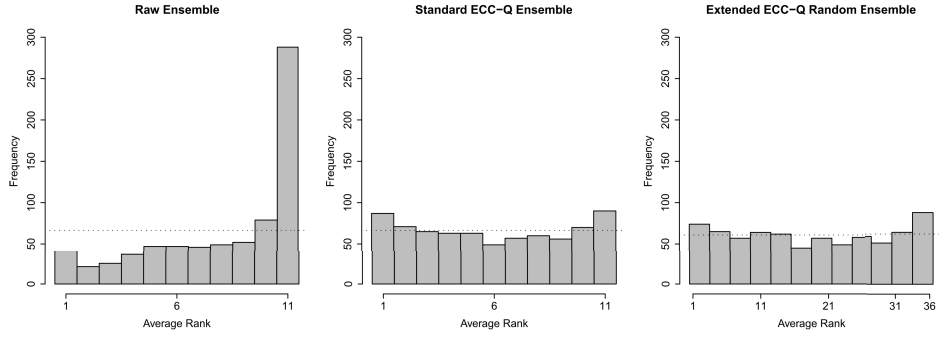
Figure 4.37: Scenario (A): (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature forecasts at Berlin and Hamburg jointly over the test period from 1 January 2011 to 31 December 2012. Based on an $M = 10$-member raw ensemble, the extended ECC-Q random and extended ECC-Q equidistant ensemble, respectively, comprises $N = 50$ members.



(a) Multivariate rank histograms



(b) Band depth rank histograms



(c) Average rank histograms

Figure 4.38: Scenario (B): (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature forecasts at Berlin and Hamburg jointly over the test period from 1 January 2011 to 31 December 2012. Based on an $M = 10$-member raw ensemble, the extended ECC-Q random ensemble comprises $N = 35$ members.

forms slightly better than extended ECC-Q equidistant throughout. Contrary to the benefits at the sampling stage, the use of equidistance has no positive effect with regard to the partitioning of the samples. In particular, Figure 4.37 reveals that extended ECC-Q random yields a somewhat better calibration than extended ECC-Q equidistant, for which the multivariate rank histogram in Figure 4.37 (a) is U-shaped and thus indicates underdispersion, for example. The reason for that is to a certain extent already hinted at and illustrated in Figure 4.32. For extended ECC-Q equidistant, the three subsamples look very much the same and are only sort of shifted, which is obvious by construction, whereas the subsample pattern of extended ECC-Q random is somehow more scattered. In some cases, the equidistant partitioning might create a sort of an unnatural overstructuring.

As N increases, the energy scores $\mathrm{ES}^{N\,\mathrm{mem}}_{\text{Ext ECC-Q}}$ of the extended ECC-Q ensembles get lower, that is, better. However, the improvements $\mathrm{ES}^{N^*\,\mathrm{mem}}_{\text{Ext ECC-Q}} - \mathrm{ES}^{N\,\mathrm{mem}}_{\text{Ext ECC-Q}}$ of the ES for an ensemble size N and a larger size $N^* > N$ are noteworthy for comparably small N, but become very minor if N gets sufficiently large.

In a nutshell, it is worth to employ extended ECC-Q approaches in cases when the original M-member raw ensemble is rather small. If doing so, it is recommended to use extended ECC-Q random, and it generally suffices to extend the initial ensemble to a size N which is not excessively larger than M. Practically, it may often be reasonable and plausible to restrict oneself to the somewhat more natural setting of Scenario (A), in which $N = r \cdot M$, and take rather low values for r, depending on the initial size M.

4.4.3 Alternative approaches

Finally, we briefly discuss two alternative approaches to the extended ECC method described in Section 4.4.1 and tested in Section 4.4.2. Both techniques also yield a postprocessed N-member ensemble greater in size than the M-member raw ensemble, that is, $N > M$. The first is based on a recycling notion, and the second uses so-called lagged ensembles.

(a) A recycling approach

First, we describe a simple and canonical recycling procedure that can be applied in the case of Scenario (A) of extended ECC only, that is, if $N = r \cdot M$, where $r \geq 2$. Assuming that we are given a postprocessed M-member standard ECC ensemble, a postprocessed, enlarged N-member ensemble can be obtained by just recycling or repeating r times the standard ECC ensemble in a redundant implementation. We refer to this method as recycled ECC in what follows. As in the case of standard ECC, we distinguish the variants (R), (T) and (Q), depending on the quantization used at the sampling stage. The recycled ECC-T or -Q ensemble comprises r times the same standard ECC-T or -Q ensemble, respectively. In contrast, the recycled ECC-R ensemble, which has also been proposed and employed by Wilks (2014), generally does not consist of r times the same sample, because repeated implementation is not necessarily redundant in the setting of scheme (R).

(b) An extended ECC variant using lagged ensembles

Our last method to create an N-member postprocessed ensemble exceeding the raw ensemble size M and being able to account for dependencies in an ECC-like manner is based on so-called lagged or time-lagged ensembles.

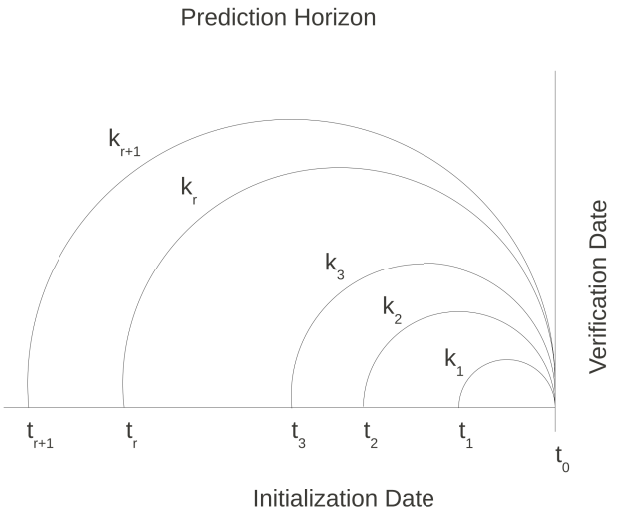


Figure 4.39: Illustration of the connection between initialization dates, verification date and prediction horizons to construct an enlarged lagged raw ensemble. A similar scheme is shown in Figure 3 in Lu et al. (2007).

Lagged ensemble forecasting has originally been proposed by Hoffman and Kalnay (1983). A forecast ensemble is thereby constructed by collecting single-valued point forecasts launched at different initialization times, but all materializing at the same verification time, thus comprising multiple prediction horizons. What Hoffman and Kalnay (1983) call a lagged average forecast is then just the average of the corresponding lagged ensemble forecasts at their proper verification time. Mostly, each of the lagged ensemble members is given equal weight (Kalnay, 2003), while more involved weighting schemes are also available (Lu et al., 2007). Meanwhile, lagged ensembles have been investigated and employed in various papers and case studies, including those of Lawrence and Hansen (2007), Lu et al. (2007), Mittermaier (2007), Bentzien and Friederichs (2012) and Scheufele et al. (2014). Lawrence and Hansen (2007) use lagged ensembles to increase ensemble size, as is also desired in our setting here. Specifically, they combine ensemble forecasts initialized at different times, but valid at the same time, rather than point forecasts. This corresponds better to our setup, and we now employ the aforementioned notions in our context. As this is meant to be a description of ideas, both the elaboration of details or modifications and the implementation of the method in a real-data case study to assess the predictive performance are issues for future work.

For weather quantity $i \in \{1, \ldots, I\}$, location $j \in \{1, \ldots, J\}$ and look-ahead time $k_1 \in \{1, \ldots, K\}$, summarized in the multi-index $\ell_1 := (i, j, k_1)$, we are given the M-member raw ensemble forecast $x_1^{\ell_1}, \ldots, x_M^{\ell_1}$ valid at verification date t_0 and initialized at date t_1. To create an enlarged ensemble of size $N = r \cdot M + P$, with $r \in \mathbb{N}$ and $P \in \mathbb{N}_0$, $0 \leq P < M$, we pad the above M-member raw ensemble with $N - M$ ensemble forecasts also valid at t_0, but launched at different initialization dates $t_2, \ldots, t_{r+1} \neq t_1$, thus corresponding to different look-ahead times $k_2, \ldots, k_{r+1} \neq k_1$ and multi-indices $\ell_2, \ldots, \ell_{r+1} \neq \ell_1$, respectively. The initialization dates $t_2, \ldots, t_{r+1}$ and the corresponding prediction horizons $k_2, \ldots, k_{r+1}$ can be selected arbitrarily. However, the dates $t_2, \ldots, t_{r+1}$ have to be in the past of t_1, and hence, the look-ahead times $k_2, \ldots, k_{r+1}$ have to be greater than k_1. Without loss of generality, we can assume that the prediction horizons $k_2, \ldots, k_{r+1}$ are increasing with the

indices $2, \ldots, r+1$ here. An illustration of the connections between initialization dates, verification date and prediction horizons to construct an enlarged lagged raw ensemble is shown in Figure 4.39, which is similar to Figure 3 in Lu et al. (2007).

As the lagged ensemble forecast is based on several initialization dates, temporal correlations among different prediction horizons can hardly be accounted for as before. Thus, the focus is on retaining dependence structures among weather variables $i \in \{1, \ldots, I\}$ and locations $j \in \{1, \ldots, J\}$, summarized in the multi-index $\ell^* := (i, j)$.

In the general case of $N = r \cdot M + P$, a lagged ensemble forecast $\boldsymbol{x}^{\ell^*}$ for each ℓ^*, based on the different raw ensemble forecasts

$$\begin{aligned}
\boldsymbol{x}^{\ell^*,[1]} &:= (x_1^{\ell^*}, \ldots, x_M^{\ell^*}) := (x_1^{\ell_1}, \ldots, x_M^{\ell_1}), \\
\boldsymbol{x}^{\ell^*,[2]} &:= (x_{M+1}^{\ell^*}, \ldots, x_{2M}^{\ell^*}) := (x_1^{\ell_2}, \ldots, x_M^{\ell_2}), \\
&\vdots \\
\boldsymbol{x}^{\ell^*,[r]} &:= (x_{(r-1)M+1}^{\ell^*}, \ldots, x_{rM}^{\ell^*}) := (x_1^{\ell_r}, \ldots, x_M^{\ell_r}), \\
\boldsymbol{x}^{\ell^*,[r+1]} &:= (x_{rM+1}^{\ell^*}, \ldots, x_{rM+P}^{\ell^*}) := (x_1^{\ell_{r+1}}, \ldots, x_P^{\ell_{r+1}}),
\end{aligned}$$

which are all valid at date t_0, can be formally obtained by

$$\begin{aligned}
\boldsymbol{x}^{\ell^*} &:= (\boldsymbol{x}^{\ell^*,[1]}, \boldsymbol{x}^{\ell^*,[2]}, \ldots, \boldsymbol{x}^{\ell^*,[r]}, \boldsymbol{x}^{\ell^*,[r+1]}) \\
&= (x_1^{\ell^*}, \ldots, x_M^{\ell^*}, x_{M+1}^{\ell^*}, \ldots, x_{2M}^{\ell^*}, \ldots, x_{(r-1)M+1}^{\ell^*}, \ldots, x_{rM}^{\ell^*}, x_{rM+1}^{\ell^*}, \ldots, x_{rM+P}^{\ell^*}) \\
&= (x_1^{\ell_1}, \ldots, x_M^{\ell_1}, x_1^{\ell_2}, \ldots, x_M^{\ell_2}, \ldots, x_1^{\ell_r}, \ldots, x_M^{\ell_r}, x_1^{\ell_{r+1}}, \ldots, x_P^{\ell_{r+1}}),
\end{aligned}$$

with $N = r \cdot M + P$. The P out of M members the ensemble $\boldsymbol{x}^{\ell^*,[r+1]}$ consists of can be chosen either randomly or according to a specific criterion, but should be fixed throughout. One could also just take the first P out of the M raw ensemble members. The order of the $r+1$ ensembles which are employed to create the extended lagged raw ensemble $\boldsymbol{x}^{\ell^*}$ can be chosen arbitrarily, while it has to remain fixed for all ℓ^* throughout.

In the special case of $P = 0$, that is, for $N = r \cdot M$, the lagged ensemble $\boldsymbol{x}^{\ell^*}$ reduces to

$$\begin{aligned}
\boldsymbol{x}^{\ell^*} &:= (\boldsymbol{x}^{\ell^*,[1]}, \boldsymbol{x}^{\ell^*,[2]}, \ldots, \boldsymbol{x}^{\ell^*,[r]}) \\
&= (x_1^{\ell^*}, \ldots, x_M^{\ell^*}, x_{M+1}^{\ell^*}, \ldots, x_{2M}^{\ell^*}, \ldots, x_{(r-1)M+1}^{\ell^*}, \ldots, x_{rM}^{\ell^*}) \\
&= (x_1^{\ell_1}, \ldots, x_M^{\ell_1}, x_1^{\ell_2}, \ldots, x_M^{\ell_2}, \ldots, x_1^{\ell_r}, \ldots, x_M^{\ell_r}).
\end{aligned}$$

The N members of the extended lagged raw ensemble $\boldsymbol{x}^{\ell^*}$ are treated as if they were exchangeable in what follows. For each ℓ^*, a permutation $\rho_{\ell^*}(n) := \text{rank}(x_n^{\ell^*})$ for $n \in \{1, \ldots, N\}$ is induced by $\boldsymbol{x}^{\ell^*}$, with ties resolved at random. In the next step, univariate ensemble postprocessing methods, such as BMA or EMOS, are employed to obatin a predictive CDF F_{ℓ^*} for each ℓ^*. Then, we draw a sample $\tilde{x}_1^{\ell^*}, \ldots, \tilde{x}_N^{\ell^*}$ of size N from each CDF F_{ℓ^*} according to quantization scheme (R) or (Q), respectively. We prefer as before the latter sampling method (Q) here and thus obtain

$$\tilde{x}_1^{\ell^*} := F_{\ell^*}^{-1}\left(\frac{1}{N+1}\right), \ldots, \tilde{x}_N^{\ell^*} := F_{\ell^*}^{-1}\left(\frac{N}{N+1}\right).$$

The samples $\tilde{x}_1^{\ell^*}, \ldots, \tilde{x}_N^{\ell^*}$ are finally reordered with respect to the permutation ρ_{ℓ^*} to conserve the spatial and inter-variable rank dependence pattern from the extended lagged raw ensemble, yielding the final postprocessed ensemble $\hat{x}_1^{\ell^*}, \ldots, \hat{x}_N^{\ell^*}$ via

$$\hat{x}_1^{\ell^*} := \tilde{x}_{(\rho_{\ell^*}(1))}^{\ell^*}, \ldots, \hat{x}_N^{\ell^*} := \tilde{x}_{(\rho_{\ell^*}(N))}^{\ell^*},$$

as in the standard ECC-Q approach.

As mentioned before, the lagged ensemble postprocessing method described above just collects ideas, and several details of the approach still have to be worked out in the future. For instance, the specific choice of the prediction horizons $k_2, \ldots, k_{r+1}$ to pad the original raw ensemble associated with look-ahead time k_1 has to be examined, in that whether there are prediction horizons that are more suitable than others for the padding process. In this context, also the development of appropriate weighting schemes for the different look-ahead times is of interest and importance. Moreover, the assumption of exchangeability for the members of the extended lagged raw ensemble might be doubtful and not fulfilled each and every time. In this light, the lagged ensemble approach might be combined with the ECC method for ensembles consisting of non-exchangeable members, which will be presented in the next Section 4.5, to provide an appropriate procedure.

With the two alternative approaches described before and the extended ECC-Q methods presented in Section 4.4.1, we can solve the shortcoming that the standard ECC ensemble always has to consist of the same number of members as the raw ensemble. However, these extension techniques still apply to raw ensembles with exchangeable members only – a problem which will be addressed with respect to standard ECC in the next Section 4.5. An extension approach that is also valid for raw ensembles comprising non-exchangeable members can be constructed by combining similarity-based ensemble methods with the Schaake shuffle presented in Section 3.2.2. This technique is described and evaluated in Chapter 8 in this thesis.

4.5 ECC for ensembles with non-exchangeable members

As noted, a further constraint of the standard ECC implementation from Section 4.1 is that it can only be applied to raw ensembles consisting of exchangeable members. We now consider situations in which at least a part of the ensemble members are statistically distinguishable, coming from different sources. For instance, we can think of a multi-model ensemble consisting of C different physical models of the atmosphere from different NWP centers, with exchangeable ensemble members arising from N_c perturbations of the initial conditions for each single model $c \in \{1, \ldots, C\}$, which are then aggregated to a final overall ensemble. This is for example realized in the TIGGE project and database (Bougeault et al., 2010). The COSMO-DE ensemble (Gebhardt et al., 2011) run by the German Weather Service is an ensemble consisting of clusters of exchangeable members.

To describe an ECC variant suitable to such settings, let $\ell := (i, j, k)$ be a multi-index referring to weather variable $i \in \{1, \ldots, I\}$, location $j \in \{1, \ldots, J\}$ and look-ahead time $k \in \{1, \ldots, K\}$. Moreover, let

$$\boldsymbol{x}^\ell := (x_1^\ell, \ldots, x_M^\ell) := \Big(\underbrace{\underbrace{x_{1,1}^\ell, \ldots, x_{1,N_1}^\ell}_{N_1\text{-member cluster 1}}, \underbrace{x_{2,1}^\ell, \ldots, x_{2,N_2}^\ell}_{N_2\text{-member cluster 2}}, \ldots, \underbrace{x_{C,1}^\ell, \ldots, x_{C,N_C}^\ell}_{N_C\text{-member cluster } C}}_{M \text{ members}} \Big)$$

be a raw ensemble forecast consisting of M members in total, but split in C clusters, or groups, comprising N_c exchangeable members for each $c \in \{1, \ldots, C\}$, respectively, such that $M = \sum_{c=1}^{C} N_c$. For each fixed margin ℓ, we proceed as follows.

1. For each cluster $c \in \{1, \ldots, C\}$, apply standard ECC based on the N_c-member raw cluster ensemble as a reference by conducting the following steps.

   - Derive the permutation

     $$\rho_\ell^c(n) := \text{rank}(x_{c,n}^\ell) \text{ for } n \in \{1, \ldots, N_c\},$$

     with ties resolved at random, induced by the corresponding raw ensemble forecast $x_{c,1}^\ell, \ldots, x_{c,N_c}^\ell$ of cluster $c \in \{1, \ldots, C\}$.
   - Perform univariate postprocessing, for example via BMA or EMOS, and obtain a calibrated and sharp predictive CDF F_ℓ^c. That is, biases and dispersion errors are corrected separately for each raw cluster ensemble.
   - Generate a sample $\tilde{x}_{c,1}^\ell, \ldots, \tilde{x}_{c,N_c}^\ell$ of size N_c from F_ℓ^c according to the sampling scheme (R), (T) or (Q), respectively.
   - Obtain the postprocessed cluster ensemble $\hat{x}_{c,1}^\ell, \ldots, \hat{x}_{c,N_c}^\ell$ by reordering the samples on the basis of ρ_ℓ^c:

     $$\hat{x}_{c,1}^\ell := \tilde{x}_{(\rho_\ell^c(1))}^\ell, \ldots, \hat{x}_{c,N_c}^\ell := \tilde{x}_{(\rho_\ell^c(N_c))}^\ell.$$

2. Aggregate the C postprocessed cluster ensembles piece by piece in the same order as in the overall raw ensemble to obtain

   $$\hat{\boldsymbol{x}}^\ell := (\hat{x}_1^\ell, \ldots, \hat{x}_M^\ell) := (\hat{x}_{1,1}^\ell, \ldots, \hat{x}_{1,N_1}^\ell, \hat{x}_{2,1}^\ell, \ldots, \hat{x}_{2,N_2}^\ell, \ldots, \hat{x}_{C,1}^\ell, \ldots, \hat{x}_{C,N_C}^\ell)$$

   as the final postprocessed M-member overall ensemble.

As hinted at in the previous section, the approach to deal with ensembles comprising non-exchangeable members described above could be combined with the lagged ensemble method from Section 4.4.3, in that for each look-ahead time $k_1, k_2, \ldots, k_{r+1}$, the respective ensemble forecasts associated therewith, where the corresponding ensemble members are assumed to be exchangeable, can be supposed to build the $C = r + 1$ clusters $c_1, c_2, \ldots, c_{r+1}$.

Chapter 5

ECC as an overarching theme

We now focus on exemplary techniques that have been already presented in the literature and show similarities to the ECC approach or contain elements thereof. In Section 5.1, we review a common and widely used member-by-member postprocessing (MBMP) method and discuss its relationship to ECC. Moreover, the predictive performance of this MBMP approach is compared in a case study to that of ECC based on the standard implementation in this thesis. Thereafter, we note further methods in the extant literaure that can also be interpreted as ECC variants in Section 5.2. All in all, ECC turns out to form an overarching theme for seemingly unrelated approaches scattered in the literature.

5.1 A member-by-member postprocessing (MBMP) method as an ECC variant

We have seen that state-of-the-art ensemble postprocessing approaches, such as Bayesian model averaging (BMA) or ensemble model output statistics (EMOS), are mostly univariate and apply to a single weather quantity at a single location and for a single prediction horizon only, thereby failing to account for potentially crucial dependence structures. To address this shortcoming, BMA and EMOS can be combined with the ECC approach, in which the postprocessed forecast ensemble inherits the spatial, temporal and inter-variable dependence structures of the unprocessed raw ensemble, as proposed in Section 4.1. In Gaussian settings, the rank dependence patterns from the raw ensemble can alternatively be retained by using member-by-member postprocessing (MBMP) methods, which are wide-spread in the literature (Doblas-Reyes et al., 2005; Johnson and Bowler, 2009; Van Schaeybroeck and Vannitsem, 2014). In this section, we review a consolidated version of the MBMP approaches and elucidate its relationships to ECC. Moreover, the predictive performances of the different methods are assessed and compared in a case study.

5.1.1 The MBMP approach

Based on the pioneering work of von Storch (1999), who proposed to use randomization in statistical downscaling, member-by-member postprocessing methods have emerged in the literature (Doblas-Reyes et al., 2005; Wood and Schaake, 2008; Johnson and Bowler, 2009; Van Schaeybroeck and Vannitsem, 2014). We fuse and complete the occasionally slightly differing formulations in these papers and present our reference member-by-member postprocessing (MBMP) approach in what follows.

In this context, we consider forecasts for a fixed weather quantity $i \in \{1, \ldots, I\}$ at a fixed location $j \in \{1, \ldots, J\}$ and for a fixed look-ahead time $k \in \{1, \ldots, K\}$, but omit for convenience the respective multi-index $\ell := (i, j, k)$. Forecasts and observations are assumed to stem from a normal distribution, as is reasonable in the case of temperature, pressure, u- and v-wind, for instance.

Let $x_1^\tau, \ldots, x_M^\tau$ denote the M-member raw ensemble forecast valid at date τ, and let y_τ be the corresponding verifying observation valid at date τ. Moreover, let the corresponding empirical ensemble mean and variance be denoted by

$$\bar{x}_\tau := \frac{1}{M} \sum_{m=1}^{M} x_m^\tau \quad \text{and} \quad s_\tau^2 := \frac{1}{M} \sum_{m=1}^{M} (x_m^\tau - \bar{x}_\tau)^2,$$

respectively.

Now let t be the verification date for which the raw ensemble forecast is to be postprocessed, and let $\{1, \ldots, D\}$ be a reference time period consisting of dates in the past of t. Using the ensemble forecast and observation data of the past period $\{1, \ldots, D\}$, we define the quantities

$$\mu_y := \frac{1}{D} \sum_{d=1}^{D} y_d, \qquad \mu_{\bar{x}} := \frac{1}{D} \sum_{d=1}^{D} \bar{x}_d, \qquad s_y^2 := \frac{1}{D} \sum_{d=1}^{D} (y_d - \mu_y)^2, \qquad s_{\bar{x}}^2 := \frac{1}{D} \sum_{d=1}^{D} (\bar{x}_d - \mu_{\bar{x}})^2$$

and

$$s_\varepsilon^2 := \frac{1}{D} \sum_{d=1}^{D} s_d^2.$$

For $\boldsymbol{y} := (y_1, \ldots, y_D)$ and $\bar{\boldsymbol{x}} := (\bar{x}_1, \ldots, \bar{x}_D)$, we additionally define the correlation coefficient $\rho := \mathrm{Corr}(\boldsymbol{y}, \bar{\boldsymbol{x}})$.

Based on the quantities defined above, the MBMP method transforms the raw ensemble forecast $x_1^t, \ldots, x_M^t$ valid at date t to a postprocessed ensemble forecast $\hat{x}_1^t, \ldots, \hat{x}_M^t$ valid at date t via

$$\begin{aligned} \hat{x}_m^t &:= \mu_y + \alpha(\bar{x}_t - \mu_{\bar{x}}) + \beta(x_m^t - \bar{x}_t) \\ &= (\mu_y - \alpha\mu_{\bar{x}}) + \alpha\bar{x}_t + \beta(x_m^t - \bar{x}_t) \end{aligned} \tag{5.1}$$

for $m \in \{1, \ldots, M\}$, with

$$\alpha := \rho \frac{s_y}{s_{\bar{x}}} \quad \text{and} \quad \beta := \sqrt{1 - \rho^2}\, \frac{s_y}{s_\varepsilon}.$$

If $\mu_y = 0$ and $\mu_{\bar{x}} = 0$, this method reduces to the approach described in Johnson and Bowler (2009), which on the other hand is a special variant for a single predictor in the general setting of Van Schaeybroeck and Vannitsem (2014) for multivariate predictors.

Elementary calculations yield that our reference MBMP ensemble has the same properties as the calibrated ensemble in Johnson and Bowler (2009), namely that the variance $s_{\hat{x}}^2$ of the calibrated ensemble is $s_{\hat{x}}^2 = \alpha^2 s_{\bar{x}}^2 + \beta^2 s_\varepsilon^2$ and that $\mathrm{Corr}(\hat{x}_m^t, \bar{x}_t) = \frac{\alpha s_{\bar{x}}}{s_{\hat{x}}}$. Hence, the specific choice of α and β can be indeed justified as in the derivations in Johnson and Bowler

(2009), and for each verification date, α and β are derived from a reference data set consisting of forecast and observation values in the past. The reference data base might contain data from several years. For comparative reasons, however, only the period of the $D = 30$ days before the verification date is used as the underlying reference data set in our case study.

For convenience, let the index t be suppressed in what follows. In our reference MBMP technique, a fixed raw ensemble member x_m is postprocessed into a calibrated ensemble member $\hat{x}_m$ via the affine transformation

$$\hat{x}_m = \mu_y - \alpha\mu_{\bar{x}} + \alpha\bar{x} + \beta(x_m - \bar{x}) = (\mu_y - \alpha\mu_{\bar{x}} + \alpha\bar{x} - \beta\bar{x}) + \beta x_m, \tag{5.2}$$

where $m \in \{1, \dots, M\}$. Hence, $\operatorname{rank}(x_m) = \operatorname{rank}(\hat{x}_m)$, and obviously the rank dependence structure of the raw ensemble is conserved in the calibrated ensemble.

These considerations remind us of the setting of sampling method (T) and Lemma 4.1, respectively, from Section 4.2. We recall that in the quantization approach (T), the raw ensemble forecast $x_1, \dots, x_M$ is transformed for each fixed multi-index ℓ, which remains suppressed in what follows, into a postprocessed ECC-T ensemble $\hat{x}_1, \dots, \hat{x}_M$ by setting

$$\hat{x}_1 := F^{-1}(S(x_1)), \dots, \hat{x}_M := F^{-1}(S(x_M)),$$

where F^{-1} denotes the inverse of a predictive CDF F obtained by univariate ensemble postprocessing, and S is a continuous CDF fitted to the raw ensemble values $x_1, \dots, x_M$. If S and F belong to the same location-scale family, that is, $S(x) = H(\frac{x-\mu}{\sigma})$ and $F(x) = H(\frac{x-\mu^*}{\sigma^*})$ for some continuous CDF H with $\mu, \mu^* \in \mathbb{R}$ and $\sigma, \sigma^* \in \mathbb{R}_+$, the transformation from x_m to $\hat{x}_m$ for $m \in \{1, \dots, M\}$ becomes affine. In particular,

$$\hat{x}_m := F^{-1}(S(x_m)) = \mu^* + \frac{\sigma^*}{\sigma}(x_m - \mu) \tag{5.3}$$

for $m \in \{1, \dots, M\}$, as has been shown in the proof of Lemma 4.1.

Thus, if we let F be the CDF of a $\mathcal{N}(\mu_y - \alpha\mu_{\bar{x}} + \alpha\bar{x}, \beta^2 s^2)$-distribution and S the CDF of a $\mathcal{N}(\bar{x}, s^2)$-distribution, with the quantities as defined before and the index t suppressed in the notation of the raw ensemble mean $\bar{x}$ and the raw ensemble variance s^2, we recover the MBMP reference approach (5.2). This can be checked by plugging in formula (5.3), obtaining

$$\hat{x}_m = \mu_y - \alpha\mu_{\bar{x}} + \alpha\bar{x} + \frac{\beta s}{s}(x_m - \bar{x}) = \mu_y - \alpha\mu_{\bar{x}} + \alpha\bar{x} + \beta(x_m - \bar{x})$$

for $m \in \{1, \dots, M\}$. When we combine the postprocessed univariate MBMP ensemble members for each location, weather quantity and prediction horizon separately in an ECC fashion, the fourth step of ECC is obsolete, as no reordering of these samples is necessary due to the affine transformation as discussed before. The index of the calibrated members is directly defined by the method itself, and the rank dependence pattern from the raw ensemble is retained by construction. In this light, the MBMP technique can be considered a variant of ECC-T.

5.1.2 Case study

In our case study, we aim at a comparison of the MBMP method and EMOS as described in Section 3.1.2 and EMOS combined with ECC, respectively, in univariate and multivariate

Table 5.1: Average CRPS (in °C) and DSS for univariate 24 hour ahead forecasts of temperature at Berlin and Hamburg. The results are averaged over the test period from 1 January 2011 to 31 December 2012.

	CRPS		DSS	
	Berlin	Hamburg	Berlin	Hamburg
Raw Ensemble	0.975	0.832	8.20	7.00
MBMP Ensemble	0.776	0.697	2.83	2.14
EMOS-T Ensemble	0.761	0.681	1.85	1.51
EMOS-Q Ensemble	0.754	0.679	1.95	1.59

settings. To this end, we focus on 24 hour ahead 50-member ECMWF ensemble forecasts for temperature at Berlin and Hamburg. The predictive performance of the different methods is evaluated over a two-year test period from 1 January 2011 to 31 December 2012, by using our usual verification tools. For comparative reasons, we use a rolling window consisting of the last 30 days before the verification date as training period or reference data set, respectively, for both the MBMP method and the EMOS postprocessing. Again, as we only focus on two distinct stations and out test period only comprises two years, our case study is not meant to be a comprehensive comparison study, but should be rather viewed as a proof-of-concept. We first show results for univariate settings and then turn to the multivariate case.

(a) Univariate setting

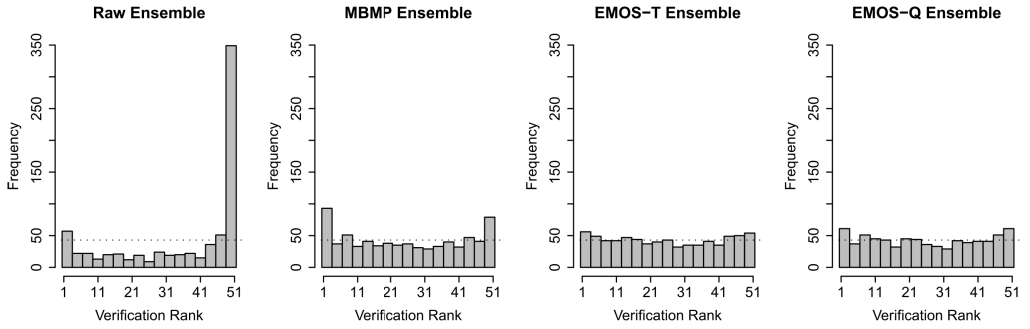
Initially, we compare the ensemble temperature forecasts at Berlin and Hamburg individually, issued by the raw ensemble, the MBMP ensemble obtained according to (5.1) and the EMOS-T and EMOS-Q ensembles, respectively.

We recall that the EMOS-T ensemble here consists of the 50 samples according to the transformation approach (T) from the corresponding univariate EMOS predictive CDF F, where the ordering is not relevant in the univariate case. In the Gaussian case of temperature here, the predictive EMOS CDF F is that of a $\mathcal{N}(a+b(x_1+\ldots+x_M), c+ds^2)$-distribution, with the raw ensemble forecast $x_1, \ldots, x_M$, the raw ensemble variance s^2 and parameters a, b, c and d estimated via minimization of the training continuous ranked probability score (CRPS), as stated in (3.5) and Section 3.1.2. The CDF S fitted to the raw ensemble is taken to be that of a $\mathcal{N}(\bar{x}, s^2)$-distribution with mean equal to the raw ensemble mean $\bar{x}$ and variance equal to the raw ensemble variance s^2. According to the affine transformation in (5.3), the postprocessed EMOS-T ensemble $\hat{x}_1, \ldots, \hat{x}_M$ is then given by

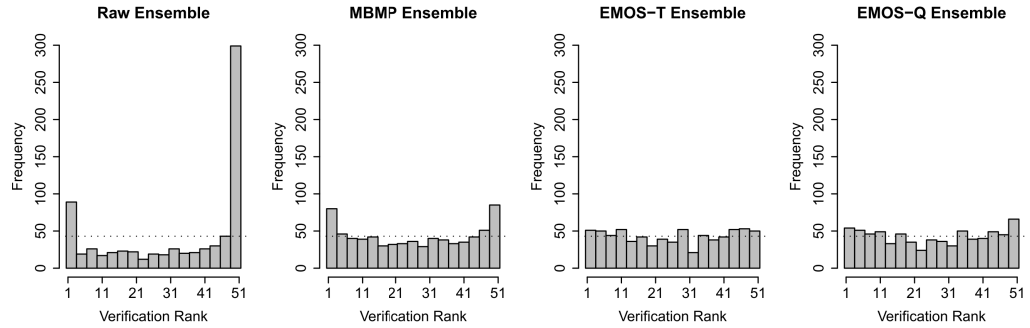
$$\hat{x}_m = a + b(x_1 + \ldots + x_M) + \frac{\sqrt{c+ds^2}}{s}(x_m - \bar{x}), \tag{5.4}$$

where $m \in \{1, \ldots, M\}$, and retains the rank dependence structures of the raw ensemble by construction.

In contrast, the EMOS-Q ensemble consists of the 50 equidistant $\frac{1}{51}$-, $\ldots$, $\frac{50}{51}$-quantiles from the corresponding univariate EMOS predictive CDF F according to sampling scheme (Q), where again the ordering is not relevant in the univariate case.



(a) Berlin



(b) Hamburg

Figure 5.1: Verification rank histograms for 24 hour ahead temperature forecasts at (a) Berlin and (b) Hamburg for the test period from 1 January 2011 to 31 December 2012

For an overall assessment of the predictive performances, summarized in Table 5.1, we employ the CRPS and the univariate Dawid-Sebastiani score (DSS), where the results are averaged over the test period. The CRPS for the ensemble forecasts is calculated via Formula (2.6), while the DSS is computed according to Formula (2.9), where we use the corresponding empirical mean $\bar{x}$ and the empirical variance s^2 of the different ensembles, respectively. For the assessment of calibration solely, we use the verification rank histograms shown in Figure 5.1.

As the focus is on the comparison of different ensembles, we do not explicitly show results for the full EMOS predictive distribution here. However, the full EMOS density generally outperforms the best EMOS-based ensemble, which is expected and intuitive, as the corresponding ensembles are discretizations of the EMOS distribution. The full EMOS density forecast very slightly outperforms the best performing EMOS-based ensemble in terms of the CRPS, namely EMOS-Q, but the results become identical when rounding to two decimal places.

According to Table 5.1, the MBMP, EMOS-T and EMOS-Q ensembles clearly outperform the raw ensemble in terms of the CRPS and DSS, and the EMOS-based ensembles outperform the MBMP approach. With respect to the CRPS, this might be partly due to the fact that EMOS is designed such that the CRPS is minimized, and MBMP is not. Moreover, the EMOS-Q ensemble performs best among the ensemble forecasts regarding the CRPS, which appears to be well in line with the results in Section 4.2. With respect to the DSS,

Table 5.2: Average ES (in °C) and DSS for 24 hour ahead forecasts of temperature at Berlin and Hamburg jointly. The results are averaged over the test period from 1 January 2011 to 31 December 2012; those for the Individual EMOS ensembles are averaged over 100 runs.

	ES	DSS
Raw Ensemble	1.400	15.15
MBMP Ensemble	1.146	5.07
Individual EMOS-T Ensemble	1.125	3.41
ECC-T Ensemble	1.125	3.34
Individual EMOS-Q Ensemble	1.122	3.60
ECC-Q Ensemble	1.121	3.54

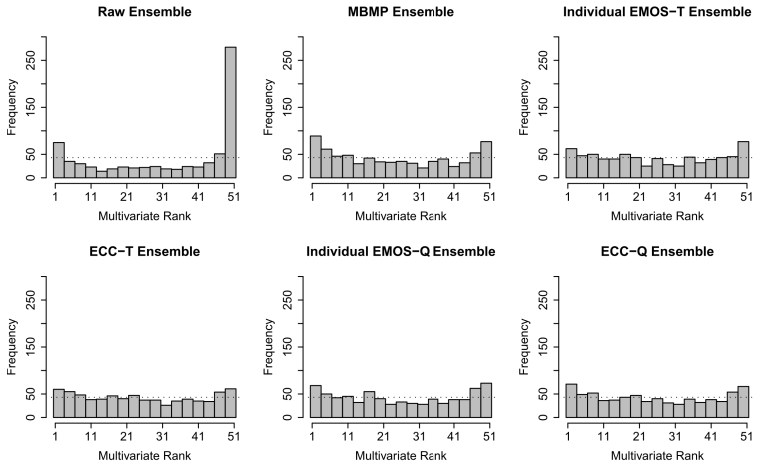
the EMOS-T ensemble outperforms the EMOS-Q ensemble. The variance of the EMOS-T ensemble turns out to be greater than that of the EMOS-Q ensemble throughout, whereas the mean is identical for both EMOS-based ensembles. Indeed, the empirical average of the 50 equidistant quantiles yields the median, hence the mean of the EMOS normal distribution due to symmetry. The empirical average of the EMOS-T ensemble forecasts is also equal to this EMOS mean, which can be verified by computing the empirical average of the ensemble member values $\hat{x}_m$ in Formula (5.4) in a straightforward calculation. Hence, optimal sampling with respect to the DSS appears to be different from that in case of the CRPS.

Regarding calibration only, all postprocessed ensembles are better calibrated than the raw ensemble, which reveals U-shaped verification rank histograms. Moreover, the EMOS-T and EMOS-Q ensembles also outperform the MBMP ensemble in terms of calibration, while showing basically an equally good performance with verification rank histograms being the closest to uniform.

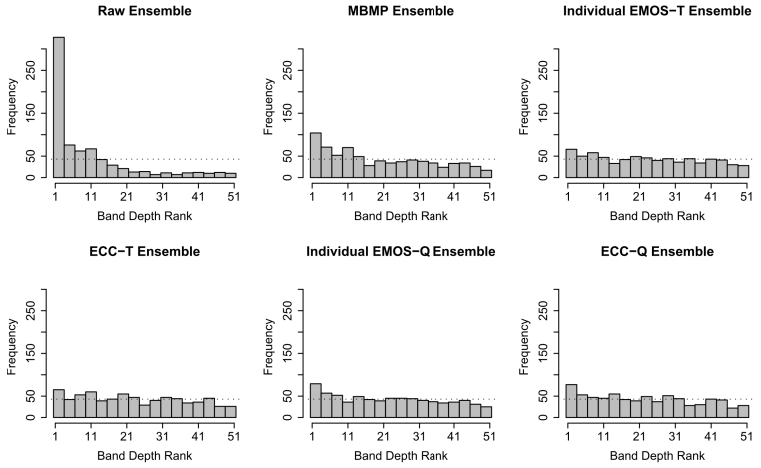
(b) Multivariate setting

Turning to the multivariate case, we now focus on spatial aspects by considering the two stations Berlin and Hamburg simultaneously. We compare

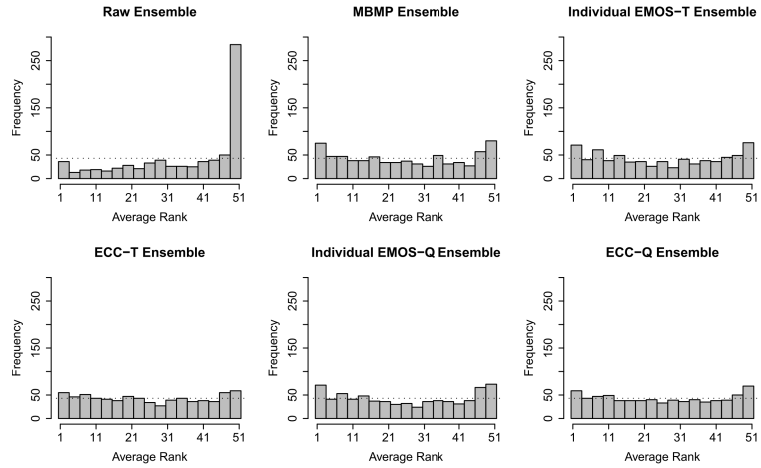
- the unprocessed ECMWF raw ensemble,
- the MBMP ensemble, in which the two univariately postprocessed ensemble forecasts obtained by MBMP via (5.1) are aggregated,
- the Individual EMOS-T ensemble, in which the order of the ensemble members obtained by EMOS-T via (5.4) is randomly shuffled, rather than being determined by the raw ensemble and remaining fixed by the monotone transformation,
- the ECC-T ensemble, which combines the univariate ensemble forecasts obtained by EMOS-T according to (5.4), where the rank dependence structure of the raw ensemble is retained by construction via the affine transformation,
- the Individual EMOS-Q ensemble, in which the order of the 50 equidistant $\frac{1}{51}$-, ..., $\frac{50}{51}$-quantiles from the EMOS predictive CDF obtained via quantization scheme (Q) is randomly shuffled, and
- the ECC-Q ensemble, which combines the univariate ensemble forecasts obtained by EMOS-Q, where the corresponding 50 equidistant quantiles are reordered according to the rank dependence structure of the raw ensemble.



(a) Multivariate rank histograms



(b) Band depth rank histograms



(c) Average rank histograms

Figure 5.2: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature forecasts at Berlin and Hamburg jointly for the test period from 1 January 2011 to 31 December 2012

We employ the energy score (ES) for ensemble forecasts according to (2.8) and the multivariate version of the DSS (2.10) as scores. For the calculation of the DSS, $\boldsymbol{\mu}$ is taken to be the empirical mean vector, and Σ is taken to be the empirical covariance matrix of the respective ensembles. In case of the Individual EMOS ensembles, the results for the scores are averaged over 100 runs. Again, all scores are averaged over the test period, and the results are reported in Table 5.2. Multivariate calibration is assessed via the multivariate, band depth and average rank histograms in Figure 5.2, where in the case of the Individual EMOS ensembles, the corresponding ranks are averages over 100 runs.

Concerning the ES and the DSS, all postprocessed ensembles outperform the raw ensemble. Among the postprocessed ensembles, the Individual EMOS and the ECC ensembles yield better scores than the MBMP ensemble, both with respect to the ES and the DSS. In this light, the results from the univariate case appear to extend to the multivariate case. This might also explain that the Individual EMOS ensembles perform better than the MBMP ensemble, although they do not conserve rank dependencies. In terms of the ES, the ECC-T and the ECC-Q ensemble, respectively, perform equally well or slightly outperform their corresponding counterparts Individual EMOS-T and Individual EMOS-Q, for which the spatial rank dependence structure is lost, with the ECC-Q ensemble showing the best ES overall, as is already the case with respect to the CRPS in the univariate setting. The minor differences between the ECC ensembles and their corresponding Individual EMOS counterparts might again be partly explained by the discrimination inability of the energy score (Pinson and Tastu, 2013), which has already been discussed in the context of Section 4.3. Regarding the DSS, the differences between the respective Individual EMOS and ECC ensembles become more obvious than with the ES, in that the ECC ensembles perform better, with the ECC-T ensemble revealing the best DSS overall, as is already the case in the univariate setting.

In terms of calibration only, all postprocessed ensembles clearly outperform the raw ensemble. In addition, all Individual EMOS and ECC ensemble variants, respectively, show a somewhat better calibration than the MBMP ensemble, where the ECC ensembles might slightly outperform their corresponding Individual EMOS counterparts.

In a nutshell, all postprocessed ensembles in our case study dealing with temperature perform notably better than the raw ensemble, while the EMOS- and ECC-based ensembles, respectively, outperform the MBMP ensemble, both in the univariate and multivariate setting. In the multivariate case, the ECC ensembles slightly outperform their Individual EMOS counterparts. As described, even though the MBMP ensemble clearly outperforms the raw ensemble, it does not yield particularly good results compared to EMOS and ECC, respectively, in the case of temperature here. However, this may turn out differently in other settings. For example, the MBMP ensemble might perform better if the MBMP parameters were estimated based on a larger reference database comprising much more than the last 30 days used for comparative reasons here.

5.2 Other examples of ECC variants in the extant literature

In addition to the MBMP approach discussed in Section 5.1, there are many other techniques, mainly in meteorological or hydrological contexts, that can be viewed as ECC variants or that contain elements of ECC. Examples are the methods of Bremnes (2007), Krzysztofowicz and Toth (2008), Todini (2008), Kann et al. (2009), Kann et al. (2011),

Flowerdew (2012), Pinson (2012), Roulin and Vannitsem (2012), Flowerdew (2014) and Van Schaeybroeck and Vannitsem (2014), while this list is not expected to be complete.

In what follows, we exemplarily investigate in detail the connections of ECC to the approaches of Pinson (2012) and Roulin and Vannitsem (2012), respectively.

5.2.1 Pinson (2012)

We recall the member-by-member postprocessing approach of Pinson (2012) for the postprocessing of raw ensemble forecasts of (u, v)-wind vectors from Section 3.2.2. At a fixed location and for a fixed prediction horizon, Pinson (2012) essentially introduces a two-dimensional translation and dilation of the sets of the u- and v-wind raw ensemble predictions $u_1, \ldots, u_M$ and $v_1, \ldots, v_M$ with empirical means $\bar{u}$ and $\bar{v}$ and empirical standard deviations s_u and s_v, respectively.

Specifically, the postprocessed u- and v-wind ensemble members $\hat{u}_1, \ldots, \hat{u}_M$ and $\hat{v}_1, \ldots, \hat{v}_M$ in Pinson's approach are given by

$$\begin{aligned} \hat{u}_m &:= \bar{u} + \tau_u + \xi_u(u_m - \bar{u}), \\ \hat{v}_m &:= \bar{v} + \tau_v + \xi_v(v_m - \bar{v}) \end{aligned}$$

for $m \in \{1, \ldots, M\}$, with translation factors τ_u and τ_v and dilation factors ξ_u and ξ_v obtained via

$$\begin{aligned} \tau_u &:= \theta_u^{(1)} + (\theta_u^{(2)} - 1)\bar{u} + \theta_u^{(3)}\bar{v}, \\ \tau_v &:= \theta_v^{(1)} + \theta_v^{(2)}\bar{u} + (\theta_v^{(3)} - 1)\bar{v}, \\ \xi_u &:= \frac{\exp(\gamma_u^{(1)})}{s_u} + \exp(\gamma_u^{(2)}), \\ \xi_v &:= \frac{\exp(\gamma_v^{(1)})}{s_v} + \exp(\gamma_v^{(2)}). \end{aligned}$$

In this context, the model parameters $\theta_u^{(1)}, \theta_u^{(2)}, \theta_u^{(3)}, \gamma_u^{(1)}, \gamma_u^{(2)}, \theta_v^{(1)}, \theta_v^{(2)}, \theta_v^{(3)}, \gamma_v^{(1)}, \gamma_v^{(2)}$ are estimated adaptively using a recursive maximum likelihood approach, with exponential forgetting of past observations.

In Pinson's approach, each postprocessed margin is essentially a translated and dilated variant of the original unprocessed margin, while the mapping is compatible with the sampling method (T) and the ECC-T scheme, where both S and F are normal. Exemplarily for u-wind, we consider the affine transformation setting in (5.3) and arrive at Pinson's approach if we let F be the CDF of a $\mathcal{N}(\bar{u} + \tau_u, \xi_u^2 s_u^2)$-distribution, that is, with $a = \tau_u$, $b = b_1 = \ldots = b_M = 1/M$, $c = 0$, $d = \xi_u^2$ and $\varepsilon = \xi_u(x_m - \bar{u})$ with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \xi_u^2 s_u^2$ in the EMOS setting (3.3) for normal distributions in Section 3.1.2, and S be the CDF of a $\mathcal{N}(\bar{u}, s_u^2)$-distribution.

The same considerations hold analogously for v-wind, and hence, Pinson's method can indeed be considered an ECC-T variant.

5.2.2 Roulin and Vannitsem (2012)

Roulin and Vannitsem (2012) employ extended logistic regression to postprocess and calibrate areal precipitation forecasts, where the parameters are estimated from a hindcast ensemble of past meteorological situations using the same model as the raw ensemble. For areal precipitation Y at a fixed location and for a fixed prediction horizon, they model the corresponding distribution via the CDF

$$F_Y(q) = \mathbb{P}(Y \leq q) = H(f(\xi) + g(q)), \tag{5.5}$$

where $H(z) = \frac{1}{1+\exp(-z)}$. They choose

$$\xi := \frac{1}{M} \sum_{m=1}^{M} x_m^{1/4} \tag{5.6}$$

as predictor, where $x_1, \ldots, x_M$ denotes the raw ensemble forecast for areal precipitation. Further, the functions f and g are set to be $f(\xi) := \beta_0 + \beta_1 \xi$ and $g(q) := \beta_2 q^{1/2}$, with parameters β_0, β_1 and β_2. Roulin and Vannitsem (2012) estimate the parameter vector $\boldsymbol{\beta} := (\beta_0, \beta_1, \beta_2)$ by maximizing the likelihood function

$$\mathcal{L}(\boldsymbol{\beta}) := \sum_{n=1}^{N} \sum_{r=1}^{R} \{y_{nr}(\beta_0 + \beta_1 \xi_n + \beta_2 q_r^{1/2}) - \log(1 + \exp(\beta_0 + \beta_1 \xi_n + \beta_2 q_r^{1/2}))\},$$

where N denotes the number of realizations in the calibration dataset, R is the number of the selected thresholds, and

$$y_{nr} := \begin{cases} 1 & \text{if the observed precipitation } y_n \text{ does not exceed the threshold } q_r, \\ 0 & \text{otherwise.} \end{cases}$$

In their specific implementation, they select $R = 7$ precipitation thresholds q_r, $r \in \{1, \ldots, 7\}$, such that these cover the upper quantiles of the distribution. For the n-th realization, Roulin and Vannitsem (2012) consider $\overline{w}_n = \xi_n + \varepsilon_n$, where

$$\overline{w}_n := \frac{1}{M} \sum_{m=1}^{M} x_{nm}^{1/4},$$

and ε_n denotes the uncertainty of the ensemble mean evaluation, which is supposed to be a random process with $\mathbb{E}[\varepsilon_n] = 0$ and $\sigma_{\varepsilon_n}^2 := \mathrm{Var}(\varepsilon_n) = \frac{\mathrm{av}(\sigma_w^2)}{M}$. In this context, σ_w^2 denotes the variance of the power-transformed ensemble members, and $\mathrm{av}(\cdot)$ the average over N realizations. Roulin and Vannitsem (2012) argue that $\mathbb{E}(\xi_n | \overline{w}_n)$ should actually be used in place of ξ_n in the extended logistic regression analysis, and they provide an approximate expression for $\mathbb{E}(\xi_n | \overline{w}_n)$.

To get postprocessed ensembles rather than separate probability distributions at each forecast step, Roulin and Vannitsem (2012) modify the raw ensembles in accordance to the results obtained by the extended logistic regression postprocessing as follows.

Roulin and Vannitsem (2012) show via a Taylor expansion that the variance of the corrected ensembles is proportional to its mean if no ensemble member predicts zero precipitation. According to (5.5), the probability of zero precipitation is $p_0 := \frac{1}{1+\exp(-\beta_0 - \beta_1 \xi)}$.

For a fixed forecast day, the raw ensemble member forecasts $x_1, \ldots, x_M$ are sorted increasingly, that is, $x_{(1)} \leq \ldots \leq x_{(M)}$, with any ties resolved at random. Each forecast $x_{(m)}$, where $m \in \{1, \ldots, M\}$, is assigned a probability $p_{(m)}$ via

$$p_{(m)} := \mathbb{P}(Y \leq x_{(m)}) = \frac{m + \frac{1}{2}}{M + 1}. \tag{5.7}$$

Then, Roulin and Vannitsem (2012) generate a postprocessed ensemble $\tilde{x}_1, \ldots, \tilde{x}_M$ by modifying the forecast values of the members ranked $m \in \{1, \ldots, M\}$ according to their probabilities via

$$\tilde{x}_{(m)} := \left(\frac{\log\left(\frac{1-p_{(m)}}{p_{(m)}}\right) - \beta_0 - \beta_1 \xi}{\beta_2} \right)^2, \tag{5.8}$$

with ξ as defined in (5.6) and parameters β_0, β_1 and β_2 as before.

If $p_{(m)} \leq p_0$, the precipitation $\tilde{x}_{(m)}$ is set to zero, that is, $\tilde{x}_{(m)} := 0$. If $p_{(m_0)} > p_0$, where m_0 denotes the number of ensemble members forecasting zero precipitation, only a fraction $r_0 := p_0/p_{(m_0)}$ of the m_0 members is set to zero at random, and the remaining members are randomly given a value between p_0 and $p_{(m_0)}$ with precipitation specified by (5.8).

Finally, Roulin and Vannitsem (2012) reassign the postprocessed values obtained via (5.8) to the corresponding members such that the raw ensemble precipitation trajectories for different prediction horizons are reconstructed.

The method of Roulin and Vannitsem (2012) can be interpreted as an ECC-Q variant, as it first extracts equidistant quantiles according to (5.7) from the postprocessed marginal predictive CDFs being of extended logistic type. Then, a reordering with respect to the rank dependence structure of the raw ensemble as in ECC is performed, with adaptations allowing to deal with a point mass at zero, to capture the temporal dependence among the raw ensemble forecasts for several look-ahead times. Although not explicitly demonstrated by Roulin and Vannitsem (2012), their method is –as is ECC– able to reconstruct spatial rank dependence patterns of the raw ensembles.

Chapter 6

Multivariate discrete copulas: The theoretical frame

In this chapter, we develop the theoretical frame of the standard ECC approach by introducing and investigating multivariate discrete copulas and thus showing that ECC can be viewed as a copula technique, as suggested by its name.

As already pointed out in Section 3.2.1, copulas (Nelsen, 2006) play important roles in probability and statistics whenever the modeling of stochastic dependence is required, have been employed in numerous application areas and are also interesting from a purely theoretical-mathematical point of view. Copulas form a special Fréchet class, where this term refers to a class of multivariate distributions with fixed uni- or multivariate margins (Fréchet, 1951; Joe, 1997). Specifically, copulas are L-variate CDFs, where $L \in \mathbb{N}$, $L \geq 2$, with standard uniform univariate margins $F_1 = \cdots = F_L = F_{\mathcal{U}([0,1])}$, thus forming the Fréchet class $\mathcal{F}(F_1, \ldots, F_L) = \mathcal{F}(F_{\mathcal{U}([0,1])}, \ldots, F_{\mathcal{U}([0,1])})$.

A special type of copulas are the so-called discrete copulas, whose properties have been studied by Mayor et al. (2005), Mesiar (2005), Kolesárová et al. (2006) and Mayor et al. (2007) in the last decade. However, the discussion in the papers mentioned above focuses on the bivariate case, and it is natural to seek a treatment of the general multivariate situation. In what follows, we generalize both the notion of discrete copulas and the most important results in this context to the multivariate case, and show to what extent they build the theoretical frame of the ECC approach (Schefzik et al., 2013) presented in Section 4.1 and also of the related Schaake shuffle (Clark et al., 2004) discussed in Section 3.2.2.

Specifically, we introduce the multivariate discrete copula concept in Section 6.1. We then point out the connection between multivariate discrete copulas and stochastic arrays (Csima, 1970; Marchi and Tarazaga, 1979) in Section 6.2 and continue with the formulation of a multivariate discrete version of Sklar's theorem in Section 6.3. Eventually, Section 6.4 deals with the relationships of the presented results to ECC and the Schaake shuffle.

This chapter is based on the findings in Schefzik (2013), with the origin lying in Schefzik (2011).

6.1 Multivariate discrete copulas

First, we extend the notions of bivariate discrete copulas (Mayor et al., 2005; Kolesárová et al., 2006) and bivariate discrete subcopulas (Mayor et al., 2007) to the general multivariate case.

Let $I_M := \left\{0, \frac{1}{M}, \frac{2}{M}, \ldots, \frac{M-1}{M}, 1\right\}$, where $M \in \mathbb{N}$, and let $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$.

Definition 6.1. A function $D : I_M^L \to [0,1]$ is a discrete copula on I_M^L if it satisfies the following conditions.

(D1) D is grounded, in that $D\left(\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right) = 0$ if $i_\ell = 0$ for at least one $\ell \in \{1, \ldots, L\}$.

(D2) $D(1, \ldots, 1, \frac{i_\ell}{M}, 1, \ldots, 1) = \frac{i_\ell}{M}$ for all $\ell \in \{1, \ldots, L\}$.

(D3) D is L-increasing, in that

$$\Delta_{i_L-1}^{i_L} \cdots \Delta_{i_1-1}^{i_1} D\left(\frac{k_1}{M}, \ldots, \frac{k_L}{M}\right) \geq 0$$

for all $i_\ell \in \{1, \ldots, M\}$, $\ell \in \{1, \ldots, L\}$, where

$$\begin{aligned}\Delta_{i_\ell-1}^{i_\ell} D\left(\frac{k_1}{M}, \ldots, \frac{k_L}{M}\right) &:= D\left(\frac{k_1}{M}, \ldots, \frac{k_{\ell-1}}{M}, \frac{i_\ell}{M}, \frac{k_{\ell+1}}{M}, \ldots, \frac{k_L}{M}\right) \\ &\quad - D\left(\frac{k_1}{M}, \ldots, \frac{k_{\ell-1}}{M}, \frac{i_\ell - 1}{M}, \frac{k_{\ell+1}}{M}, \ldots, \frac{k_L}{M}\right).\end{aligned}$$

Definition 6.2. A discrete copula $D : I_M^L \to [0,1]$ is irreducible if it has minimal range, that is, $\mathrm{Ran}(D) = I_M$.

Following Fréchet (1951) and Chapter 3 in Joe (1997), a multivariate discrete copula can be interpreted as a multivariate distribution in the Fréchet class $\mathcal{F}(F_{\mathcal{U}(I_M)}, \ldots, F_{\mathcal{U}(I_M)})$, where $F_{\mathcal{U}(I_M)}$ is the CDF of a uniformly distributed random variable on I_M.

Definition 6.3. A function $D^* : J_M^{(1)} \times \cdots \times J_M^{(L)} \to [0,1]$ with $\{0,1\} \subset J_M^{(1)}, \ldots, J_M^{(L)} \subset I_M$ is a discrete subcopula if it satisfies the following conditions.

(S1) $D^*\left(\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right) = 0$ if $i_\ell = 0$ for at least one $\ell \in \{1, \ldots, L\}$.

(S2) $D^*(1, \ldots, 1, \frac{i_\ell}{M}, 1, \ldots, 1) = \frac{i_\ell}{M}$ for all $\frac{i_\ell}{M} \in J_M^{(\ell)}$.

(S3)

$$\Delta_{i_L}^{j_L} \cdots \Delta_{i_1}^{j_1} D^*\left(\frac{k_1}{M}, \ldots, \frac{k_L}{M}\right) \geq 0$$

for all $(\frac{i_1}{M}, \ldots, \frac{i_L}{M}), (\frac{j_1}{M}, \ldots, \frac{j_L}{M}) \in J_M^{(1)} \times \cdots \times J_M^{(L)}$ such that $i_\ell \leq j_\ell$ for $\ell \in \{1, \ldots, L\}$, where

$$\begin{aligned}\Delta_{i_\ell}^{j_\ell} D^*\left(\frac{k_1}{M}, \ldots, \frac{k_L}{M}\right) &:= D^*\left(\frac{k_1}{M}, \ldots, \frac{k_{\ell-1}}{M}, \frac{j_\ell}{M}, \frac{k_{\ell+1}}{M}, \ldots, \frac{k_L}{M}\right) \\ &\quad - D^*\left(\frac{k_1}{M}, \ldots, \frac{k_{\ell-1}}{M}, \frac{i_\ell}{M}, \frac{k_{\ell+1}}{M}, \ldots, \frac{k_L}{M}\right).\end{aligned}$$

Definition 6.4. A discrete subcopula $D^* : J_M^{(1)} \times \cdots \times J_M^{(L)} \to [0,1]$ is irreducible if $\mathrm{Ran}(D^*) = I_M$.

The definition of discrete (sub)copulas can be generalized. A discrete copula need not necessarily have domain I_M^L, but can generally be defined on $I_{M_1} \times \cdots \times I_{M_L}$, where $M_1, \ldots, M_L \in \mathbb{N}$ might take distinct values. Then, the axioms (D1), (D2) and (D3) apply analogously to this case. Similarly, discrete subcopulas can generally be defined on $J_{M_1}^{(1)} \times \cdots \times J_{M_L}^{(L)}$ for possibly distinct numbers $M_1, \ldots, M_L \in \mathbb{N}$, satisfying the conditions in Definition 6.3.

For convenience and in view of ECC and the Schaake shuffle as applications in Section 6.4, we confine ourselves to the case of $M := M_1 = \cdots = M_L$ as in the above Definitions 6.1 to 6.4 in what follows.

Tailored to the applications to be discussed in Section 6.4, we defined the multivariate discrete copula on points that are equally spaced across the set I_M^L. However, when considering a multivariate distribution with discrete margins, the points where the copula is of interest typically do not need to be equidistant across I_M^L, but rather are heterogeneously spaced across the marginals. Such more general situations are studied in Genest and Nešlehová (2007) and Genest et al. (2014), for instance.

We now give first explicit examples of multivariate discrete copulas.

Example 6.5. Let $i_1, \ldots, i_M \in \{0, 1, \ldots, M\}$.

(a) $\Pi\left(\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right) := \prod_{\ell=1}^{L} \frac{i_\ell}{M}$ is a discrete copula, the so-called product or independence copula.

(b) $\mathcal{M}\left(\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right) := \min\left\{\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right\}$ is an irreducible discrete copula.

As they represent the restrictions of two well-known standard copulas defined on $[0,1]^L$ to the discrete set I_M^L, Π and $\mathcal{M}$ are indeed multivariate discrete copulas.

Example 6.6. Another example for an irreducible discrete copula is given by the so-called empirical copula (Rüschendorf, 1976), which has already been defined in Section 3.2.1 and is now recalled in our context here. The empirical copula, which has also become popular under the term "empirical dependence function" (Deheuvels, 1979), is extremely important and relevant in view of both ECC and the Schaake shuffle, for which it provides the theoretical background, as we will see in Section 6.4.

Let $\mathcal{S} := \{(x_1^1, \ldots, x_1^L), \ldots, (x_M^1, \ldots, x_M^L)\}$, where $x_m^\ell \in \mathbb{R}$ for all $m \in \{1, \ldots, M\}$ and $\ell \in \{1, \ldots, L\}$ with $x_m^1 \neq x_\mu^1, \ldots, x_m^L \neq x_\mu^L$ for $m, \mu \in \{1, \ldots, M\}$, $m \neq \mu$. That is, we assume for simplicity that there are no ties among the respective samples. Moreover, let $x_{(1)}^1 < \ldots < x_{(M)}^1, \ldots, x_{(1)}^L < \ldots < x_{(M)}^L$ be the marginal order statistics of the collections $\{x_1^1, \ldots, x_M^1\}, \ldots, \{x_1^L, \ldots, x_M^L\}$, respectively.

Then, the empirical copula $E_M : I_M^L \to I_M$ defined from $\mathcal{S}$ is given by

$$E_M\left(\frac{i_1}{M},\ldots,\frac{i_L}{M}\right)$$
$$:=\begin{cases}0 & \text{if } i_\ell=0 \text{ for at least one } \ell\in\{1,\ldots,L\},\\ \frac{\#\{(x_m^1,\ldots,x_m^L)\in\mathcal{S}\,|\,x_m^1\le x_{(i_1)}^1,\ldots,x_m^L\le x_{(i_L)}^L\}}{M} & \text{if } i_\ell\in\{1,\ldots,M\} \text{ for all } \ell\in\{1,\ldots,L\}.\end{cases}$$

Equivalently,

$$E_M\left(\frac{i_1}{M},\ldots,\frac{i_L}{M}\right) = \frac{1}{M}\sum_{m=1}^{M}\mathbb{1}_{\{\mathrm{rank}(x_m^1)\le i_1,\ldots,\mathrm{rank}(x_m^L)\le i_L\}} = \frac{1}{M}\sum_{m=1}^{M}\prod_{\ell=1}^{L}\mathbb{1}_{\{\mathrm{rank}(x_m^\ell)\le i_\ell\}},$$

as in Section 3.2.1.

Obviously, the empirical copula is an irreducible discrete copula. Conversely, any irreducible discrete copula is the empirical copula of some set $\mathcal{S}$, as will be discussed in Example 6.10 (c) in Section 6.2.

Asymptotic theory for the corresponding empirical processes is provided in Rüschendorf (1976), van der Vaart and Wellner (1996), Fermanian et al. (2004) and Rüschendorf (2009), for instance.

6.2 A characterization of multivariate discrete copulas using stochastic arrays

According to Mayor et al. (2005) and Kolesárová et al. (2006), there is a one-to-one correspondence between discrete copulas and bistochastic matrices in the bivariate case. We now formulate a similar characterization for multivariate discrete copulas. To this end, the notion of stochastic arrays (Csima, 1970; Marchi and Tarazaga, 1979) is required.

Definition 6.7. An array $A := (a_{i_1\ldots i_L})_{i_1,\ldots,i_L=1}^{M}$ is an L-dimensional stochastic array, or an L-stochastic matrix, of order M if the following conditions hold.

(A1) $a_{i_1\ldots i_L}\ge 0$ for all $i_1,\ldots,i_L\in\{1,\ldots,M\}$.

(A2) $\sum_{i_1=1}^{M}\cdots\sum_{i_{\ell-1}=1}^{M}\sum_{i_{\ell+1}=1}^{M}\cdots\sum_{i_L=1}^{M}a_{i_1\ldots i_{\ell-1}i_\ell i_{\ell+1}\ldots i_L}=1$ for $i_\ell\in\{1,\ldots,M\}$, $\ell\in\{1,\ldots,L\}$.

As a special case, an L-dimensional stochastic array A is an L-dimensional permutation array, or an L-permutation matrix, if the entries of A only take the values 0 and 1, that is, $a_{i_1\ldots i_L}\in\{0,1\}$ for all $i_1,\ldots,i_L\in\{1,\ldots,M\}$.

Theorem 6.8. Let $D: I_M^L\to[0,1]$. Then, the following statements are equivalent.

(1) D is a discrete copula.

(2) There exists an L-dimensional stochastic array $A := (a_{i_1\ldots i_L})_{i_1,\ldots,i_L=1}^{M}$ of order M such that

$$D\left(\frac{i_1}{M},\ldots,\frac{i_L}{M}\right)=\frac{1}{M}\sum_{\nu_1=1}^{i_1}\cdots\sum_{\nu_L=1}^{i_L}a_{\nu_1\ldots\nu_L} \tag{6.1}$$

for $i_1,\ldots,i_L\in\{0,1,\ldots,M\}$.

Proof. (1) ⇒ (2):
If $i_\ell = 0$ for at least one $\ell \in \{1, \dots, L\}$, $D(\frac{i_1}{M}, \dots, \frac{i_L}{M}) = 0$ by axiom (D1) in Definition 6.1, in accordance with setting the empty sum equal to zero by convention, that is, $\sum_{n=1}^{0} c_n := 0$ for some sequence $(c_n)_{n\in\mathbb{N}}$. For $i_1, \dots, i_L \in \{1, \dots, M\}$, we set

$$a_{i_1 \dots i_L} := M \cdot \left(\Delta_{i_L - 1}^{i_L} \cdots \Delta_{i_1 - 1}^{i_1} D\left(\frac{k_1}{M}, \dots, \frac{k_L}{M} \right) \right),$$

with $\Delta_{i_\ell - 1}^{i_\ell} D$ as defined in axiom (D3) in Definition 6.1, and $A := (a_{i_1 \dots i_L})_{i_1, \dots, i_L = 1}^{M}$ and show that A satisfies the axioms (A1) and (A2) from Definition 6.7 and therefore is an L-dimensional stochastic array of order M.

(A1) Since D is a discrete copula, D is L-increasing, that is, $\Delta_{i_L - 1}^{i_L} \cdots \Delta_{i_1 - 1}^{i_1} D(\frac{k_1}{M}, \dots, \frac{k_L}{M}) \geq 0$. Hence, $a_{i_1 \dots i_L} \geq 0$ by definition, and (A1) is fulfilled.

(A2) Now let $\ell \in \{1, \dots, L\}$ be fixed. We have to show that

$$S_1 := \sum_{i_1 = 1}^{M} \cdots \sum_{i_{\ell-1} = 1}^{M} \sum_{i_{\ell+1} = 1}^{M} \cdots \sum_{i_L = 1}^{M} a_{i_1 \dots i_{\ell-1} i_\ell i_{\ell+1} \dots i_L} = 1.$$

To this end, let $\lambda \in \{1, \dots, L\} \setminus \{\ell\}$ and first consider the sum $S_2 := \sum_{i_\lambda = 1}^{M} a_{i_1 \dots i_L}$. By using the above definition of $a_{i_1 \dots i_L}$, writing down the sum S_2 explicitly yields that all of the $M \cdot 2^L$ addends $D(\cdot, \dots, \cdot)$ of S_2 cancel out except for those 2^L having 0 or 1 in the λ-th component. Since discrete copulas are grounded according to axiom (D1) in Definition 6.1, all the 2^{L-1} terms of S_2 that have a 0 in the λ-th component vanish, and the 2^{L-1} terms

$$S_2 \;=\; M \cdot \left(\Delta_{i_L - 1}^{i_L} \cdots \Delta_{i_{\lambda+1} - 1}^{i_{\lambda+1}} \Delta_{i_{\lambda-1} - 1}^{i_{\lambda-1}} \cdots \Delta_{i_1 - 1}^{i_1} D\left(\frac{k_1}{M}, \dots, \frac{k_{\lambda-1}}{M}, 1, \frac{k_{\lambda+1}}{M}, \dots, \frac{k_L}{M} \right) \right)$$

remain. By writing down the multiple sum S_1 explicitly, iteratively applying the above considerations for the calculation of a sum of the type S_2 and accounting for the fact that discrete copulas are grounded due to axiom (D1) in Definition 6.1, all but two of the terms $D(\cdot, \dots, \cdot)$ of S_1 vanish or cancel out, such that

$$\begin{aligned} S_1 &= M \cdot \left(\Delta_{i_\ell - 1}^{i_\ell} D\left(1, \dots, 1, \frac{k_\ell}{M}, 1, \dots, 1 \right) \right) \\ &= M \cdot \left(D\left(1, \dots, 1, \frac{i_\ell}{M}, 1, \dots, 1 \right) - D\left(1, \dots, 1, \frac{i_\ell - 1}{M}, 1, \dots, 1 \right) \right) \\ &= M \cdot \left(\frac{i_\ell}{M} - \frac{i_\ell - 1}{M} \right) \\ &= 1, \end{aligned}$$

where the axiom (D2) in Definition 6.1 is employed in the third equality. Hence, (A2) is fulfilled.

Thus, A is an L-dimensional stochastic array of order M.

Finally, the definition of A gives the structure of D in (6.1). Indeed, similar arguments as in the proof of (A2) above yield that for fixed $\lambda \in \{1, \dots, L\}$, the sum $S_3 := \sum_{\nu_\lambda = 1}^{i_\lambda} a_{\nu_1 \dots \nu_L}$ can be calculated as

$$S_3 = M \cdot \left(\Delta_{\nu_L - 1}^{\nu_L} \cdots \Delta_{\nu_{\lambda+1} - 1}^{\nu_{\lambda+1}} \Delta_{\nu_{\lambda-1} - 1}^{\nu_{\lambda-1}} \cdots \Delta_{\nu_1 - 1}^{\nu_1} D\left(\frac{k_1}{M}, \dots, \frac{k_{\lambda-1}}{M}, \frac{i_\lambda}{M}, \frac{k_{\lambda+1}}{M}, \dots, \frac{k_L}{M} \right) \right).$$

For fixed $\ell \in \{1,\dots,L\}$ and by using the expression for S_3, similar calculations as in the proof of (A2) before yield

$$S_4 := \sum_{\nu_1=1}^{i_1} \cdots \sum_{\nu_{\ell-1}=1}^{i_{\ell-1}} \sum_{\nu_{\ell+1}=1}^{i_{\ell+1}} \cdots \sum_{\nu_L=1}^{i_L} a_{j_1\dots j_L} = M \cdot \left(\Delta_{\nu_\ell - 1}^{\nu_\ell} D\left(\frac{i_1}{M}, \dots, \frac{i_{\ell-1}}{M}, \frac{k_\ell}{M}, \frac{i_{\ell+1}}{M}, \dots, \frac{i_L}{M} \right) \right).$$

We then employ S_4 to calculate

$$\begin{aligned}
& \frac{1}{M} \sum_{\nu_1=1}^{i_1} \cdots \sum_{\nu_\ell=1}^{i_\ell} \cdots \sum_{\nu_L=1}^{i_L} a_{\nu_1\dots\nu_L} = \frac{1}{M} \sum_{\nu_\ell=1}^{i_\ell} \sum_{\nu_1=1}^{i_1} \cdots \sum_{\nu_L=1}^{i_L} a_{\nu_1\dots\nu_L} \\
= \; & \frac{1}{M} \sum_{\nu_\ell=1}^{i_\ell} M \cdot \left(\Delta_{\nu_\ell - 1}^{\nu_\ell} D\left(\frac{i_1}{M}, \dots, \frac{i_{\ell-1}}{M}, \frac{k_\ell}{M}, \frac{i_{\ell+1}}{M}, \dots, \frac{i_L}{M} \right) \right) \\
= \; & \sum_{\nu_\ell=1}^{i_\ell} \Delta_{\nu_\ell - 1}^{\nu_\ell} D\left(\frac{i_1}{M}, \dots, \frac{i_{\ell-1}}{M}, \frac{k_\ell}{M}, \frac{i_{\ell+1}}{M}, \dots, \frac{i_L}{M} \right) \\
= \; & \sum_{\nu_\ell=1}^{i_\ell} \left[D\left(\frac{i_1}{M}, \dots, \frac{i_{\ell-1}}{M}, \frac{\nu_\ell}{M}, \frac{i_{\ell+1}}{M}, \dots, \frac{i_L}{M} \right) - D\left(\frac{i_1}{M}, \dots, \frac{i_{\ell-1}}{M}, \frac{\nu_\ell - 1}{M}, \frac{i_{\ell+1}}{M}, \dots, \frac{i_L}{M} \right) \right] \\
= \; & D\left(\frac{i_1}{M}, \dots, \frac{i_\ell}{M}, \dots, \frac{i_L}{M} \right) - \underbrace{D\left(\frac{i_1}{M}, \dots, \frac{1-1}{M}, \dots, \frac{i_L}{M} \right)}_{=0, \text{ as } D \text{ is grounded}} \\
= \; & D\left(\frac{i_1}{M}, \dots, \frac{i_L}{M} \right),
\end{aligned}$$

where the second last equality takes advantage of a telescoping sum. Thus, (6.1) holds.

(2) ⇒ (1):
Let the function D be defined as in (6.1). Obviously, D has domain $\mathrm{Dom}(D) = I_M^L$. Since A is an L-dimensional stochastic array of order M, and according to the rules for multiple sums, we have

$$0 \le \frac{1}{M} \sum_{\nu_1=1}^{i_1} \cdots \sum_{\nu_\ell=1}^{i_\ell} \cdots \sum_{\nu_L=1}^{i_L} a_{\nu_1\dots\nu_L} = \frac{1}{M} \sum_{\nu_\ell=1}^{i_\ell} \sum_{\nu_1=1}^{i_1} \cdots \sum_{\nu_L=1}^{i_L} a_{\nu_1\dots\nu_L} \le \frac{1}{M} \sum_{\nu_\ell=1}^{i_\ell} 1 = \frac{1}{M} \cdot i_\ell \le 1$$

for fixed $i_\ell \in \{0,1,\dots,M\}$, $\ell \in \{1,\dots,L\}$, and hence get the range $\mathrm{Ran}(D) = [0,1]$. Moreover, we have to check the axioms (D1), (D2) and (D3) in Definition 6.1 for D.

(D1) Let $i_\ell = 0$ for some $\ell \in \{1,\dots,L\}$. Since the empty sum is equal to zero by convention, we get

$$D\left(\frac{i_1}{M}, \dots, 0, \dots, \frac{i_L}{M} \right) = \frac{1}{M} \sum_{\nu_1=1}^{i_1} \cdots \sum_{\nu_\ell=1}^{0} \cdots \sum_{\nu_L=1}^{i_L} a_{\nu_1\dots\nu_L} = 0.$$

Clearly, this is also the case if there are two or more $\ell \in \{1,\dots,L\}$ such that $i_\ell = 0$. Hence, D is grounded.

(D2) Let $\ell \in \{1,\dots,L\}$. Then,

$$D\left(1, \dots, 1, \frac{i_\ell}{M}, 1, \dots, 1 \right) = D\left(\frac{M}{M}, \dots, \frac{M}{M}, \frac{i_\ell}{M}, \frac{M}{M}, \dots, \frac{M}{M} \right)$$

$$
\begin{aligned}
&= \frac{1}{M} \sum_{\nu_1=1}^{M} \cdots \sum_{\nu_{\ell-1}=1}^{M} \sum_{\nu_\ell=1}^{i_\ell} \sum_{\nu_{\ell+1}=1}^{M} \cdots \sum_{\nu_L=1}^{M} a_{\nu_1 \ldots \nu_L} \\
&= \frac{1}{M} \sum_{\nu_\ell=1}^{i_\ell} \underbrace{\sum_{\nu_1=1}^{M} \cdots \sum_{\nu_{\ell-1}=1}^{M} \sum_{\nu_{\ell+1}=1}^{M} \cdots \sum_{\nu_L=1}^{M} a_{\nu_1 \ldots \nu_L}}_{=1, \text{ as } A \text{ is a stochastic array}} \\
&= \frac{i_\ell}{M},
\end{aligned}
$$

according to the rules for multiple sums.

(D3) We have to show that D is L-increasing, that is,

$$
V := \Delta_{i_L-1}^{i_L} \cdots \Delta_{i_1-1}^{i_1} D\left(\frac{k_1}{M}, \ldots, \frac{k_L}{M}\right) \geq 0.
$$

By definition, V involves 2^L terms of the form $D(\cdot, \ldots, \cdot)$, where 2^{L-1} of them have positive sign and 2^{L-1} negative sign. Moreover, each of the L arguments of a term $D(\cdot, \ldots, \cdot)$ is either of the form $\frac{i_\ell}{M}$ or of the form $\frac{i_\ell - 1}{M}$ for $\ell \in \{1, \ldots, L\}$.

Let $\ell \in \{1, \ldots, L\}$ be fixed. In addition, let the arguments $\frac{k_\lambda}{M}$ for $\lambda \in \{1, \ldots, L\} \setminus \{\ell\}$ also be fixed, that is, k_λ is either equal to i_λ or equal to $i_\lambda - 1$. First,

$$
\begin{aligned}
& D\left(\frac{k_1}{M}, \ldots, \frac{i_\ell}{M}, \ldots, \frac{k_L}{M}\right) - D\left(\frac{k_1}{M}, \ldots, \frac{i_\ell - 1}{M}, \ldots, \frac{k_L}{M}\right) \\
= \; & \frac{1}{M} \sum_{\nu_1=1}^{k_1} \cdots \sum_{\nu_\ell=1}^{i_\ell} \cdots \sum_{\nu_L=1}^{k_L} a_{\nu_1 \ldots \nu_\ell \ldots \nu_L} - \frac{1}{M} \sum_{\nu_1=1}^{k_1} \cdots \sum_{\nu_\ell=1}^{i_\ell - 1} \cdots \sum_{\nu_L=1}^{k_L} a_{\nu_1 \ldots \nu_\ell \ldots \nu_L} \quad (6.2) \\
= \; & \frac{1}{M} \sum_{\nu_1=1}^{k_1} \cdots \sum_{\nu_\ell=i_\ell}^{i_\ell} \cdots \sum_{\nu_L=1}^{k_L} a_{\nu_1 \ldots \nu_\ell \ldots \nu_L} \\
= \; & \frac{1}{M} \sum_{\nu_1=1}^{k_1} \cdots \sum_{\nu_{\ell-1}=1}^{k_{\ell-1}} \sum_{\nu_{\ell+1}=1}^{k_{\ell+1}} \cdots \sum_{\nu_L=1}^{k_L} a_{\nu_1 \ldots \nu_{\ell-1} i_\ell \nu_{\ell+1} \ldots \nu_L},
\end{aligned}
$$

where to some extent, the multiple sum is now reduced due to the fact that the index $\nu_\ell = i_\ell$ is fixed.

By using the definition of V and writing down V explicitly in terms of the 2^L terms $D(\cdot, \ldots, \cdot)$ step by step, we obtain 2^{L-1} such differences as described above within the expression for V. Having calculated all those 2^{L-1} differences in the way as proposed before, we get 2^{L-2} new differences of the same type as in (6.2), where the index $\nu_1 = i_1$ in the multiple sums becomes fixed, and can thus proceed as before. By applying this scheme successively, we finally end up with

$$
V = \frac{1}{M} \sum_{\nu_L=1}^{i_L} a_{i_1 i_2 \ldots i_{L-1} \nu_L} - \frac{1}{M} \sum_{\nu_L=1}^{i_L - 1} a_{i_1 i_2 \ldots i_{L-1} \nu_L} = \frac{1}{M} a_{i_1 i_2 \ldots i_{L-1} i_L}.
$$

Since $A = (a_{i_1 \ldots i_L})_{i_1 \ldots i_L}^{M}$ is an L-dimensional stochastic array of order M by assumption, $a_{i_1 i_2 \ldots i_{L-1} i_L} \geq 0$ for all $i_\ell \in \{1, \ldots, M\}$, $\ell \in \{1, \ldots, L\}$. Hence, $V \geq 0$, and D is L-increasing.

Thus, D is indeed a discrete copula. □

Corollary 6.9. D is an irreducible discrete copula if and only if there is an L-dimensional permutation array $A := (a_{i_1 \ldots i_L})_{i_1,\ldots,i_L=1}^{M}$ such that (6.1) holds for $i_1, \ldots, i_L \in \{0, 1, \ldots, M\}$.

We stress that Theorem 6.8 can also be interpreted as a reformulation of the relation between the CDF and the probability mass function (PMF) (Xu, 1996) because the stochastic array in Definition 6.7 can be identified with M times the PMF.

Essentially, Theorem 6.8 yields the equivalences
Discrete copula
$\Leftrightarrow$ Marginal distributions concentrated on I_M
$\Leftrightarrow$ Probability masses on $\left\{\frac{1}{M}, \frac{2}{M}, \ldots, 1\right\}^L$
$\Leftrightarrow$ Stochastic array.

In the situation of Corollary 6.9, we have the equivalences
Irreducible discrete copula
$\Leftrightarrow$ Empirical copula
$\Leftrightarrow$ M point masses of $\frac{1}{M}$ each
$\Leftrightarrow$ Permutation array
$\Leftrightarrow$ Latin hypercube of order M in L dimensions (Gupta, 1974).

Illustrations of these equivalences are given in Section 6.4, where we discuss their relevance with respect to ECC and the Schaake shuffle, respectively.

Example 6.10.

(a) The discrete product copula $\Pi\left(\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right) := \prod_{\ell=1}^{L} \frac{i_\ell}{M}$ on I_M^L in Example 6.5 (a) corresponds to the L-dimensional stochastic array $A := \left(\frac{1}{M^{L-1}}\right)_{i_1,\ldots,i_L=1}^{M}$ of order M whose entries are all equal to $\frac{1}{M^{L-1}}$. Indeed,

$$\begin{aligned} \frac{1}{M} \sum_{\nu_1=1}^{i_1} \cdots \sum_{\nu_L=1}^{i_L} \frac{1}{M^{L-1}} &= \frac{1}{M} \sum_{\nu_1=1}^{i_1} \cdots \sum_{\nu_{L-1}=1}^{i_{L-1}} \frac{i_L}{M^{L-1}} = \frac{1}{M} \cdot \frac{i_1 \cdot \ldots \cdot i_{L-1} \cdot i_L}{M^{L-1}} \\ &= \prod_{\ell=1}^{L} \frac{i_\ell}{M} = \Pi\left(\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right). \end{aligned}$$

(b) The irreducible discrete copula $\mathcal{M}\left(\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right) := \min\left\{\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right\}$ on I_M^L in Example 6.5 (b) corresponds to the L-dimensional identity stochastic array

$$\mathbb{I} := (a_{i_1 \ldots i_L})_{i_1,\ldots,i_L=1}^{M}, \quad \text{where } a_{i_1 \ldots i_L} := \begin{cases} 1 & \text{if } i_1 = \cdots = i_L \\ 0 & \text{otherwise} \end{cases},$$

of order M. Indeed, employing the definition and writing down the corresponding multiple sum explicitly yields

$$\frac{1}{M} \sum_{\nu_1=1}^{i_1} \cdots \sum_{\nu_L=1}^{i_L} a_{\nu_1 \ldots \nu_L} = \frac{1}{M} \cdot \min\{i_1, \ldots, i_L\} = \min\left\{\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right\}$$

$$= \mathcal{M}\left(\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right).$$

(c) The empirical copula E_M in Example 6.6, which is an irreducible discrete copula, corresponds to the L-dimensional permutation array $A := (a_{i_1 \ldots i_L})_{i_1,\ldots,i_L=1}^{M}$ of order M with

$$a_{i_1 \ldots i_L} := \begin{cases} 1 & \text{if } (x^1_{(i_1)}, \ldots, x^L_{(i_L)}) \in \mathcal{S}, \\ 0 & \text{if } (x^1_{(i_1)}, \ldots, x^L_{(i_L)}) \notin \mathcal{S}, \end{cases}$$

with $\mathcal{S}$ as defined in Example 6.6.

Conversely, for an irreducible discrete copula D with associated L-dimensional permutation array $A := (a_{i_1 \ldots i_L})_{i_1,\ldots,i_L=1}^{M}$ of order M, we consider the sets $\mathcal{X}_1 := \{x^1_1 < \ldots < x^1_M\}, \ldots, \mathcal{X}_L := \{x^L_1 < \ldots < x^L_M\}$. Then, D is the empirical copula of the set $\mathcal{S} := \{(x^1_{i_1}, \ldots, x^L_{i_L}) | a_{i_1 \ldots i_L} = 1\}$.

6.3 A multivariate discrete version of Sklar's theorem

The key result in the context of copulas undoubtably is Sklar's theorem (Sklar, 1959; Nelsen, 2006), as formulated in Theorem 3.3. We now aim at stating and proving a multivariate discrete version thereof.

In the continuous case, an established proof of Sklar's theorem employs an extension lemma, stating that every subcopula can be extended to a copula. The extension lemma in turn is shown via a multivariate interpolation argument (Nelsen, 2006, and references therein). We are guided by this idea and first formulate and prove an extension lemma in a multivariate discrete setting, which provides the main ingredient to showing a multivariate discrete variant of Sklar's theorem. In the proof, which is in some way straightforward, but involves rather tedious calculations, we employ the one-to-one correspondence of discrete copulas to stochastic arrays from Theorem 6.8. A bivariate variant of the discrete extension lemma has been shown by Mayor et al. (2007).

Lemma 6.11. (Extension lemma) For each irreducible discrete subcopula $D^* : J_M^{(1)} \times \cdots \times J_M^{(L)} \to I_M$, there is an irreducible discrete copula $D : I_M^L \to I_M$ such that

$$D|_{J_M^{(1)} \times \cdots \times J_M^{(L)}} = D^*,$$

that is, the restriction of D to $J_M^{(1)} \times \cdots \times J_M^{(L)}$ coincides with D^*.

Proof. Let

$$J_M^{(\ell)} := \left\{ 0 =: \frac{b_0^{(\ell)}}{M} < \frac{b_1^{(\ell)}}{M} < \cdots < \frac{b_{r_\ell}^{(\ell)}}{M} < \frac{b_{r_\ell+1}^{(\ell)}}{M} := 1 \right\}$$

for $\ell \in \{1, \ldots, L\}$, with the corresponding equivalent sets

$$K_M^{(\ell)} := \{0 =: a_0^{(\ell)} < a_1^{(\ell)} < \cdots < a_{r_\ell}^{(\ell)} < a_{r_\ell+1}^{(\ell)} := M\}.$$

According to Theorem 6.8, it suffices to construct an L-dimensional permutation array A of order M to get an irreducible discrete extension copula D of an irreducible discrete subcopula

D^*. The array A has to be such that each block specified by the positions $(a_{s_1}^{(1)}, a_{s_2}^{(2)}, \ldots, a_{s_L}^{(L)})$ and $(a_{s_1+1}^{(1)}, a_{s_2+1}^{(2)}, \ldots, a_{s_L+1}^{(L)})$, which consists of the rows from $a_{s_1}^{(1)}+1$ to $a_{s_1+1}^{(1)}$, from $a_{s_2}^{(2)}+1$ to $a_{s_2+1}^{(2)}$, and so forth, up to the row from $a_{s_L}^{(L)}+1$ to $a_{s_L+1}^{(L)}$, contains a number of 1's equal to the volume

$$M \cdot \left(\Delta_{a_{s_L}^{(L)}}^{a_{s_L+1}^{(L)}} \cdots \Delta_{a_{s_1}^{(1)}}^{a_{s_1+1}^{(1)}} D^* \left(\frac{k_1}{M}, \ldots, \frac{k_L}{M} \right) \right),$$

where $s_\ell \in \{0, \ldots, r_\ell\}$ and $\ell \in \{1, \ldots, L\}$.

To show the existence of such a permutation array A, let $\ell \in \{1, \ldots, L\}$ be fixed and consider the subarray which contains all the blocks determined by the positions $(a_{s_1}^{(1)}, \ldots, a_{s_L}^{(L)})$ and $(a_{s_1+1}^{(1)}, \ldots, a_{s_L+1}^{(L)})$ for all $s_\lambda \in \{0, \ldots, r_\lambda\}$, where $\lambda \in \{1, \ldots, L\} \setminus \{\ell\}$.

We need to show that the number $a_{s_\ell+1}^{(\ell)} - a_{s_\ell}^{(\ell)}$ of rows in this subarray is equal to the number of 1's corresponding to all those blocks. This indeed holds, as

$$\begin{aligned}
&\sum_{s_1=0}^{r_1} \cdots \sum_{s_{\ell-1}=0}^{r_{\ell-1}} \sum_{s_{\ell+1}=0}^{r_{\ell+1}} \cdots \sum_{s_L=0}^{r_L} M \cdot \left(\Delta_{a_{s_L}^{(L)}}^{a_{s_L+1}^{(L)}} \cdots \Delta_{a_{s_1}^{(1)}}^{a_{s_1+1}^{(1)}} D^* \left(\frac{k_1}{M}, \ldots, \frac{k_L}{M} \right) \right) \\
&\quad = M \cdot \sum_{s_1=0}^{r_1} \cdots \sum_{s_{\ell-1}=0}^{r_{\ell-1}} \sum_{s_{\ell+1}=0}^{r_{\ell+1}} \cdots \sum_{s_L=0}^{r_L} \left(\Delta_{a_{s_L}^{(L)}}^{a_{s_L+1}^{(L)}} \cdots \Delta_{a_{s_1}^{(1)}}^{a_{s_1+1}^{(1)}} D^* \left(\frac{k_1}{M}, \ldots, \frac{k_L}{M} \right) \right) \\
&\quad = M \cdot \left(\Delta_{a_{s_\ell}^{(\ell)}}^{a_{s_\ell+1}^{(\ell)}} D^* \left(1, \ldots, 1, \frac{k_\ell}{M}, 1, \ldots, 1 \right) \right) \\
&\quad = M \cdot \left(D^* \left(1, \ldots, 1, \frac{a_{s_\ell+1}^{(\ell)}}{M}, 1, \ldots, 1 \right) - D^* \left(1, \ldots, 1, \frac{a_{s_\ell}^{(\ell)}}{M}, 1, \ldots, 1 \right) \right) \\
&\quad = M \cdot \left(\frac{a_{s_\ell+1}^{(\ell)}}{M} - \frac{a_{s_\ell}^{(\ell)}}{M} \right) \\
&\quad = a_{s_\ell+1}^{(\ell)} - a_{s_\ell}^{(\ell)},
\end{aligned}$$

where we use axiom (S2) in Definition 6.3 for the second last equality.

To see the second equality explicitly, we proceed analogously as in the proof of axiom (A2) in Theorem 6.8, part (1) $\Rightarrow$ (2). We set

$$S := \sum_{s_1=0}^{r_1} \cdots \sum_{s_{\ell-1}=0}^{r_{\ell-1}} \sum_{s_{\ell+1}=0}^{r_{\ell+1}} \cdots \sum_{s_L=0}^{r_L} \left(\Delta_{a_{s_L}^{(L)}}^{a_{s_L+1}^{(L)}} \cdots \Delta_{a_{s_1}^{(1)}}^{a_{s_1+1}^{(1)}} D^* \left(\frac{k_1}{M}, \ldots, \frac{k_L}{M} \right) \right),$$

let $\lambda \in \{1, \ldots, L\} \setminus \{\ell\}$ be fixed and first consider the sum

$$T := \sum_{s_\lambda=0}^{r_\lambda} \left(\Delta_{a_{s_L}^{(L)}}^{a_{s_L+1}^{(L)}} \cdots \Delta_{a_{s_1}^{(1)}}^{a_{s_1+1}^{(1)}} D^* \left(\frac{k_1}{M}, \ldots, \frac{k_L}{M} \right) \right).$$

The $(r_\lambda + 1) \cdot 2^L$ addends $D(\cdot, \ldots, \cdot)$ of T cancel except for those 2^L having a 0 or a 1 in the λ-th component, which indeed occurs as $a_0^{(\lambda)} = 0$ and $a_{r_\lambda+1}^{(\lambda)} = M$. According to axiom (S1) in Definition 6.3, all the 2^{L-1} terms having a 0 in the λ-th component vanish, and we

obtain

$$T = \Delta_{a_{s_L}^{(L)}}^{a_{s_L+1}^{(L)}} \cdots \Delta_{a_{s_{\lambda+1}}^{(\lambda+1)}}^{a_{s_{\lambda+1}+1}^{(\lambda+1)}} \Delta_{a_{s_{\lambda-1}}^{(\lambda-1)}}^{a_{s_{\lambda-1}+1}^{(\lambda-1)}} \cdots \Delta_{a_{s_1}^{(1)}}^{a_{s_1+1}^{(1)}} D^* \left(\frac{k_1}{M}, \ldots, \frac{k_{\lambda-1}}{M}, 1, \frac{k_{\lambda+1}}{M}, \ldots, \frac{k_L}{M} \right).$$

Applying this iteratively and using again axiom (S1) in Definition 6.3, all but two of the terms $D(\cdot, \ldots, \cdot)$ of S vanish or cancel out, such that

$$S = \Delta_{a_{s_\ell}^{(\ell)}}^{a_{s_\ell+1}^{(\ell)}} D^* \left(1, \ldots, 1, \frac{k_\ell}{M}, 1, \ldots, 1 \right),$$

as desired. □

Generally, the extension proposed in Lemma 6.11 is not uniquely determined.

The following definitions of finite distribution functions are standard. Nevertheless, we recapitulate them, as they are needed for the formulation of Sklar's theorem in the multivariate discrete case.

Definition 6.12.

1. A function $F : \overline{\mathbb{R}} \to [0, 1]$ is a finite CDF if F is non-decreasing and right-continuous, and $\mathrm{Ran}(F)$ is a finite set containing $\{0, 1\}$.

2. A function $H : \overline{\mathbb{R}}^L \to [0, 1]$ is a finite L-dimensional CDF if it satisfies the following conditions.

(DF1) H is L-increasing, that is,

$$\Delta_{y_L}^{z_L} \cdots \Delta_{y_1}^{z_1} H(x_1, \ldots, x_L) \geq 0,$$

where

$$\begin{aligned} \Delta_{y_\ell}^{z_\ell} H(x_1, \ldots, x_L) \quad := \quad & H(x_1, \ldots, x_{\ell-1}, z_\ell, x_{\ell+1}, \ldots, x_L) \\ & - H(x_1, \ldots, x_{\ell-1}, y_\ell, x_{\ell+1}, \ldots, x_L) \end{aligned}$$

for all $y_1, \ldots, y_L \in \overline{\mathbb{R}}$ and $z_1, \ldots, z_L \in \overline{\mathbb{R}}$ such that $y_\ell \leq z_\ell$ for all $\ell \in \{1, \ldots, L\}$.

(DF2) H is non-decreasing and right-continuous in each argument.

(DF3) $\mathrm{Ran}(H)$ is a finite set, $H(\infty, \ldots, \infty) = 1$, and H is grounded in the sense that $H(y_1, \ldots, y_L) = 0$ if $y_\ell = -\infty$ for at least one $\ell \in \{1, \ldots, L\}$.

3. For a finite L-dimensional CDF H, the CDFs F_ℓ for $\ell \in \{1, \ldots, L\}$ with $F_\ell(y_\ell) := H(\infty, \ldots, \infty, y_\ell, \infty, \ldots, \infty)$ are the marginal CDFs of H.

It is also well-known that for a random variable $Y : \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$,

$$F_Y(y) := \mathbb{P}(Y \leq y) = \mathbb{P}(\{\omega \in \Omega | Y(\omega) \leq y\})$$

is a CDF corresponding to Y. If Y is a discrete random variable with finite range, then F_Y is a finite CDF. Analogously, for a random vector $\boldsymbol{Y} := (Y_1, \ldots, Y_L)$, the function

$$H_{\boldsymbol{Y}}(y_1, \ldots, y_L) := \mathbb{P}(Y_1 \leq y_1, \ldots, Y_L \leq y_L)$$

is an L-dimensional CDF corresponding to $Y_1, \ldots, Y_L$ with marginal CDFs F_ℓ of Y_ℓ, where $\ell \in \{1, \ldots, L\}$. If $Y_1, \ldots, Y_L$ are discrete random variables with finite ranges, then $H_{\boldsymbol{Y}}$ is a finite L-dimensional CDF.

With this and Lemma 6.11, we are ready to state and prove a multivariate discrete version of Sklar's theorem. For the bivariate case, such a result can be found in Mayor et al. (2007).

Theorem 6.13. (Sklar's theorem in the multivariate discrete case)

1. Let $F_1, \ldots, F_L$ be finite univariate CDFs with $\mathrm{Ran}(F_\ell) \subseteq I_M$ for all $\ell \in \{1, \ldots, L\}$. If D is an irreducible discrete copula on I_M^L, the function

$$H(y_1, \ldots, y_L) := D(F_1(y_1), \ldots, F_L(y_L)) \tag{6.3}$$

   for $y_1, \ldots, y_L \in \overline{\mathbb{R}}$ is a finite L-dimensional CDF with $\mathrm{Ran}(H) \subseteq I_M$, having $F_1, \ldots, F_L$ as marginal CDFs.

2. Conversely, if H is a finite L-dimensional CDF with marginal finite univariate CDFs $F_1, \ldots, F_L$ and $\mathrm{Ran}(H) \subseteq I_M$, there exists an irreducible discrete copula D on I_M^L such that

$$H(y_1, \ldots, y_L) = D(F_1(y_1), \ldots, F_L(y_L))$$

   for $y_1, \ldots, y_L \in \overline{\mathbb{R}}$. Furthermore, D is uniquely determined if $\mathrm{Ran}(F_\ell) = I_M$ for all $\ell \in \{1, \ldots, L\}$.

Proof.

1. This is just a special case of the common Sklar's theorem. The claim follows straightforwardly by checking the axioms of a finite L-dimensional CDF in Definition 6.12 for H as defined in (6.3).

2. Let H be a finite L-dimensional CDF with $\mathrm{Ran}(H) \subseteq I_M$ having univariate marginal CDFs $F_1, \ldots, F_L$. Set

$$J_M^{(\ell)} := \left\{ \frac{i_\ell}{M} \in I_M \,\middle|\, \frac{i_\ell}{M} \in \mathrm{Ran}(F_\ell) \right\} \supseteq \{0, 1\}$$

   for $\ell \in \{1, \ldots, L\}$ and define

$$D^* : J_M^{(1)} \times \cdots \times J_M^{(L)} \to I_M, \;\; D^*\left(\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right) := H(y_1, \ldots, y_L),$$

   where y_ℓ satisfies $F_\ell(y_\ell) = \frac{i_\ell}{M}$ for $\ell \in \{1, \ldots, L\}$.

   We now show that D^* is an irreducible discrete subcopula. First, $\mathrm{Ran}(H) \subseteq I_M$ by assumption, and D^* is well-defined, due to the well-known fact that $H(y_1, \ldots, y_L) = H(z_1, \ldots, z_L)$ for points $y_1, \ldots, y_L \in \overline{\mathbb{R}}$ and $z_1, \ldots, z_L \in \overline{\mathbb{R}}$ such that $F(y_\ell) = F(z_\ell)$ for all $\ell \in \{1, \ldots, L\}$. Furthermore, the axioms (S1), (S2) and (S3) for discrete subcopulas in Definition 6.3 are fulfilled, as shown in what follows.

   (S1) Let $i_\ell = 0$ for an $\ell \in \{1, \ldots, L\}$. Then,

$$D^*\left(\frac{i_1}{M}, \ldots, \frac{i_{\ell-1}}{M}, 0, \frac{i_{\ell+1}}{M}, \ldots, \frac{i_L}{M}\right) = H(y_1, \ldots, y_L)$$

with $F_\ell(y_\ell) = \frac{0}{M} = 0$ and $F_\lambda(y_\lambda) = \frac{i_\lambda}{M}$ for all $\lambda \in \{1, \dots, L\} \setminus \{\ell\}$. However,

$$F_\ell(y_\ell) = H(\infty, \dots, \infty, y_\ell, \infty, \dots, \infty) = 0,$$

and since H is non-decreasing in each argument, we have $H(y_1, \dots, y_\ell, \dots, y_L) = 0$, and hence

$$D^*\left(\frac{i_1}{M}, \dots, \frac{i_{\ell-1}}{M}, 0, \frac{i_{\ell+1}}{M}, \dots, \frac{i_L}{M}\right) = 0$$

for all $\frac{i_\lambda}{M} \in J_M^{(\lambda)}$, where $\lambda \in \{1, \dots, L\} \setminus \{\ell\}$. Clearly, this is also true if $i_\ell = 0$ for two or more $\ell \in \{1, \dots, L\}$.

(S2) For $\ell \in \{1, \dots, L\}$, consider

$$D^*\left(1, \dots, 1, \frac{i_\ell}{M}, 1, \dots, 1\right) = H(y_1, \dots, y_L)$$

with $F_\ell(y_\ell) = \frac{i_\ell}{M}$ and $F_\lambda(y_\lambda) = 1$ for $\lambda \in \{1, \dots, L\} \setminus \{\ell\}$. Set $y_\lambda := \infty$ for $\lambda \in \{1, \dots, L\} \setminus \{\ell\}$. Then,

$$D^*\left(1, \dots, 1, \frac{i_\ell}{M}, 1, \dots, 1\right) = H(\infty, \dots, \infty, y_\ell, \infty, \dots, \infty) = F_\ell(y_\ell) = \frac{i_\ell}{M}$$

for all $\frac{i_\ell}{M} \in J_M^{(\ell)}$.

(S3) To show that D^* is L-increasing, we use the L-increasingness of H as a multivariate CDF and obtain

$$\Delta_{i_L}^{j_L} \cdots \Delta_{i_1}^{j_1} D^*\left(\frac{k_1}{M}, \dots, \frac{k_L}{M}\right) = \Delta_{y_L}^{z_L} \cdots \Delta_{y_1}^{z_1} H(x_1, \dots, x_L) \geq 0$$

for all $\frac{i_\ell}{M}, \frac{j_\ell}{M} \in J_M^{(\ell)}$ such that $\frac{i_\ell}{M} \leq \frac{j_\ell}{M}$, where $F_\ell(y_\ell) = \frac{i_\ell}{M}$ and $F_\ell(z_\ell) = \frac{j_\ell}{M}$ for all $y_\ell, z_\ell \in \overline{\mathbb{R}}$ and $\ell \in \{1, \dots, L\}$. Hence, D^* is L-increasing.

Thus, D^* is indeed a subcopula.

According to Lemma 6.11, D^* can be extended to a discrete copula D, which satisfies

$$D(F_1(y_1), \dots, F_L(y_L)) = D\left(\frac{i_1}{M}, \dots, \frac{i_L}{M}\right) = D^*\left(\frac{i_1}{M}, \dots, \frac{i_L}{M}\right) = H(y_1, \dots, y_L)$$

for $y_1, \dots, y_L \in \overline{\mathbb{R}}$. Hence, $H(y_1, \dots, y_L) = D(F_1(y_1), \dots, F_L(y_L))$.

If $\mathrm{Ran}(F_\ell) = I_M$ for all $\ell \in \{1, \dots, L\}$, then the discrete subcopula D^* has domain I_M^L, and thus we have $D = D^*$, that is, D is uniquely determined. □

Theorem 6.13 is tailored to situations with empirical copulas for data without ties. It is for instance relevant in the context of ECC and the Schaake shuffle, respectively, as will be discussed in Section 6.4.

The statement of Theorem 6.13, part 2, can very likely be extended and accentuated to the effect that the irreducible discrete copula D is uniquely determined if and only if $\mathrm{Ran}(F_\ell) = I_M$ for all $\ell \in \{1, \dots, L\}$. Following the procedure of Mayor et al. (2007) in the bivariate case, a rigorous proof of the new implication in the above equivalence would

require the construction of a smallest and largest discrete extension copula, respectively. The corresponding algorithms to achieve this appear to be supplied by modifying those for the bivariate case in Mayor et al. (2007). However, due to its expected length and technical character, a detailed study of this question is not conducted in this thesis.

Example 6.14. We now illustrate the second part of Theorem 6.13, that is, we give an example for obtaining a discrete copula associated to a finite L-dimensional CDF with given univariate marginal CDFs.

For this purpose, we focus on the case of $L = 3$, $M = 6$ and $I_6 = \left\{0, \frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, 1\right\}$ and assume that we throw a fair die once. Let Y_1 be the random variable that gives the result of the die roll. Moreover, let Y_2 and Y_3 be the $\{0,1\}$- and $\{1,2\}$-valued random variables, respectively, defined by

$$Y_2 := \begin{cases} 0 & \text{if } Y_1 \text{ takes the value 1 ,2, 3 or 4,} \\ 1 & \text{if } Y_1 \text{ takes the value 5 or 6} \end{cases}$$

and

$$Y_3 := \begin{cases} 1 & \text{if } Y_1 \text{ takes an odd value,} \\ 2 & \text{if } Y_1 \text{ takes an even value.} \end{cases}$$

The finite three-dimensional CDF $H_{\boldsymbol{Y}}$ for the random vector $\boldsymbol{Y} := (Y_1, Y_2, Y_3)$ is then given by

$$H_{\boldsymbol{Y}}(y_1, y_2, y_3) = \mathbb{P}(Y_1 \leq y_1, Y_2 \leq y_2, Y_3 \leq y_3),$$

where $y_\ell \in \overline{\mathbb{R}}$, $\ell \in \{1,2,3\}$, and H has the marginal CDFs F_1 of Y_1 with $\mathrm{Ran}(F_1) = I_6$, F_2 of Y_2 with $\mathrm{Ran}(F_2) = \left\{0, \frac{4}{6}, 1\right\}$ and F_3 of Y_3 with $\mathrm{Ran}(F_3) = \left\{0, \frac{3}{6}, 1\right\}$. We have $\mathrm{Ran}(H) = I_6$, and our goal is now to get a discrete copula corresponding to H and F_1, F_2 and F_3. First, we define the discrete subcopula D^* via $D^*\left(\frac{i_1}{6}, \frac{i_2}{6}, \frac{i_3}{6}\right) := H(y_1, y_2, y_3)$, where $y_\ell \in \overline{\mathbb{R}}$, $\ell \in \{1,2,3\}$, are such that $F_1(y_1) = \frac{i_1}{6}$, $F_2(y_2) = \frac{i_2}{6}$ and $F_3(y_3) = \frac{i_3}{6}$. The domain of this discrete subcopula is $J_6^{(1)} \times J_6^{(2)} \times J_6^{(3)}$, where $J_6^{(1)} := I_6$, $J_6^{(2)} := \left\{0, \frac{4}{6}, 1\right\}$ and $J_6^{(3)} := \left\{0, \frac{3}{6}, 1\right\}$. Due to Lemma 6.11, we can extend the discrete subcopula D^* defined on $J_6^{(1)} \times J_6^{(2)} \times J_6^{(3)}$ to a discrete copula D defined on $I_6^3 = I_6 \times I_6 \times I_6$. However, the discrete extension copula D is not uniquely determined, and there are multiple possibilities to complete the missing values of D which are not covered by the values of D^*.

A possible discrete extension copula for our scenario is shown in Table 6.1, in which we give explicit values $D\left(\frac{i_1}{6}, \frac{i_2}{6}, \frac{i_3}{6}\right) = \frac{i}{6}$, where $i_1 \in \{1, \ldots, 6\}$ is fixed for each subtable. For convenience, we present the corresponding values of i instead of $i/6$ in each case. Moreover, we abstain from showing a table for the case of $i_1 = 0$, because $D\left(0, \frac{i_2}{6}, \frac{i_3}{6}\right) = 0$ for all $\frac{i_2}{6} \in I_6$ and $\frac{i_3}{6} \in I_6$, due to axiom (D1) in Definition 6.1. In our setting, the values $D\left(\frac{i_1}{6}, 0, 0\right)$, $D\left(\frac{i_1}{6}, 0, \frac{3}{6}\right)$, $D\left(\frac{i_1}{6}, 0, 1\right)$, $D\left(\frac{i_1}{6}, \frac{4}{6}, 0\right)$, $D\left(\frac{i_1}{6}, \frac{4}{6}, \frac{3}{6}\right)$, $D\left(\frac{i_1}{6}, \frac{4}{6}, 1\right)$, $D\left(\frac{i_1}{6}, 1, 0\right)$, $D\left(\frac{i_1}{6}, 1, \frac{3}{6}\right)$ and $D\left(\frac{i_1}{6}, 1, 1\right)$ for all $\frac{i_1}{6} \in I_6$ are uniquely determined by the corresponding values of the discrete subcopula D^*, indicated by the bold font in Table 6.1, whereas the other values have to be chosen in confirmity with the axioms for discrete copulas.

$i_1 = 1$

i_2 \ i_3	0	1	2	3	4	5	6
0	**0**	0	0	**0**	0	0	**0**
1	0	1	1	1	1	1	1
2	0	1	1	1	1	1	1
3	0	1	1	1	1	1	1
4	**0**	1	1	**1**	1	1	**1**
5	0	1	1	1	1	1	1
6	**0**	1	1	**1**	1	1	**1**

$i_1 = 2$

i_2 \ i_3	0	1	2	3	4	5	6
0	**0**	0	0	**0**	0	0	**0**
1	0	1	1	1	1	1	1
2	0	1	1	1	2	2	2
3	0	1	1	1	2	2	2
4	**0**	1	1	**1**	2	2	**2**
5	0	1	1	1	2	2	2
6	**0**	1	1	**1**	2	2	**2**

$i_1 = 3$

i_2 \ i_3	0	1	2	3	4	5	6
0	**0**	0	0	**0**	0	0	**0**
1	0	1	1	1	1	1	1
2	0	1	2	2	2	2	2
3	0	1	2	2	3	3	3
4	**0**	1	2	**2**	3	3	**3**
5	0	1	2	2	3	3	3
6	**0**	1	2	**2**	3	3	**3**

$i_1 = 4$

i_2 \ i_3	0	1	2	3	4	5	6
0	**0**	0	0	**0**	0	0	**0**
1	0	1	1	1	1	1	1
2	0	1	2	2	2	2	2
3	0	1	2	2	3	3	3
4	**0**	1	2	**2**	3	4	**4**
5	0	1	2	3	4	4	4
6	**0**	1	2	**3**	4	4	**4**

$i_1 = 5$

i_2 \ i_3	0	1	2	3	4	5	6
0	**0**	0	0	**0**	0	0	**0**
1	0	1	1	1	1	1	1
2	0	1	2	2	2	2	2
3	0	1	2	2	3	3	3
4	**0**	1	2	**2**	3	4	**4**
5	0	1	2	3	4	5	5
6	**0**	1	2	**3**	4	5	**5**

$i_1 = 6$

i_2 \ i_3	0	1	2	3	4	5	6
0	**0**	0	0	**0**	0	0	**0**
1	0	1	1	1	1	1	1
2	0	1	2	2	2	2	2
3	0	1	2	2	3	3	3
4	**0**	1	2	**2**	3	4	**4**
5	0	1	2	3	4	5	5
6	**0**	1	2	**3**	4	5	**6**

Table 6.1: Example of a discrete extension copula D for the scenario in Example 6.14: Explicit values $D\left(\frac{i_1}{6}, \frac{i_2}{6}, \frac{i_3}{6}\right) = \frac{i}{6}$, where the corresponding values of i are shown. The bold values are uniquely determined by the corresponding values of the discrete subcopula D^*.

6.4 ECC and the Schaake shuffle as multivariate discrete copula approaches

Now we relate the concepts and results presented beforehand in this chapter to the ECC approach (Schefzik et al., 2013) from Section 4.1 and the Schaake shuffle (Clark et al., 2004) discussed in Section 3.2.2.

(a) Multivariate discrete copulas and ECC

We start with ECC and deepen the theoretical considerations in Section 4.2 in Schefzik et al. (2013). As indicated by its name, ECC has strong connections to copulas, particularly to the notions and results presented before, which is hinted at by Schefzik et al. (2013) and investigated in more detail in what follows.

To this end, let $X_1, \ldots, X_L$ be discrete random variables taking values in $\{x_1^1, \ldots, x_M^1\}, \ldots, \{x_1^L, \ldots, x_M^L\}$, respectively, where $x_1^\ell, \ldots, x_M^\ell$ is the M-member raw ensemble forecast for a multi-index $\ell := (i, j, k)$ pointing at a fixed weather quantity $i \in \{1, \ldots, I\}$, location $j \in \{1, \ldots, J\}$ and look-ahead time $k \in \{1, \ldots, K\}$, and $L := I \times J \times K$. For convenience,

we assume that there are no ties among the corresponding raw ensemble margins. Concerning the multivariate random vector $\boldsymbol{X} := (X_1, \ldots, X_L)$, the corresponding univariate CDFs $R_1, \ldots, R_L$ take values in I_M, that is, $\mathrm{Ran}(R_1) = \cdots = \mathrm{Ran}(R_L) = I_M$. Moreover, we have $\mathrm{Ran}(R) = I_M$ for the multivariate CDF $R : \mathbb{R}^L \to I_M$ of $\boldsymbol{X}$. According to the multivariate discrete version of Sklar's theorem tailored to the ECC framework in Theorem 6.13, there exists a uniquely determined irreducible discrete, and hence a uniquely determined empirical, copula $E_M : I_M^L \to I_M$ such that

$$R(y_1, \ldots, y_L) = E_M(R_1(y_1), \ldots, R_L(y_L)) \tag{6.4}$$

for $y_1, \ldots, y_L \in \overline{\mathbb{R}}$, that is, the multivariate distribution R is connected to its univariate margins $R_1, \ldots, R_L$ via E_M. Conversely, if we take E_M to be the empirical copula defined from the raw ensemble forecast $\{x_1^1, \ldots, x_M^1\}, \ldots, \{x_1^L, \ldots, x_M^L\}$ and $R_1, \ldots, R_L$ to be the univariate CDFs of the raw ensemble margins, then R as constructed in (6.4) is a multivariate CDF.

Following and generalizing the statistical interpretation of discrete copulas for the bivariate case by Mesiar (2005),

$$E_M\left(\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right) = \mathbb{P}(R \in [-\infty, y_1] \times \cdots \times [-\infty, y_L]),$$

where $y_1, \ldots, y_L \in \overline{\mathbb{R}}$ such that $R_1(y_1) = \mathbb{P}(X_1 \leq y_1) = \frac{i_1}{M}, \ldots, R_L(y_L) = \mathbb{P}(X_L \leq y_L) = \frac{i_L}{M}$, that is, $E_M\left(\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right) = \mathbb{P}(X_1 \leq y_1, \ldots, X_L \leq y_L)$. To describe the discrete probability distribution of the random vector $\boldsymbol{X}$, we set $\alpha_{i_1 \ldots i_L} := \mathbb{P}(X_1 = x_{(i_1)}^1, \ldots, X_L = x_{(i_L)}^L)$, where $x_{(i_\ell)}^\ell$ for $i_\ell \in \{1, \ldots, M\}$ and $\ell \in \{1, \ldots, L\}$ denote the corresponding order statistics of the values $X_1, \ldots, X_L$ attain. Then, $\alpha_{i_1 \ldots i_L} \in \left\{0, \frac{1}{M}\right\}$ for all $i_1, \ldots, i_L \in \{1, \ldots, M\}$. Hence, $a_{i_1 \ldots i_L} := M\alpha_{i_1 \ldots i_L} \in \{0, 1\}$ for $i_1, \ldots, i_L \in \{1, \ldots, M\}$, $A := (a_{i_1 \ldots i_L})_{i_1, \ldots, i_L = 1}^M$ is a permutation array of order M, and

$$\frac{1}{M} \sum_{\nu_1 = 1}^{i_1} \cdots \sum_{\nu_L = 1}^{i_L} a_{\nu_1 \ldots \nu_L} = E_M\left(\frac{i_1}{M}, \ldots, \frac{i_L}{M}\right),$$

in accordance with Theorem 6.8.

In the setting of the standard ECC implementation, the above considerations hold analogously for both an individually postprocessed ensemble $\tilde{x}_1^\ell, \ldots, \tilde{x}_M^\ell$ and the ECC ensemble $\hat{x}_1^\ell, \ldots, \hat{x}_M^\ell$. In obvious notation, let $\tilde{F}$ and $\hat{F}$ be the respective multivariate empirical CDFs. Moreover, let $\tilde{F}_1, \ldots, \tilde{F}_L$ denote the marginal empirical CDFs of the individually postprocessed ensemble, with $\tilde{E}_M$ being the corresponding copula. Then,

$$\tilde{F}(y_1, \ldots, y_L) = \tilde{E}_M(\tilde{F}_1(y_1), \ldots, \tilde{F}_L(y_L)) \tag{6.5}$$

and

$$\hat{F}(y_1, \ldots, y_L) = E_M(\tilde{F}_1(y_1), \ldots, \tilde{F}_L(y_L)) \tag{6.6}$$

for $y_1, \ldots, y_L \in \overline{\mathbb{R}}$.

Comparing Equations (6.4), (6.5) and (6.6), the individually postprocessed ensemble and the ECC ensemble have the same marginal distributions, whereas the raw ensemble and the ECC ensemble are associated with the same empirical copula modeling the dependence, due

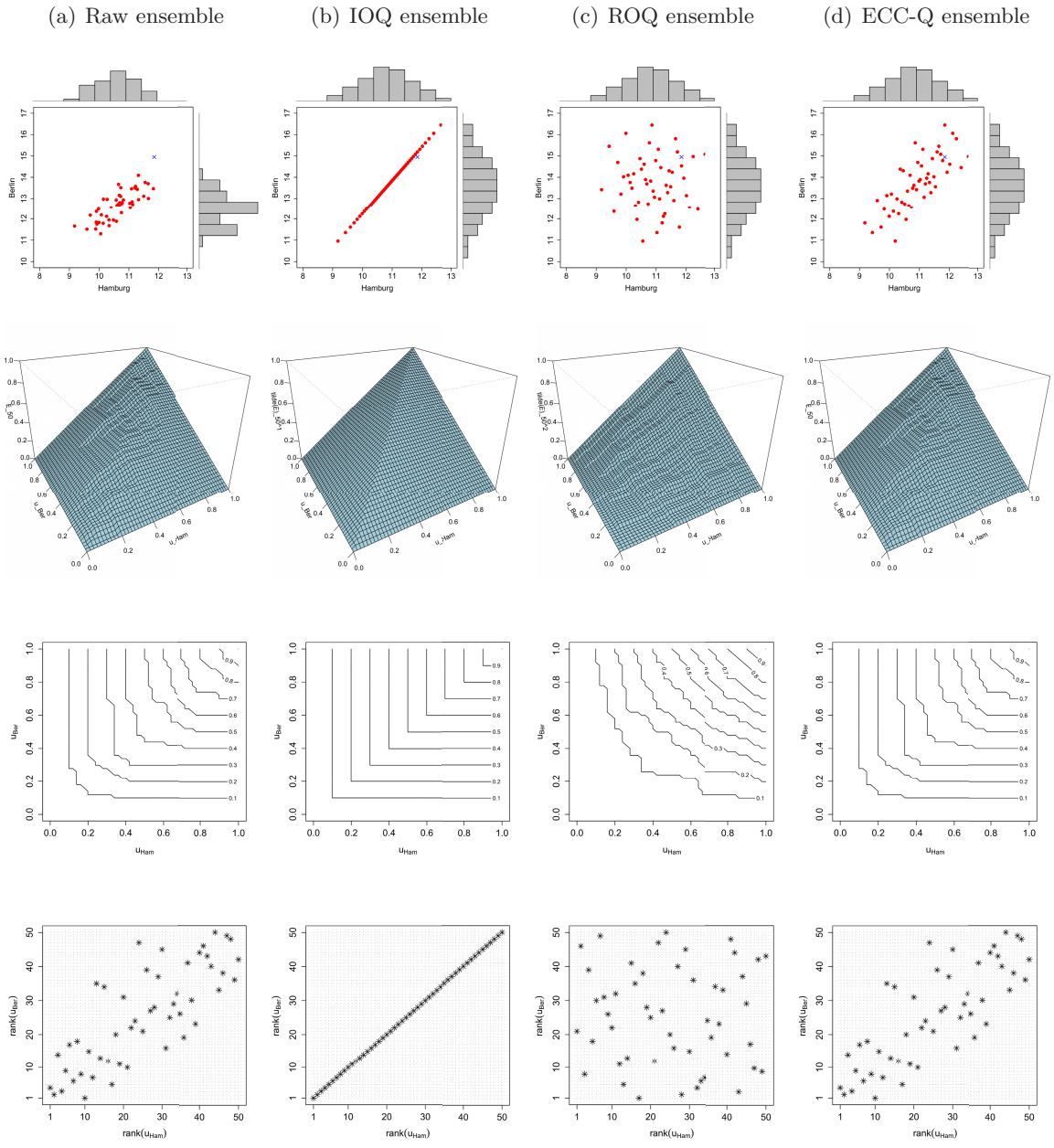


Figure 6.1: Different ensemble prediction approaches comprising the (a) ECMWF raw, (b) IOQ, (c) ROQ and (d) ECC-Q ensemble. First row: Scatterplots with marginal histograms of 24 hour ahead temperature forecasts (in °C) at Berlin (Ber) and Hamburg (Ham), valid 2:00 am on 27 June 2010. The red dots show the respective 50 ensemble member forecasts, and the verifying observation is indicated by the blue cross. Second row: Perspective plots of the corresponding empirical copulas. Third row: Contour plots of the corresponding empirical copulas. Fourth row: Corresponding Latin squares.

to the design of ECC aiming at retaining the rank dependence pattern from the raw ensemble. In particular, the ECC ensemble conserves the bivariate Spearman rank correlation coefficients in the raw ensemble output.

Although our focus has been on the general multivariate case in this chapter, we consider for illustrative purposes a bivariate example, that is, $L = 2$, in Figure 6.1 dealing with 24 hour ahead forecasts for temperature at Berlin and Hamburg, based on the $M = 50$-member ECMWF ensemble and valid 2:00 am on 27 June 2010, thus following the setting of Figure 4.17 in Section 4.3.1. Univariate postprocessing is performed via BMA, using a rolling training period of 30 days. In the left panel of the first row, the unprocessed raw ensemble forecast is shown, revealing a pronounced positive correlation. The plots in the middle of the first row present the IOQ and the ROQ ensemble, respectively, as described in Section 4.3.1. While they both correct for biases and dispersion errors, the IOQ ensemble assumes a maximal possible correlation, whereas the ROQ ensemble essentially provides no correlation structure, in that the bivariate rank order characteristics of the unprocessed forecasts from the left panel are lost. Finally, the postprocessed ECC-Q ensemble in the right panel corrects for biases and dispersion errors as the IOQ and the ROQ ensemble do, but additionally conserves the rank dependence pattern given by the raw ensemble. Thus, although the IOQ, ROQ and ECC-Q ensembles have the same marginal distributions, they differ drastically in their multivariate rank dependence structures. In the second and third row of Figure 6.1, the perspective and stabilized contour plots, respectively, of the empirical copula linked to the different ensembles in our illustrative example are shown. As discussed before, the raw and the ECC-Q ensemble are associated with the same empirical copula E_{50}. On the other hand, the IOQ and the ROQ ensemble are linked to distinct empirical copulas $\tilde{E}^1_{50}$ and $\tilde{E}^2_{50}$, respectively. While the perspective and contour plots of $\tilde{E}^1_{50}$ essentially resemble those of the copula $\mathcal{M}$ from Example 6.5 (b) modeling perfect positive dependence, the plots of $\tilde{E}^2_{50}$ are not far away from those of the independence copula Π introduced in Example 6.5 (a). Hence, the specific design of the respective ensembles is reflected. According to the equivalences discussed in Section 6.2, the raw and the ECC ensembles are also related to the same Latin square, which is a Latin hypercube (Gupta, 1974) in $L = 2$ dimensions, of order $M = 50$, while the IOQ and ROQ ensembles are not, as is illustrated in the fourth row in Figure 6.1.

In a nutshell, ECC indeed can be considered as a copula approach, as it comes up with a postprocessed, discrete L-dimensional distribution, which is according to Theorem 6.13 and Equation (6.6), respectively, constructed from the L empirical CDFs $\tilde{F}_1, \ldots, \tilde{F}_L$ given by the samples drawn from the predictive CDFs $F_1, \ldots, F_L$ obtained by univariate postprocessing on the one hand and the empirical copula E_M induced by the raw ensemble on the other hand. Conversely, each multivariate distribution with fixed univariate margins yields a uniquely determined empirical copula, which defines the rank dependence structure in our setting. Although several multivariate copula-based methods for discrete data have been proposed, for instance by Panagiotelis et al. (2012) using vine and pair copulas, ECC as a discrete copula approach still provides an appropriate and useful alternative to these methods, with benefits as discussed in Section 4.1. As we have seen, the notion of discrete copulas arises naturally in the context of the ECC approach.

The above considerations apply to the standard ECC implementation as introduced in Section 4.1. Starting from those, a next step for further work is to study the theoretical frame

of the ECC modifications presented in this thesis, such as the ECC variant for ensembles consisting of non-exchangeable members from Section 4.5, in whose context new questions concerning the combination of copulas may arise.

(b) Multivariate discrete copulas and the Schaake shuffle

In close analogy to ECC, the Schaake shuffle (Clark et al., 2004) presented in Section 3.2.2 can also be interpreted as a discrete copula approach (Schefzik et al., 2013), with very similar justifications as in part (a). Letting $\ell^* := (i, j)$ denote a multi-index comprising a fixed weather variable $i \in \{1, \dots, I\}$ at a fixed location $j \in \{1, \dots, J\}$, the main difference however is that the corresponding empirical copula O_N in the Schaake shuffle method is defined based on historical verifying observations $y_1^{\ell^*}, \dots, y_N^{\ell^*}$, rather than on raw ensemble forecasts $x_1^\ell, \dots, x_M^\ell$ as in ECC, where N does not need to equal M. Apart from this, the derivations in part (a) hold analogously. In particular, the multivariate empirical CDF $\hat{F}$ associated with the Schaake shuffle ensemble is given by analogy with Equation (6.6), with E_M replaced by O_N, while the empirical marginal CDFs $\tilde{F}_1, \dots, \tilde{F}_L$ are derived based on N samples from the same predictive CDFs $F_1, \dots, F_L$ obtained by univariate postprocessing as employed in the ECC method. Hence, the reordered forecast in the Schaake shuffle ensemble inherits the multivariate rank dependence pattern as well as the pairwise Spearman rank correlation coefficients from the underlying historical weather record, rather than from the raw ensemble as in the ECC approach.

An illustration is given in Figure 6.2, where we consider 24 hour ahead pressure forecasts at Berlin and Hamburg based on the $M = 50$-member ECMWF ensemble, which are valid 1:00 am on 9 March 2011. Univariate postprocessing is performed via BMA, using a sliding training period of 30 days. Since the amount of our available observation data is comparably limited, and for reasons of clarity, we here employ the corresponding verifying observations of the last $N = 30$ days before the verification date of 9 March 2011, that is, of the period from 7 February 2011 to 8 March 2011, as the underlying historical record database. In the original Schaake shuffle approach (Clark et al., 2004), the historical observation database, which the relevant dates close to the verification date in an annual cyclic sense are randomly chosen from, usually comprises several years, if not decades.

Figure 6.2 compares (a) the 50-member unprocessed raw ensemble, (b) the historical observation record, (c) a 30-member ROQ ensemble and (d) the 30-member Schaake shuffle ensemble. In the first row, the scatterplots with the corresponding marginal histograms are shown. The historical observations reveal a strong positive correlation, which is reflected well in the raw ensemble forecast. Hence, it is en passant illustrated that this reasonable assumption for the ECC technique is actually fulfilled. Turning back to the Schaake shuffle, the corresponding ensemble has the same margins as the ROQ ensemble, while conserving the dependence pattern of the historical observation record, which cannot be achieved by the ROQ ensemble completely losing correlation structure. The perspective plots, stabilized contour plots and the Latin squares for the observation record and the different ensembles are shown in the second, third and fourth row, respectively, of Figure 6.2. By construction, the historical observations and the Schaake shuffle ensemble are associated with the same empirical copula O_{30} and Latin square, respectively, whereas the raw ensemble is linked to its empirical copula E_{50}, and the ROQ ensemble to its empirical copula $\tilde{E}_{30}$.

(a) Raw ensemble (b) Historical observations (c) ROQ ensemble (d) Schaake shuffle ensemble

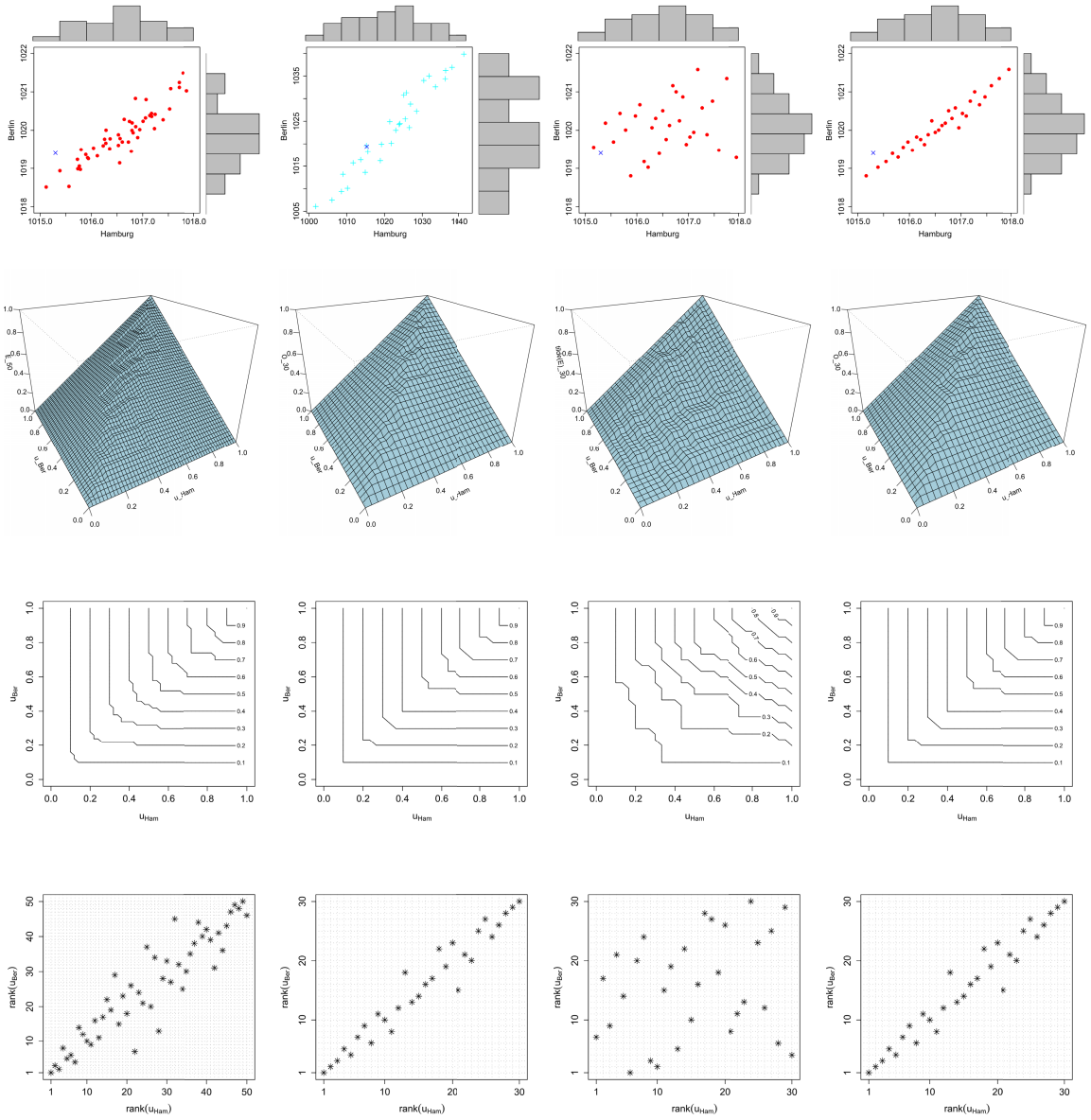


Figure 6.2: Pressure at Berlin (Ber) and Hamburg (Ham) (in hPa): (a) 24 hour ahead ECMWF raw ensemble forecast valid 1:00 am on 9 March 2011, (b) past observations from 7 February 2011 to 8 March 2011, (c) ROQ ensemble forecast valid 1:00 am on 9 March 2011 and (d) Schaake shuffle ensemble valid 1:00 am on 9 March 2011. First row: Scatterplots with marginal histograms. Second row: Perspective plots of the corresponding empirical copulas. Third row: Contour plots of the corresponding empirical copulas. Fourth row: Corresponding Latin squares. The red dots refer to ensemble member forecasts, the blue cross to the verifying observation, and the cyan daggers to the historical observation record.

Chapter 7

Combining low-dimensional postprocessing methods in an ECC-like manner

In the standard ECC approach introduced in Section 4.1, univariate ensemble postprocessing techniques are employed to get calibrated and sharp predictive distributions for each weather quantity, location and look-ahead time separately, and samples thereof are aggregated by using the empirical copula induced by the unprocessed raw ensemble. This leads to physically coherent probabilistic forecasts across space, time and variables.

The approach presented in this chapter aims at combining multiple low-dimensional, that is, not necessarily univariate, postprocessing methods in an ECC-like manner, with the goal of using the specific benefits of all involved techniques. Examples for such low-dimensional methods have been already discussed in Section 3.2.2, including the Gaussian copula approach of Möller et al. (2013) or the bivariate EMOS approach for (u, v)-wind vectors of Schuhen et al. (2012) in inter-variable settings and the spatial BMA and EMOS approaches (Berrocal et al., 2007; Feldmann et al., 2014) for purely spatial settings.

In the context of our new method, the issues of multivariate quantiles and the sampling from a multivariate distribution, respectively, arise and are discussed in Section 7.1. The new approach itself is then described in Section 7.2, and the chapter closes with a case study in Section 7.3.

7.1 Multivariate quantiles

We first discuss the notion of quantiles in multivariate settings, which will be relevant for the new approach.

In the univariate case, the definition and interpretation of an α-quantile of a random variable X, where $\alpha \in (0, 1)$, is well-known from probability theory. The notion of equidistant quantiles is then also straightforward and has already been used in the sampling scheme (Q) of the ECC approach in Section 4.2.

Alternatively, the α-quantile $x_\alpha \in \mathbb{R}$ of a real-valued random variable X with $\mathbb{E}[|X|] < \infty$,

where $\alpha \in (0,1)$, can be characterized based on a minimization notion via

$$x_\alpha := \arg\min_{\xi \in \mathbb{R}} \mathbb{E}[(2\alpha - 1)(X - \xi) + |X - \xi|] \tag{7.1}$$

(Ferguson, 1967; Serfling, 2002b).

According to Abdous and Theodorescu (1992), the above assumption that $\mathbb{E}[|X|] < \infty$ can be circumvented by defining the α-quantile $x_\alpha \in \mathbb{R}$ via

$$x_\alpha := \arg\min_{\xi \in \mathbb{R}} \mathbb{E}\left[\frac{|X - \xi| + (2\alpha - 1)(X - \xi)}{2} - \frac{|X| + (2\alpha - 1)X}{2}\right]. \tag{7.2}$$

The common univariate median is just the (1/2)-quantile.

For a generalized definition of quantiles in the multivariate case, there are various possibilities, with Serfling (2002b) providing an overview. We exemplarily discuss the notions of Abdous and Theodorescu (1992) and Chaudhuri (1996), which extend the univariate definitions in (7.1) and (7.2), respectively, for our purposes here.

Abdous and Theodorescu (1992) first define a norm-like function $||\cdot||_{q,\alpha} : \mathbb{R}^L \to \mathbb{R}$ via

$$||\boldsymbol{x}||_{q,\alpha} := ||(x_1, \ldots, x_L)||_{q,\alpha} := \left|\left|\left(\frac{|x_1| + (2\alpha - 1)x_1}{2}, \ldots, \frac{|x_L| + (2\alpha - 1)x_L}{2}\right)\right|\right|_q,$$

where $\alpha \in (0,1)$, and $||\cdot||_q$ denotes the usual ℓ^q-norm on $\mathbb{R}^L$ for $q \in [1, \infty]$. An α-quantile $\boldsymbol{x}_\alpha \in \mathbb{R}^L$ of an $\mathbb{R}^L$-valued random vector $\boldsymbol{X}$ is then given by

$$\boldsymbol{x}_\alpha := \arg\min_{\boldsymbol{\xi} \in \mathbb{R}^L} \mathbb{E}[||\boldsymbol{X} - \boldsymbol{\xi}||_{q,\alpha} - ||\boldsymbol{X}||_{q,\alpha}].$$

With the above objective function, Abdous and Theodorescu (1992) circumvent the assumption of $\mathbb{E}[||\boldsymbol{X}||] < \infty$, as is shown by Kemperman (1987) for the case of the multivariate median associated with $\alpha = 1/2$.

Chaudhuri (1996) proposes an alternative characterization of multivariate quantiles based on a generalization of (7.1), which is first rewritten as

$$x_u := \arg\min_{\xi \in \mathbb{R}} \mathbb{E}[u(X - \xi) + |X - \xi|], \tag{7.3}$$

where $u := (2\alpha - 1) \in (-1, 1)$. Thus, the univariate α-quantiles $x_\alpha \in \mathbb{R}$ for $\alpha \in (0,1)$ are re-indexed by $u \in (-1,1)$, yielding u-quantiles $x_u \in \mathbb{R}$.

In the multivariate L-dimensional setting, the index set $(-1,1)$ is extended to the open unit ball $\mathbb{B}^{(L)}(\mathbf{0}) := \{\boldsymbol{z} \in \mathbb{R}^L \,|\, ||\boldsymbol{z}|| < 1\}$ with $\mathbf{0} := (0, \ldots, 0)$. Then, a $\boldsymbol{u}$-quantile $\boldsymbol{x_u} \in \mathbb{R}^L$ of an $\mathbb{R}^L$-valued random vector $\boldsymbol{X}$ is defined as

$$\boldsymbol{x_u} := \arg\min_{\boldsymbol{\xi} \in \mathbb{R}^L} \mathbb{E}[\Phi(\boldsymbol{u}, \boldsymbol{X} - \boldsymbol{\xi}) - \Phi(\boldsymbol{u}, \boldsymbol{X})] \tag{7.4}$$

for $\boldsymbol{u} \in \mathbb{B}^{(L)}(\mathbf{0})$, where $\Phi(\boldsymbol{u}, \boldsymbol{t}) := ||\boldsymbol{t}|| + \langle \boldsymbol{u}, \boldsymbol{t} \rangle$ with the usual Euclidean norm $||\cdot||$ and the usual Euclidean inner product $\langle \cdot, \cdot \rangle$. The subtraction term $-\Phi(\boldsymbol{u}, \boldsymbol{X})$ in (7.4) eliminates the

need for the assumption of $\mathbb{E}[||\boldsymbol{X}||] < \infty$, in that $\Phi(\boldsymbol{u}, \boldsymbol{X} - \boldsymbol{\xi}) - \Phi(\boldsymbol{u}, \boldsymbol{X})$ always has a finite expectation, even though the expectation of $\boldsymbol{X}$ may not always be finite. However, if the expectation of $\boldsymbol{X}$ is finite, then $\boldsymbol{x_u}$ becomes a minimizer of $\mathbb{E}[\Phi(\boldsymbol{u}, \boldsymbol{X} - \boldsymbol{\xi})]$ (Chaudhuri, 1996).

A $\boldsymbol{u}$-quantile $\boldsymbol{x_u}$ has both direction and magnitude, its existence is guaranteed for any $\boldsymbol{u} \in \mathbb{B}^{(L)}(\boldsymbol{0})$, and it is unique given that the distribution of $\boldsymbol{X}$ is not supported on a single straight line. Thus, uniqueness is achieved whenever $\boldsymbol{X}$ has an absolutely continuous distribution on $\mathbb{R}^L$ (Chaudhuri, 1996).

Alternatively to the characterization via (7.4), the $\boldsymbol{u}$-quantile $\boldsymbol{x_u}$ may be represented as the solution $\boldsymbol{x_u} := \boldsymbol{\xi}$ of the equation

$$\mathbb{E}\left[\frac{\boldsymbol{X} - \boldsymbol{\xi}}{||\boldsymbol{X} - \boldsymbol{\xi}||}\right] + \boldsymbol{u} = \boldsymbol{0} \tag{7.5}$$

(Chaudhuri, 1996; Serfling, 2002a). If we set $\boldsymbol{u} := \boldsymbol{0}$ as a special case in the above considerations in the sense of Chaudhuri (1996), we obtain the multivariate median $\boldsymbol{x_0}$.

In the context of ensembles, the sample version of the multivariate median as a point forecast has already been discussed in Section 2.3 and used in previous case studies. It is in accordance with the above concepts, where the finite expectiation assumption is fulfilled and the subtraction terms in the above definitions become superfluous. Various algorithms to compute the multivariate median are discussed in Fritz et al. (2012).

As we have seen in this thesis, the use of equally spaced quantiles is useful and appealing at the sampling stage in univariate settings. Thus, it is natural to seek a similar concept for multivariate scenarios. However, an extension of the notion of equidistant quantiles from the univariate to the multivariate case is not obvious. A very first idea would be to employ the concept of equal outlyingness (Serfling, 2002a) as outlined in what follows.

In the univariate case, each point $\xi \in \mathbb{R}$ has a quantile representation $\xi = x_u$ for some choice of $u \in (-1, 1)$. Analogously, each point $\boldsymbol{\xi} \in \mathbb{R}^L$ has a quantile representation $\boldsymbol{\xi} = \boldsymbol{x_u}$ with $\boldsymbol{u} \in \mathbb{B}^{(L)}(\boldsymbol{0})$ in the multivariate case (Serfling, 2002a). From (7.5), it follows that "central" quantiles $\boldsymbol{x_u}$ correspond to indices $\boldsymbol{u}$ with $||\boldsymbol{u}||$ being close to zero, where the multivariate median for $\boldsymbol{u} := \boldsymbol{0}$ with $||\boldsymbol{u}|| = 0$ represents the "most central" quantile. In contrast, "extreme" quantiles are associated with indices $\boldsymbol{u}$ for which $||\boldsymbol{u}||$ is close to one. Hence, the multivariate quantile $\boldsymbol{x_u}$ can be interpreted as indexed by a directional outlyingness parameter $\boldsymbol{u}$, whose magnitude quantifies the outlyingness (Serfling, 2002a). In this spirit, the set of quantiles of equal outlyingness is given by $\{\boldsymbol{x_u} \,|\, ||\boldsymbol{u}|| = c\}$ for fixed $c \in [0, 1)$. As an analog concept to equally spaced quantiles in the univariate case, we can consider multivariate quantiles with equidistant outlyingness when dealing with higher dimensions.

To generate M quantiles $\boldsymbol{x_{u_1}}, \ldots, \boldsymbol{x_{u_M}}$ of equidistant outlyingness as a sample from a multivariate distribution, we first have to create M indices $\boldsymbol{u}_m$, where $m \in \{1, \ldots, M\}$, for which the values of $c_m := ||\boldsymbol{u}_m|| \in [0, 1)$ are equally spaced. We can for instance choose $c_m \in \left\{\frac{m}{M+1} \,\middle|\, m = 1, \ldots, M\right\}$ to obtain this, similar to quantization scheme (Q) from Section 4.2. Alternatively, $c_m \in \left\{\frac{m}{M} \,\middle|\, m = 0, \ldots, M-1\right\}$ can be employed, including the multivariate median for $c_m = 0$, which can be derived as discussed before. In the case of $c_m \in (0, 1)$, we can generate an index $\boldsymbol{u}_m := (u_m^1, \ldots, u_m^L)$ for $m \in \{1, \ldots, M\}$ such that $||\boldsymbol{u}_m|| = c_m$

according to the following steps.

(i) We draw $L-1$ random variates $u_m^1, \ldots, u_m^{L-1}$ from a uniform distribution on $\left(-\frac{c_m}{\sqrt{L-1}}, \frac{c_m}{\sqrt{L-1}}\right)$.

(ii) We consider the equation

$$||\boldsymbol{u}_m|| = \sqrt{(u_m^1)^2 + \ldots + (u_m^{L-1})^2 + (u_m^L)^2} = c_m$$

with $u_m^1, \ldots, u_m^{L-1}$ from step (i) and solve it for u_m^L. Setting $\gamma := (u_m^1)^2 + \cdots + (u_m^{L-1})^2$, elementary algebra yields

$$u_m^L = \pm\sqrt{c_m^2 - \gamma},$$

as $(c_m^2 - \gamma) > 0$ since $\gamma < (L-1) \cdot \left(\frac{c_m}{\sqrt{L-1}}\right)^2 = c_m^2$ due to $u_m^\ell \in \left(-\frac{c_m}{\sqrt{L-1}}, \frac{c_m}{\sqrt{L-1}}\right)$ for $\ell \in \{1, \ldots, L-1\}$. That is, we get two solutions for u_m^L differing only in the sign and can randomly choose the positive or negative one for our procedure.

(iii) From the corresponding values in (i) and (ii), respectively, we build the final index $\boldsymbol{u}_m := (u_m^1, \ldots, u_m^{L-1}, u_m^L)$.

With these indices $\boldsymbol{u}_m$ for $m \in \{1, \ldots, M\}$, the corresponding $\boldsymbol{u}_m$-quantiles $\boldsymbol{x}_{\boldsymbol{u}_m}$ can be derived according to (7.4) or (7.5), respectively. However, we note that the indices $\boldsymbol{u}_m$ obtained via the above construction scheme are not uniquely determined and subject to several random components.

In a nutshell, while the employment of equidistant quantiles has advantages and benefits when it comes to sampling in the univariate case, a generalization of this concept to multivariate settings is not straightforward. It is not entirely clear how to proceed, with the above considerations only providing initial suggestions. Moreover, it is unknown whether such a generalization yields optimal sampling results similar to the univariate case. In addition, the respective implementation might be rather involved, whereas schemes for drawing random samples from multivariate distributions are partly already implemented and available in R. For instance, the R package `MASS` provides the command `mvrnorm()` for randomly sampling from multivariate Gaussian distributions. Hence, we use random multivariate samples for the initial examples in our case study in this chapter. Nevertheless, further resarch into the direction of generalization of equidistance and optimal sampling in the multivariate case is strongly encouraged.

7.2 Combining low-dimensional postprocessing methods in an ECC-like manner: The LDP-ECC approach

We now introduce a new approach which combines Low-Dimensional Postprocessing (LDP) methods on the one hand and the reordering notion of ECC on the other hand. It will thus be referred to as the LDP-ECC approach in what follows.

Let $i \in \{1, \ldots, I\}$ be a weather variable, $j \in \{1, \ldots, J\}$ a location and $k \in \{1, \ldots, K\}$ a look-ahead time. As before, let $\ell := (i, j, k)$ denote the corresponding multi-index, let $L := I \times J \times K$, and let x_m^ℓ be the forecast of raw ensemble member $m \in \{1, \ldots, M\}$ for fixed ℓ.

In addition, let V be the number of low-dimensional postprocessing methods intended to be applied. For each method $\nu \in \{1, \ldots, V\}$, let S_ν denote the number of cases in which the corresponding technique shall be applied. Moreover, let L_ν be the corresponding dimension of the predictive CDF $F_{\nu,s}$ obtained by approach $\nu \in \{1, \ldots, V\}$ in case $s \in \{1, \ldots, S_\nu\}$. We can thus decompose the dimension L via

$$L = \sum_{\nu=1}^{V} S_\nu L_\nu.$$

The choice of the involved low-dimensional postprocessing methods should be as plausible as possible and tailored to the specific setting one is interested in. It might be reasonable to postprocess variables with an expected pronounced correlation jointly, for instance via the methods from Section 3.2.2. At individual stations, (u, v)-wind vectors could be postprocessed simultaneously via the bivariate EMOS approach of Schuhen et al. (2012) or the method of Pinson (2012), for example, while temperature and pressure could be treated jointly via the Gaussian copula approach of Möller et al. (2013). On the other hand, purely spatial settings could be handled by Spatial BMA (Berrocal et al., 2007) or Spatial EMOS (Feldmann et al., 2014), respectively.

The description of the LDP-ECC approach is divided into three parts in what follows.

Part I

1. Let $\ell^{\nu,s,1}, \ldots, \ell^{\nu,s,L_\nu}$ be the corresponding multi-indices of the raw ensemble forecasts that shall be jointly postprocessed via method $\nu \in \{1, \ldots, V\}$ in case $s \in \{1, \ldots, S_\nu\}$. Cases in which one should work with standardized forecasts are explixitly stated in what follows.

   We are given the raw ensemble forecast $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \in \mathbb{R}^L$, with the ensemble members assumed to be exchangeable, as the output of an NWP system, where for an ensemble member $m \in \{1, \ldots, M\}$, we have $\boldsymbol{x}_m := (\boldsymbol{x}_m^1, \boldsymbol{x}_m^2, \ldots, \boldsymbol{x}_m^V)$ with

$$\begin{aligned} \boldsymbol{x}_m^\nu &:= (\boldsymbol{x}_m^{\nu,1}, \boldsymbol{x}_m^{\nu,2}, \ldots, \boldsymbol{x}_m^{\nu,S_\nu}) \\ &:= \Bigg(\underbrace{\left(x_m^{\ell^{\nu,1,1}}, \ldots, x_m^{\ell^{\nu,1,L_\nu}} \right)}_{\text{case } 1}, \underbrace{\left(x_m^{\ell^{\nu,2,1}}, \ldots, x_m^{\ell^{\nu,2,L_\nu}} \right)}_{\text{case } 2}, \ldots, \underbrace{\left(x_m^{\ell^{\nu,S_\nu,1}}, \ldots, x_m^{\ell^{\nu,S_\nu,L_\nu}} \right)}_{\text{case } S_\nu} \Bigg) \end{aligned}$$

   for $\nu \in \{1, \ldots, V\}$.

   Our goal is to come up with a coherent, postprocessed forecast ensemble $\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_M \in \mathbb{R}^L$.

Part II

The following is applied to each method $\nu \in \{1, \ldots, V\}$ and each case $s \in \{1, \ldots, S_\nu\}$ separately.

2. While the common univariate ordering is applied in the standard ECC method, we here first need to define a multivariate ranking structure. To this end, we derive for each method ν and case s a multivariate ranking characteristic $R_m^{\nu,s}$ for each ensemble member $m \in \{1,\ldots,M\}$. For instance, such a characteristic $R_m^{\nu,s}$ based on the raw ensemble forecast $\boldsymbol{x}_1^{\nu,s},\ldots,\boldsymbol{x}_M^{\nu,s}$ can be given by one of the expressions considered in the following subitems (a), (b) and (c), where the corresponding choice has to be consistent for all involved cases.

   (a) We can use the multivariate pre-rank following Gneiting et al. (2008) and set

   $$R_m^{\nu,s} := \sum_{\mu=1}^{M} \mathbb{1}_{\{\boldsymbol{x}_\mu^{\nu,s} \preceq \boldsymbol{x}_m^{\nu,s}\}}, \tag{7.6}$$

   where $\boldsymbol{x}_\mu^{\nu,s} \preceq \boldsymbol{x}_m^{\nu,s}$ if and only if $x_\mu^{\ell^{\nu,s,\lambda}} \leq x_m^{\ell^{\nu,s,\lambda}}$ for all $\lambda \in \{1,\ldots,L_\nu\}$.

   (b) For each univariate margin, that is, for each $\lambda \in \{1,\ldots,L_\nu\}$, we can build the univariate order statistics $x_{(1)}^{\ell^{\nu,s,\lambda}} \leq \ldots \leq x_{(M)}^{\ell^{\nu,s,\lambda}}$, with ties resolved at random, defining the permutations

   $$\sigma_{\ell^{\nu,s,\lambda}}(m) := \operatorname{rank}\left(x_m^{\ell^{\nu,s,\lambda}}\right)$$

   of $\{1,\ldots,M\}$, that is, each member $m \in \{1,\ldots,M\}$ is assigned its rank via $\sigma_{\ell^{\nu,s,\lambda}}$. Then, we can derive the sum of the marginal ranks of the ensemble members for all involved multi-indices and set

   $$R_m^{\nu,s} := \sum_{\lambda=1}^{L_\nu} \sigma_{\ell^{\nu,s,\lambda}}(m). \tag{7.7}$$

   (c) We can use the Euclidean norm and set

   $$R_m^{\nu,s} := ||\boldsymbol{x}_m^{\nu,s}|| = \sqrt{\sum_{\lambda=1}^{L_\nu} \left(x_m^{\ell^{\nu,s,\lambda}}\right)^2}. \tag{7.8}$$

   In this case, it is advisable to use standardized raw ensemble forecast values.

   After the calculation of these M multivariate ranking characteristics $R_m^{\nu,s}$, we compute their order statistics $R_{(1)}^{\nu,s} \leq \ldots \leq R_{(M)}^{\nu,s}$, with ties resolved at random. The corresponding permutation $\tau^{\nu,s}$ of $\{1,\ldots,M\}$ is then given by $\tau^{\nu,s}(m) := \operatorname{rank}(R_m^{\nu,s})$.

3. We apply the low-dimensional postprocessing method ν in the case s to obtain a calibrated and sharp L_ν-variate predictive CDF $F_{\nu,s}$.

4. From the predictive distribution $F_{\nu,s}$, we draw a sample $\tilde{\boldsymbol{x}}_1^{\nu,s},\ldots,\tilde{\boldsymbol{x}}_M^{\nu,s} \in \mathbb{R}^{L_\nu}$ of size M, where

   $$\tilde{\boldsymbol{x}}_m^{\nu,s} := \left(\tilde{x}_m^{\ell^{\nu,s,1}},\ldots,\tilde{x}_m^{\ell^{\nu,s,L_\nu}}\right)$$

   for $m \in \{1,\ldots,M\}$.

5. We reorder the sample $\tilde{\boldsymbol{x}}_1^{\nu,s},\ldots,\tilde{\boldsymbol{x}}_M^{\nu,s}$ from the previous step by using the permutation $\tau^{\nu,s}$ determined in step 2 to get the ensemble $\hat{\boldsymbol{x}}_1^{\nu,s},\ldots,\hat{\boldsymbol{x}}_M^{\nu,s} \in \mathbb{R}^{L_\nu}$, with

   $$\hat{\boldsymbol{x}}_m^{\nu,s} := \left(\hat{x}_m^{\ell^{\nu,s,1}},\ldots,\hat{x}_m^{\ell^{\nu,s,L_\nu}}\right)$$

for $m \in \{1, \ldots, M\}$.

To this end, we derive analogously to step 2 the M multivariate ranking characteristics $\tilde{R}_m^{\nu,s}$, where $m \in \{1, \ldots, M\}$, associated to the sample $\tilde{\boldsymbol{x}}_1^{\nu,s}, \ldots, \tilde{\boldsymbol{x}}_M^{\nu,s}$. This induces the corresponding order statistics $\tilde{R}_{(1)_\sim}^{\nu,s} \leq \ldots \leq \tilde{R}_{(M)_\sim}^{\nu,s}$, with any ties resolved at random, using the notation $(\cdot)_\sim$ to emphasize that the ordering is based on the sample values. The corresponding permutation $\tilde{\tau}^{\nu,s}$ of $\{1, \ldots, M\}$ is then given by $\tilde{\tau}^{\nu,s}(m) := \text{rank}(\tilde{R}_m^{\nu,s})$, and the reordered ensemble $\hat{\boldsymbol{x}}_1^{\nu,s}, \ldots, \hat{\boldsymbol{x}}_M^{\nu,s}$ is finally obtained via

$$\hat{\boldsymbol{x}}_m^{\nu,s} := \tilde{\boldsymbol{x}}_{(\tilde{\tau}^{\nu,s})^{-1}(\tau^{\nu,s}(m))}^{\nu,s}$$

for $m \in \{1, \ldots, M\}$. In other words using order statistic notation, the reordered ensemble $\hat{\boldsymbol{x}}_1^{\nu,s}, \ldots, \hat{\boldsymbol{x}}_M^{\nu,s}$ is given by

$$\hat{\boldsymbol{x}}_m^{\nu,s} := \tilde{\boldsymbol{x}}_{(\tau^{\nu,s}(m))_\sim}^{\nu,s}$$

for $m \in \{1, \ldots, M\}$.

That is, the sample member with the index $(\tilde{\tau}^{\nu,s})^{-1}(\tau^{\nu,s}(m))$ from step 4 forms the member m in the reordered ensemble. Therefore, the multivariate sample members $\tilde{\boldsymbol{x}}_1^{\nu,s}, \ldots, \tilde{\boldsymbol{x}}_M^{\nu,s} \in \mathbb{R}^{L_\nu}$ from step 4 have to stay as they are along the components, and only their order among themselves and hence the indices $m \in \{1, \ldots, M\}$ are changed, according to the multivariate ranking structure of the raw ensemble.

Part III

6. We aggregate all reordered ensembles from step 5 in order to get the final LDP-ECC ensemble $\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_M \in \mathbb{R}^{L_\nu}$. Precisely, we first aggregate all cases $s \in \{1, \ldots, S_\nu\}$ of a fixed method $\nu \in \{1, \ldots, V\}$ for each member $m \in \{1, \ldots, M\}$ by setting

$$\hat{\boldsymbol{x}}_m^\nu := \Bigg(\underbrace{\left(\hat{x}_m^{\ell^{\nu,1,1}}, \ldots, \hat{x}_m^{\ell^{\nu,1,L_\nu}} \right)}_{\text{case } 1}, \underbrace{\left(\hat{x}_m^{\ell^{\nu,2,1}}, \ldots, \hat{x}_m^{\ell^{\nu,2,L_\nu}} \right)}_{\text{case } 2}, \ldots, \underbrace{\left(\hat{x}_m^{\ell^{\nu,S_\nu,1}}, \ldots, \hat{x}_m^{\ell^{\nu,S_\nu,L_\nu}} \right)}_{\text{case } S_\nu} \Bigg).$$

Then, we aggregate all methods ν for each member m via

$$\hat{\boldsymbol{x}}_m := (\hat{\boldsymbol{x}}_m^1, \ldots, \hat{\boldsymbol{x}}_m^V),$$

with the multi-indices $\ell^{\nu,s,\lambda}$, where $\lambda \in \{1, \ldots, L_\nu\}$, in the same order as in the raw ensemble in step 1, and obtain the final LDP-ECC ensemble $\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_M$.

If postprocessing method ν is univariate, such as standard EMOS or BMA, the procedure in Part II is just that of the standard ECC approach. For univariate methods, we have seen that ist is most convenient and advisable to use equally spaced quantiles in the style of ECC-Q in the quantization step 4. In contrast, a generalization of equidistance to the multivariate case and its concrete implementation are not straightforward, as hinted at in the previous section. For our purposes here, we thus suggest to employ random samples from the corresponding multivariate predictive distributions from today's point of view. However, the development of optimal sampling schemes for multivariate settings is an important issue for future work.

With respect to the usual scores used for the evaluation of predictive performances such as the energy score, the reordering in step 5 has no effect and thus becomes irrelevant for scenarios in which only one multivariate method is employed for one case. In these situations, our approach reduces to applying the corresponding method and drawing a sample from the obtained multivariate predictive CDF.

Example 7.1. As a first illustrative example for a setting in which the LDP-ECC approach can be employed, and to get familiar with our notation, assume that we are interested in joint 24 and 48 hour ahead forecasts for temperature, pressure, precipitation and (u, v)-wind vectors jointly at Berlin, Hamburg and Frankfurt simultaneously, made by an ensemble consisting of M members. Hence, we have $I = 5$, $J = 3$ and $K = 2$, thus being confronted with an $L = 5 \times 3 \times 2 = 30$-dimensional problem. As concerns the low-dimensional postprocessing, we can think of employing the Gaussian copula approach of Möller et al. (2013) as method 1 for the inter-variable postprocessing of pressure and temperature jointly, that is, $L_1 = 2$, but individually for each station and prediction horizon, that is, we have to deal with $S_1 = 3 \times 2 = 6$ cases for method 1. The postprocessing of the (u, v)-wind vectors might be performed via the bivariate EMOS approach of Schuhen et al. (2012) as method 2 with dimension $L_2 = 2$, again for each station and prediction horizon separately, that is, $S_2 = 6$. Finally, univariate postprocessing for precipitation could be made by the BMA variant of Sloughter et al. (2007) as method 3 with dimension $L_3 = 1$ for each individual location and look-ahead time, respectively, that is, $S_3 = 6$. In a nutshell, we would then have

$$30 = L = \underbrace{3 \times 2}_{\substack{= 6 = S_1 \\ \text{(3 locations,} \\ \text{2 prediction} \\ \text{horizons)}}} \times \underbrace{2}_{\substack{= L_1 \\ \text{(Gaussian copula} \\ \text{method for pressure} \\ \text{and temperature} \\ \text{jointly)}}} + \underbrace{3 \times 2}_{\substack{= 6 = S_2 \\ \text{(3 locations,} \\ \text{2 prediction} \\ \text{horizons)}}} \times \underbrace{2}_{\substack{= L_2 \\ \text{(bivariate EMOS} \\ \text{for } (u,v)\text{-wind} \\ \text{vectors)}}} + \underbrace{3 \times 2}_{\substack{= 6 = S_3 \\ \text{(3 locations,} \\ \text{2 prediction} \\ \text{horizons)}}} \times \underbrace{1}_{\substack{= L_3 \\ \text{(BMA for} \\ \text{precipitation)}}} .$$

Example 7.2. To further illustrate the LDP-ECC approach in more detail, we test it in the following real-data case study using the $M = 50$-member ECMWF ensemble. We consider 48 hour ahead predictions for (u, v)-wind vectors at Berlin (Ber) and Hamburg (Ham), such that $I = 2$, $J = 2$ and $K = 1$, and thus we face an $L = I \times J \times K = 4$–dimensional scenario with the raw ensemble forecast $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{50} \in \mathbb{R}^4$. The bivariate EMOS approach of Schuhen et al. (2012) is applied to postprocess the (u, v)-wind vectors jointly at the individual stations of Berlin and Hamburg, respectively. Hence, our setting is

$$\boldsymbol{x}_m := (\boldsymbol{x}_m^{\text{Ber 48}}, \boldsymbol{x}_m^{\text{Ham 48}}) := \underbrace{(\underbrace{(x_m^{u,\,\text{Ber 48}}, x_m^{v,\,\text{Ber 48}})}_{\text{case 1 (Ber 48)}}, \underbrace{(x_m^{u,\,\text{Ham 48}}, x_m^{v,\,\text{Ham 48}})}_{\text{case 2 (Ham 48)}})}_{\text{method 1 (bivariate EMOS for } (u,v)\text{-wind)}}$$

for each raw ensemble member $m \in \{1, \ldots, M\}$. We later use the whole $M = 50$-member ensemble in the evaluation of our method over a longer test period. However, for reasons of clarity and comprehensibility, we demonstrate the procedure for only $M = 10$ raw ensemble members by just taking the first 10 members of the 50-member ECMWF ensemble as the underlying raw ensemble in the following illustration.

For the two cases involved in our setting, Table 7.1 shows for each member m the raw ensemble values $\boldsymbol{x}_m$ valid 2:00 am on 15 September 2010, the multivariate characteristics $R_m^{\text{Ber 48}}$ and $R_m^{\text{Ham 48}}$ based on the bivariate pre-ranks according to (7.6), as well as the induced permutations $\tau^{\text{Ber 48}}$ and $\tau^{\text{Ham 48}}$, with ties having been resolved at random, and thus

Table 7.1: Raw ensemble values (in m/s), corresponding multivariate ranking characteristics according to (7.6) and permutations in the setting of Example 7.2, that is, for 48 hour ahead (u, v)-wind vector forecasts at Berlin (Ber) and Hamburg (Ham) separately, valid 2:00 am on 15 September 2010.

	method $\nu = 1$ (bivariate EMOS for (u, v)-wind)							
	case $s_1 = 1$ (Ber 48)				case $s_1 = 2$ (Ham 48)			
m	$x_m^{u,\,\text{Ber 48}}$	$x_m^{v,\,\text{Ber 48}}$	$R_m^{\text{Ber 48}}$	$\tau^{\text{Ber 48}}(m)$	$x_m^{u,\,\text{Ham 48}}$	$x_m^{v,\,\text{Ham 48}}$	$R_m^{\text{Ham 48}}$	$\tau^{\text{Ham 48}}(m)$
1	4.99	6.12	5	9	7.01	1.39	8	10
2	5.60	4.85	3	7	6.87	0.59	2	3
3	5.74	4.78	3	8	6.84	0.85	3	8
4	4.30	5.60	2	5	5.69	0.71	1	2
5	4.52	4.27	1	3	5.86	1.99	2	6
6	5.95	4.99	5	10	9.30	0.12	2	4
7	4.63	4.77	2	4	6.58	1.28	2	7
8	4.26	5.26	1	2	6.62	−0.01	1	1
9	5.25	5.16	3	6	6.91	1.39	6	9
10	6.46	3.51	1	1	6.94	0.52	2	5

Table 7.2: Bivariate samples (in m/s) from $F_{\text{Ber 48}}$ and $F_{\text{Ham 48}}$, respectively, corresponding multivariate ranking characteristics according to (7.6) and permutations for the setting of Example 7.2.

	method $\nu = 1$ (bivariate EMOS for (u, v))							
	case $s_1 = 1$ (Ber 48)				case $s_1 = 2$ (Ham 48)			
m	$\tilde{x}_m^{u,\,\text{Ber 48}}$	$\tilde{x}_m^{v,\,\text{Ber 48}}$	$\tilde{R}_m^{\text{Ber 48}}$	$\tilde{\tau}^{\text{Ber 48}}(m)$	$\tilde{x}_m^{u,\,\text{Ham 48}}$	$\tilde{x}_m^{v,\,\text{Ham 48}}$	$\tilde{R}_m^{\text{Ham 48}}$	$\tilde{\tau}^{\text{Ham 48}}(m)$
1	4.01	3.71	6	9	4.57	1.67	2	6
2	3.52	3.89	3	5	4.80	2.61	3	7
3	4.14	3.15	4	7	5.40	−0.29	2	5
4	3.34	1.73	1	1	7.32	−2.70	1	1
5	4.70	3.69	7	10	5.33	−0.13	2	3
6	6.91	2.74	3	6	5.03	1.60	3	8
7	3.59	3.01	2	2	4.98	0.67	2	4
8	3.70	2.68	2	3	5.79	1.38	5	9
9	3.35	3.16	2	4	4.18	−0.85	1	2
10	3.75	3.52	5	8	6.23	2.47	8	10

Table 7.3: The LDP-ECC ensemble: Reordered bivariate samples (in m/s) from Table 7.2 according to the permutations $\tau^{\text{Ber 48}}$ and $\tau^{\text{Ham 48}}$, respectively. As a check, the corresponding multivariate ranking characteristics and the indices $(\tilde{\tau}^{\text{Ber 48}})^{-1}(\tau^{\text{Ber 48}}(m))$ and $(\tilde{\tau}^{\text{Ham 48}})^{-1}(\tau^{\text{Ham 48}}(m))$, respectively, are given.

	method $\nu = 1$ (bivariate EMOS for (u, v))							
	case $s_1 = 1$ (Ber 48)				case $s_1 = 2$ (Ham 48)			
m	$\hat{x}_m^{u,\text{Ber 48}}$	$\hat{x}_m^{v,\text{Ber 48}}$	$\hat{R}_m^{\text{Ber 48}}$	$(\tilde{\tau}^{\text{Ber 48}})^{-1}$ $(\tau^{\text{Ber 48}}(m))$	$\hat{x}_m^{u,\text{Ham 48}}$	$\hat{x}_m^{v,\text{Ham 48}}$	$\hat{R}_m^{\text{Ham48}}$	$(\tilde{\tau}^{\text{Ham 48}})^{-1}$ $(\tau^{\text{Ham 48}}(m))$
1	4.01	3.71	6	1	6.23	2.47	8	10
2	4.14	3.15	4	3	5.33	−0.13	2	5
3	3.75	3.52	5	10	5.03	1.60	3	6
4	3.52	3.89	3	2	4.18	−0.85	1	9
5	3.70	2.68	2	8	4.57	1.67	2	1
6	4.70	3.69	7	5	4.98	0.67	2	7
7	3.35	3.16	2	9	4.80	2.61	3	2
8	3.59	3.01	2	7	7.32	−2.70	1	4
9	6.91	2.74	3	6	5.79	1.38	5	8
10	3.34	1.73	1	4	5.40	−0.29	2	3

covers steps 1 and 2 of the LDP-ECC approach.

The postprocessing of the (u, v)-wind vector forecasts at Berlin and Hamburg separately according to the EMOS method of Schuhen et al. (2012) yields bivariate predictive CDFs $F_{\text{Ber 48}}$ and $F_{\text{Ham 48}}$, respectively. Using the `mvrnorm()` command in the R package `MASS`, we generate a sample $\tilde{\boldsymbol{x}}_1^{\text{Ber 48}}, \ldots, \tilde{\boldsymbol{x}}_{10}^{\text{Ber 48}}$ from $F_{\text{Ber 48}}$ and a sample $\tilde{\boldsymbol{x}}_1^{\text{Ham 48}}, \ldots, \tilde{\boldsymbol{x}}_{10}^{\text{Ham 48}}$ from $F_{\text{Ham 48}}$, where $\tilde{\boldsymbol{x}}_m^{\text{Ber 48}} := (\tilde{x}_m^{u,\text{Ber 48}}, \tilde{x}_m^{v,\text{Ber 48}})$ and $\tilde{\boldsymbol{x}}_m^{\text{Ham 48}} := (\tilde{x}_m^{u,\text{Ham 48}}, \tilde{x}_m^{v,\text{Ham 48}})$, respectively, for $m \in \{1, \ldots, 10\}$. Table 7.2 lists these samples, together with the corresponding multivariate characteristics $\tilde{R}_m^{\text{Ber 48}}$ and $\tilde{R}_m^{\text{Ham 48}}$ according to (7.6), which are derived analogously as for the raw ensemble, and the permutations $\tilde{\tau}^{\text{Ber 48}}$ and $\tilde{\tau}^{\text{Ham 48}}$. Thus, it comprises steps 3 and 4 of the LDP-ECC method.

The reordered postprocessed ensembles of step 5 for the two cases, as well as the corresponding values for $\hat{R}_m^{\text{Ber 48}}$ and $\hat{R}_m^{\text{Ham 48}}$, again derived analogously as for the raw ensemble, are shown in Table 7.3. As a check, also the indices $(\tilde{\tau}^{\text{Ber 48}})^{-1}(\tau^{\text{Ber 48}}(m))$ and $(\tilde{\tau}^{\text{Ham 48}})^{-1}(\tau^{\text{Ham 48}}(m))$, respectively, indicating which member of the unordered random samples in Table 7.2 builds member m in the reordered ensemble, are listed in Table 7.3. The final LDP-ECC ensemble according to step 6 is then given by

$$\hat{\boldsymbol{x}}_m := (\hat{\boldsymbol{x}}_m^{\text{Ber 48}}, \hat{\boldsymbol{x}}_m^{\text{Ham 48}}) := (\underbrace{(\hat{x}_m^{u,\text{Ber 48}}, \hat{x}_m^{v,\text{Ber 48}})}_{\text{case 1 (Ber 48)}}, \underbrace{(\hat{x}_m^{u,\text{Ham 48}}, \hat{x}_m^{v,\text{Ham 48}})}_{\text{case 2 (Ham 48)}}),$$

with $\hat{\boldsymbol{x}}_m^{\text{Ber 48}} := \tilde{\boldsymbol{x}}_{(\tau^{\text{Ber 48}}(m))_\sim}^{\text{Ber 48}}$ and $\hat{\boldsymbol{x}}_m^{\text{Ham 48}} := \tilde{\boldsymbol{x}}_{(\tau^{\text{Ham 48}}(m))_\sim}^{\text{Ham 48}}$, respectively.

An illustration of our approach based on this example and Tables 7.1 to 7.3 is given in Figure 7.1. For Berlin and Hamburg, respectively, the raw ensemble values of Table 7.1 are shown in the left panel, where each member indicated by a red dot is labeled with its index and its corresponding bivariate pre-rank according to (7.6) in brackets. In the mid-panel, we see the unordered LDP ensemble, that is, the bivariate random samples from Table 7.2, and in the right panel the reordered samples according to Table 7.3, which form the final

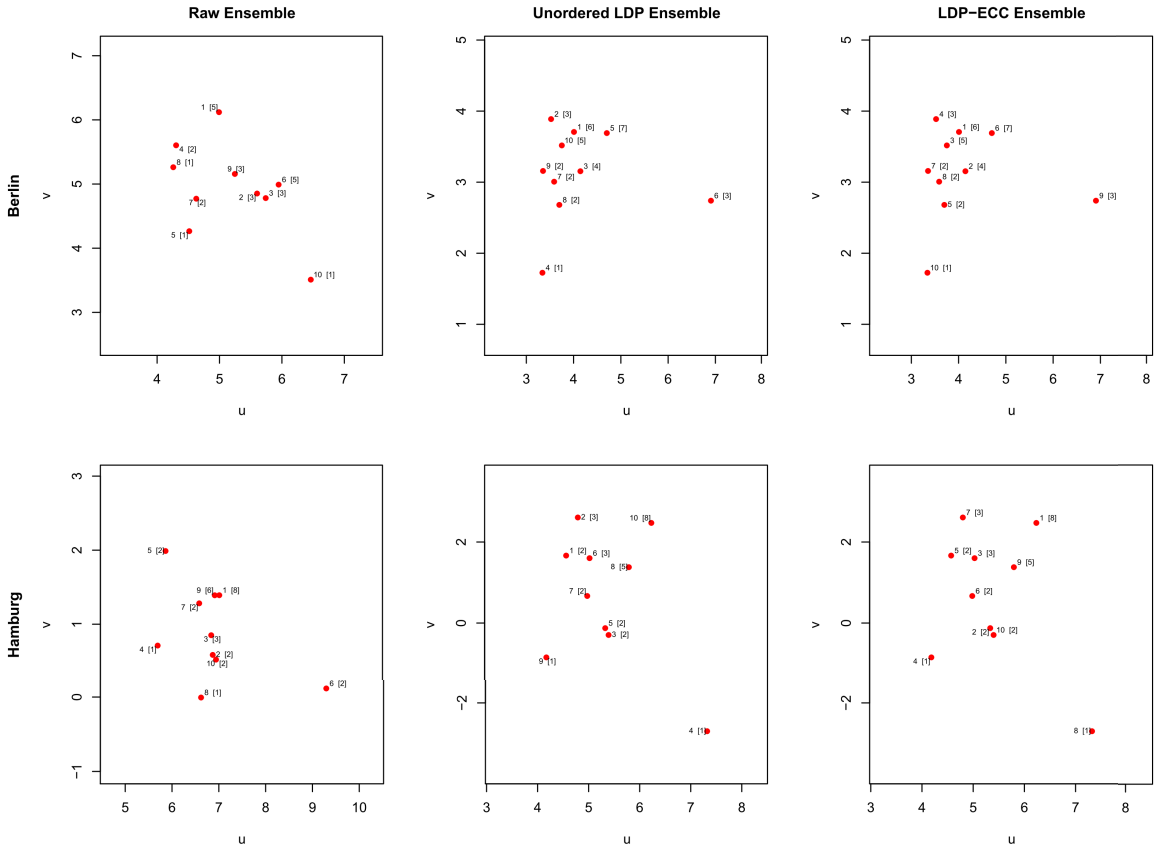


Figure 7.1: Illustration of the LDP-ECC approach based on Example 7.2 and Tables 7.1 to 7.3. Each member indicated by a red dot is labeled with its index and its corresponding bivariate pre-rank according to (7.6) in brackets. For reasons of clarity, the scales on the axes are different.

LDP-ECC ensemble.

The pairs of (u, v)-values and their corresponding bivariate pre-ranks for Berlin and Hamburg, respectively, are exactly the same for both the unordered LDP (middle) and the LDP-ECC (right) ensembles. However, using the LDP-ECC approach, the pairs get other member indices according to the bivariate pre-rank structure in the raw ensemble. That is, a final LDP-ECC ensemble member m for Berlin and Hamburg, respectively, should have a bivariate pre-rank that is near to or comparable to the bivariate pre-rank which the raw ensemble member with the index m has. An exact confirmity of the bivariate pre-rank pattern of the raw and the LDP-ECC ensemble obvoiusly is not always possible, as in the bivariate samples, other bivariate pre-ranks with different frequencies than in the raw ensemble might occur.

However, the LDP-ECC approach tries to conserve the given structure as much as possible. For instance, in our example for Berlin, the raw ensemble members 1 and 6 have both the highest bivariate pre-rank, namely 5, and each of the members 5, 8 and 10 the lowest, namely 1. This is respected quite effectively by LDP-ECC, in that the LDP-ECC ensemble member 6 has the highest bivariate pre-rank, namely 7, and member 1 the second highest, namely 6, while member 10 holds the lowest bivariate pre-rank, namely 1, and members

5 and 8 are two out of three members with the second lowest bivariate pre-rank, namely 2. In contrast, this structure gets lost for the unordered LDP ensemble in Table 7.2 and the mid-panel in Figure 7.1, respectively, where for instance member 6 has the bivariate pre-rank 3, which is by far not one of the highest, whereas for example member 10 has an unsuitably high bivariate pre-rank, namely 5.

Generally, several raw ensemble members might have the same bivariate pre-rank, as is also the case in our example here. Thus, the randomization in the allocation of ties when determining the permutations $\tau^{\text{Ber }48}$ and $\tau^{\text{Ham }48}$, respectively, also has an impact on the form of the final LDP-ECC ensemble. However, this random component can be expected to have an inferior influence when using the alternative multivariate characteristics according to (7.7) or (7.8), respectively, instead of the bivariate pre-rank structure. In these cases, the appearance of ties should be smaller. More sophisticated allocation methods in case of ties tailored to the setting of the LDP-ECC approach could be developed in the future.

The LDP-ECC approach can also be interpreted in terms of copulas, in that for each method ν and each case s, the corresponding empirical copula is defined by the multivariate ranking characteristics $R_m^{\nu,s}$, where $m \in \{1, \dots, M\}$, derived from the respective raw ensemble values. The aggregation of these empirical copulas in a final step raises similar questions concerning the combination of empirical copulas as the ECC variant for ensembles consisting of non-exchangeable members in Section 4.5, as hinted at in Section 6.4.

In the LDP-ECC method, we seek for a multivariate predictive distribution with fixed uni- or multivariate margins. It would be nice if a result similar to Sklar's theorem could be stated, but this is a non-trivial open question for further research. For some special cases, such as for fixed bi- or trivariate margins, theory is already available to some extent (Joe, 1997).

7.3 Case study

We now evaluate the LDP-ECC approach in a case study following the introductory Example 7.2. That is, we consider 48 hour ahead joint (u, v)-wind vector forecasts at Berlin and Hamburg simultaneously, provided by the $M = 50$-member ECMWF ensemble. The (u, v)-wind vector predictions are postprocessed jointly via bivariate EMOS (Schuhen et al., 2012), using a sliding training window of 30 days, individually at each station. In our case study here, we throughout use $\nu = 2$ as period in the bivariate EMOS model for the correlation coefficient ρ in (3.10), which is suggested by empirical considerations. Moreover, we employ the local bivariate EMOS method throughout, in which the parameters are estimated based on training data only from the individual station of interest, thus yielding a distinct set of parameters for each location (Schuhen et al., 2012). Samples from the corresponding bivariate EMOS predictive distributions are drawn randomly, using the `mvrnorm()` command in the R package `MASS`. Our one-year test period ranges from 1 May 2010 to 30 April 2011, and we compare the predictive performances of

- the unprocessed raw ensemble,
- the standard ECC-Q ensemble, where standard EMOS with a rolling training window of 30 days is used for univariate postprocessing at each station and for each wind vector component separately, and the equidistant $\frac{1}{51}$-, $\dots$, $\frac{50}{51}$-quantiles are sampled from each resulting univariate predictive CDF,

Table 7.4: Average ES over the test period from 1 May 2010 to 30 April 2011 for 48 hour ahead joint u- and v-wind ensemble forecasts at Berlin and Hamburg simultaneously. Univariate postprocessing for the ECC-Q and ECC-R ensemble is performed via EMOS, while bivariate postprocessing for the unordered LDP and the LDP-ECC ensemble variants is done via the bivariate EMOS approach for wind vectors of Schuhen et al. (2012). The ES values for the ECC-R, the unordered LDP and the three LDP-ECC ensembles, respectively, are averaged over 100 runs.

	ES
Raw Ensemble	1.869
ECC-Q Ensemble	1.694
ECC-R Ensemble	1.707
Unordered LDP Ensemble	1.698
LDP-ECC Ensemble (a)	1.697
LDP-ECC Ensemble (b)	1.696
LDP-ECC Ensemble (c)	1.698

- the ECC-R ensemble, where as for ECC-Q, standard EMOS is employed for univariate postprocessing, but the samples are drawn randomly from each univariate predictive CDF,

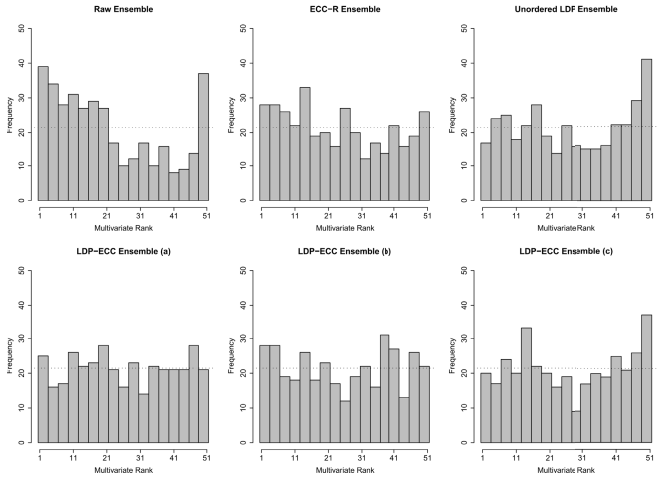
- the unordered LDP ensemble consisting of the unordered bivariate samples obtained from the bivariate EMOS postprocessing, that is, we are in the situation illustrated in the mid-panel of Figure 7.1, for instance, and

- the LPD-ECC ensemble obtained from the bivariate EMOS postprocessing, with variants based on the different multivariate ranking characteristics according to (7.6), (7.7) and (7.8), as described in the subitems (a), (b) and (c), respectively, in the previous section. In what follows, we refer to these variants as LDP-ECC (a), (b) and (c), respectively.

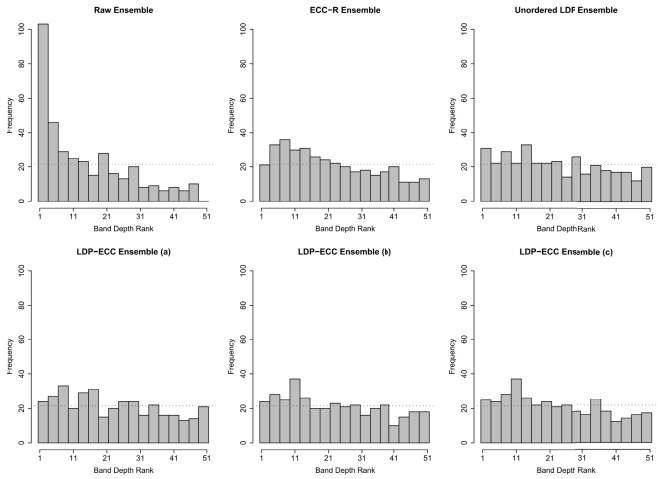
The average energy scores (ES) as overall performance measures for the different methods in our setting are shown in Table 7.4. To account for the random component in the sampling procedure, the scores for the ECC-R, the unordered LDP and the three LDP-ECC ensembles are reported as averages over 100 runs. Calibration is checked via the multivariate, band depth and average rank histograms, respectively, in Figure 7.2, omitting the results for the ECC-Q ensemble for reasons to be discussed soon.

In terms of the ES, each postprocessed ensemble, be it ECC-type or unordered LDP-type or LDP-ECC-type, outperforms the unprocessed raw ensemble, which is clearly uncalibrated.

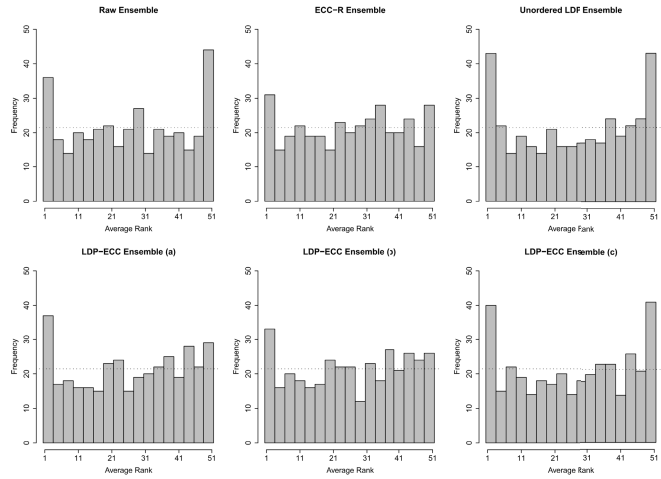
The unordered LDP and the three LDP-ECC ensembles show a better ES than the ECC-R ensemble, but are outperformed by the ECC-Q ensemble. However, it appears to be somewhat unfair to compare the LDP ensembles, which consist of random samples from bivariate distributions, to ECC-Q working with equidistant samples, since the sampling method has an impact on the predictive performance. As we have seen, equidistant sampling typically leads to notably better scores than random sampling. In this regard, it is only reasonable to focus on a comparison of the unordered LDP and the LDP-ECC ensembles to the ECC-R ensemble. With that said, the results for the ECC-Q ensemble are omitted with respect to the rank histograms in Figure 7.2. The LDP ensembles perform a bit better with respect to the ES than the ECC-R ensemble, which moreover is not calibrated that well in terms of the band depth rank histogram.



(a) Multivariate rank histograms



(b) Band depth rank histograms



(c) Average rank histograms

Figure 7.2: (a) Multivariate, (b) band depth and (c) average rank histograms for 48 hour ahead joint u- and v-wind forecasts at Berlin and Hamburg simultaneously, that is, $L = 4$, over the test period from 1 May 2010 to 30 April 2011.

In addition, the unordered LDP ensemble performs a bit worse than the LDP-ECC ensembles (a) and (b). While the differences are minor with regard to the ES, which might be again explained by the findings of Pinson and Tastu (2013), they become obvious in the multivariate and average rank histograms, respectively, considering calibration solely. Hence, a reordering in an ECC-manner proves to be indeed useful and is thus essential for our new approach.

In our case study, the rank-based LDP-ECC methods (a) and (b) perform better than the LDP-ECC approach (c) based on Euclidean norms. Especially in terms of calibration checked via the multivariate and average rank histogram, respectively, the LDP-ECC ensemble (c) appears to be worse than the other two. This is possibly due to a systematic error that might occur when using Euclidean norms in the specific context of (u, v)-wind vector forecasts here. By definition, u- or v-wind is a signed quantity. However, by squaring in the Euclidean norm expression, this property is eliminated, which might lead to failures in the reordering step. Hence, LDP-ECC method (c) can be expected to work better for other, for example positive-valued, variables.

In a nutshell, the LDP-ECC ensemble variants (a) and (b) based on the multivariate characteristics (7.6) and (7.7), respectively, yield good results and turn out to be useful and appealing alternatives to the standard ECC approach. In our setting dealing with (u, v)-wind vectors here, the LDP-ECC ensemble (c) based on Euclidean norms according to (7.8) does not perform that well. However, it is expected to be useful and more appropriate for other weather quantities.

Since the reordering of the bivariate samples is actually only a spatial issue in our case study here, the rather minor differences between the unordered LDP ensemble and the LDP-ECC ensemble variants in the ES might also be explained by the relatively low spatial correlation structure for u- and v-wind, respectively, between the two sites of Berlin and Hamburg. Thus, it would be interesting to conduct a case study on grid-based regions similar to those in Section 4.3.2 (b), where spatial correlations can be expected to be more pronounced.

As hinted at in the previous section, the multivariate ranking characteristics used in the LDP-ECC approach might be subject to a random component due to the random allocation of possible ties, which might have an influence when using the multivariate pre-rank based variant according to (7.6). In such a case, one could repeat the respective case study several times and report average results. This has however not been done in the case study here, as tests showed no necessity to do so, with several initial runs yielding the same ES results to the relevant decimal places.

Chapter 8

Combining similarity-based ensemble methods with the Schaake shuffle

As described in Section 3.2.2, the Schaake shuffle (Clark et al., 2004) generates a postprocessed ensemble inheriting the rank dependence structure from historical observations, and we have shown in Section 6.4 that it can be regarded as an emprical copula-based approach. However, the standard Schaake shuffle fails to condition the multivariate dependence pattern on current or predicted atmospheric states. To address this shortcoming, Clark et al. (2004, page 260) propose to develop an extension of the Schaake shuffle, driven by the idea

> "to preferentially select dates from the historical record that resemble forecasted atmospheric conditions and use the spatial correlation structure from this subset of dates to reconstruct the spatial variability for a specific forecast."

Inspired by this suggestion, we introduce an alternative approach for physically consistent ensemble postprocessing, which essentially fuses the notion of the Schaake shuffle and the idea of similarity- or analog-based ensemble methods, in this chapter.

In similarity- or analog-based ensemble methods, one seeks ensemble forecasts in an archive of past data that are similar to the current one. The basic idea is that the states of the atmosphere corresponding to such an analog ensemble can be assumed to be similar to the state to be predicted (Hamill and Whitaker, 2006). Such methods have been gaining popularity recently, as witnessed by the papers of Bannayan and Hoogenboom (2008), Klausner et al. (2009), Hall et al. (2010), Delle Monache et al. (2011) and Messner and Mayr (2011), among others. In this context, the question of the choice of appropriate similarity measures in a nearest neighbor sense arises, with the papers above providing some proposals.

Our new technique, which will be called the SimSchaake method, is described in Section 8.1 and then tested and assessed in a case study in Section 8.2.

8.1 Combining similarity-based ensemble methods with the Schaake shuffle: The SimSchaake method

The following method combines the idea of similarity-based ensembles and the Schaake shuffle. It is referred to as the SimSchaake approach in what follows. In contrast to standard

ECC, it can be applied to any ensemble, regardless of whether consisting of exchangeable or non-exchangeable members, and the size of the final postprocessed ensemble is not restricted to equal that of the raw ensemble.

Let $i \in \{1, \ldots, I\}$ be a weather variable, $j \in \{1, \ldots, J\}$ a location and $k \in \{1, \ldots, K\}$ a look-ahead time. For simplicity, let $\ell := (i, j, k)$ and $\ell^* := (i, j)$ denote the corresponding multi-indices, and let $L := I \times J \times K$ and $L^* := I \times J$, respectively. Moreover, let M denote the number of raw ensemble members, N the desired number of members in the postprocessed ensemble obtained by the SimSchaake technique, Λ the length of the training period required for the univariate ensemble postprocessing step later on, and t the verification date. Let further D be the number of dates in the past of t for which ensemble forecasts and observation data are available. For the feasibility of the SimSchaake approach, it is required that both ensemble forecast and observation data are available for at least $\max\{N, \Lambda\}$ dates in the past of t, that is, $D \geq \max\{N, \Lambda\}$. In the data set employed in our case study later, we only deal with data valid 00:00 UTC. The SimSchaake approach then proceeds as follows.

1. We are given the possibly standardized raw ensemble forecasts $\boldsymbol{x}^{\ell,t} := (x_1^{\ell,t}, \ldots, x_M^{\ell,t})$ valid on date t for fixed ℓ. Moreover, let

   $$\boldsymbol{x}^t := (\boldsymbol{x}^{1,t}, \boldsymbol{x}^{2,t}, \ldots, \boldsymbol{x}^{L,t}) = (x_1^{1,t}, \ldots, x_M^{1,t}, x_1^{2,t}, \ldots, x_M^{2,t}, \ldots, x_1^{L,t}, \ldots, x_M^{L,t})$$

   denote the $(L \times M)$-tuple consisting of the M-member ensemble forecasts of all L combinations of variable, location and prediction horizon.

   If weather variables with distinct units or magnitudes are involved, the components of $\boldsymbol{x}^{\ell,t}$ should be standardized for each ℓ. This can be for instance done as described in Section 4.3.3.

2. For each date t_d in the past of t, where $d \in \{1, \ldots, D\}$, we compute a suitable fixed similarity measure $\Delta^{t_d} := \Delta^{t_d}(\boldsymbol{x}^t, \boldsymbol{x}^{t_d}) \in \mathbb{R}_+^0$ between $\boldsymbol{x}^t$ and $\boldsymbol{x}^{t_d}$, which is taken to be negatively oriented, that is, the lower the similarity measure the more similar the ensemble forecasts.

3. Choose those N dates $\tau_1, \ldots, \tau_N \in \{t_d \,|\, d \in \{1, \ldots, D\}\}$ among the D dates $t_1, \ldots, t_D$ in the past of t for which the data is most similar to that for the date t in the sense that the corresponding values of Δ^{τ_n} for $n \in \{1, \ldots, N\}$ are the smallest among the values of Δ^{t_d} for $d \in \{1, \ldots, D\}$.

   Thereby, let τ_1 be the date corresponding to the "nearest" ensemble relative to that of date t, and so forth, and τ_N be the date corresponding to the "furthermost" ensemble relative to that of date t with respect to the respective similarity measure.

4. For each ℓ^*, let $y^{\ell^*,\tau_1}, \ldots, y^{\ell^*,\tau_N}$ denote the corresponding N historical verifying observations valid on the dates $\tau_1, \ldots, \tau_N$ determined in step 3. With those, we build the data vector $\boldsymbol{y}^{\ell^*,t} := (y_1^{\ell^*,t}, \ldots, y_N^{\ell^*,t})$ assigned to the verification date t of interest by setting

   $$y_1^{\ell^*,t} := y^{\ell^*,\tau_1}, \ldots, y_N^{\ell^*,t} := y^{\ell^*,\tau_N}.$$

   For each ℓ^*, the corresponding order statistics $y_{(1)}^{\ell^*,t} \leq \ldots \leq y_{(N)}^{\ell^*,t}$ induce a permutation π_{ℓ^*} of $\{1, \ldots, N\}$ via $\pi_{\ell^*}(n) := \mathrm{rank}(y_n^{\ell^*,t})$, where $n \in \{1, \ldots, N\}$, with ties resolved at random.

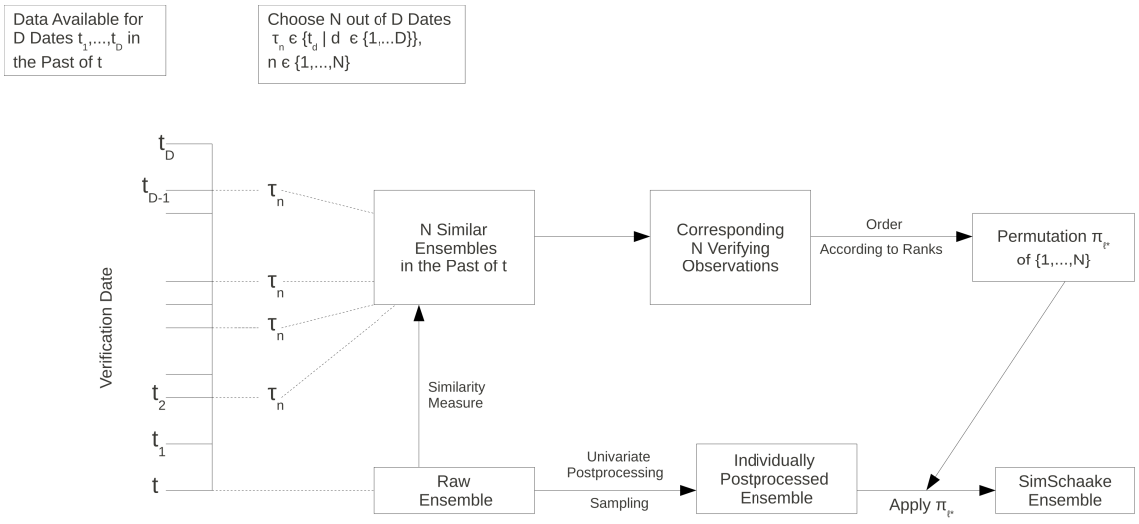


Figure 8.1: Scheme of the SimSchaake approach

5. For each ℓ separately, we apply state-of-the-art univariate methods to postprocess the raw ensemble $\boldsymbol{x}^{\ell,t}$, such as BMA or EMOS, and obtain a predictive CDF F_ℓ^t valid on date t in each case. Then, we draw a sample $\tilde{x}_1^{\ell,t}, \ldots, \tilde{x}_N^{\ell,t}$ of size N from each obtained F_ℓ^t. As usual, we prefer quantization scheme (Q), in which the sample is given by

$$\tilde{x}_1^{\ell,t} := (F_\ell^t)^{-1}\left(\frac{1}{N+1}\right), \ldots, \tilde{x}_N^{\ell,t} := (F_\ell^t)^{-1}\left(\frac{N}{N+1}\right),$$

   forming the postprocessed ensemble $\tilde{\boldsymbol{x}}^{\ell,t} := (\tilde{x}_1^{\ell,t}, \ldots, \tilde{x}_N^{\ell,t})$.

6. For each ℓ, the final SimSchaake ensemble is then given by $\hat{\boldsymbol{x}}^{\ell,t} := (\hat{x}_1^{\ell,t}, \ldots, \hat{x}_N^{\ell,t})$, where

$$\hat{x}_1^{\ell,t} := \tilde{x}_{(\pi_{\ell^*}(1))}^{\ell,t}, \ldots, \hat{x}_N^{\ell,t} := \tilde{x}_{(\pi_{\ell^*}(N))}^{\ell,t},$$

   that is, the sample from step 5 is reordered with respect to the corresponding observation ranks of $\boldsymbol{y}^{\ell^*,t}$ derived in step 4.

Steps 1 to 3 employ information of all ℓ simultaneously to determine the dates $t_1, \ldots, t_N$. However, steps 4 to 6 are applied to each ℓ and ℓ^*, respectively, individually.

A schematic diagram of the SimSchaake approach is given in Figure 8.1.

An appropriate choice of the similarity measure Δ^{t_d} in step 2, which is consistently used throughout the whole SimSchaake approach, is crucial. For example, we can take

$$\Delta_1^{t_d} := \sqrt{\frac{1}{L}\sum_{\ell=1}^{L}(\bar{x}^{\ell,t} - \bar{x}^{\ell,t_d})^2 + \frac{1}{L}\sum_{\ell=1}^{L}(s^{\ell,t} - s^{\ell,t_d})^2}, \tag{8.1}$$

where

$$\bar{x}^{\ell,\tau} := \frac{1}{M}\sum_{m=1}^{M} x_m^{\ell,\tau} \quad \text{and} \quad s^{\ell,\tau} := \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(x_m^{\ell,\tau} - \bar{x}^{\ell,\tau})^2}$$

denote the empirical mean and standard deviation, respectively, of the ensemble forecast $\boldsymbol{x}^{\ell,\tau}$ for fixed indices ℓ and τ. As it does not depend on how the ensemble members are labeled, the similarity measure $\Delta_1^{t_d}$ can be applied to ensembles consisting of exchangeable members and is thus suitable for the ECMWF forecasts used in the case study in the next section. Alternative proposals for similarity measures, also for the case of ensembles with non-exchangeable members, can be found in the references mentioned at the beginning of this chapter.

The similarity measure $\Delta_1^{t_d}$ can be modified by multiplying it with a weighting function w_{t_d} as a chronological term. By proceeding so, we get

$$\Delta_2^{t_d} := w_{t_d} \cdot \Delta_1^{t_d}, \tag{8.2}$$

and there are various possibilities to design the time-weighting function w_{t_d}. One variant tailored to our specific data set at hand is to employ a linear weighting scheme, in that

$$w_{t_d} := \frac{\delta_d}{\sum\limits_{d=1}^{D} \delta_d}, \tag{8.3}$$

where $\delta_d \in \mathbb{N}$ denotes the temporal distance in days between the dates t and t_d. That is, the closer the past date t_d lies to the date of interest t the lower is the corresponding weight.

However, scheme (8.3) does not take account of seasonality. Alternatively, we can address this for instance by setting

$$w_{t_d} := \frac{A - \cos\left(\frac{2\pi \cdot \delta_d}{365}\right)}{\sum\limits_{d=1}^{D} \left(A - \cos\left(\frac{2\pi \cdot \delta_d}{365}\right)\right)}, \tag{8.4}$$

where $\delta_d \in \mathbb{N}$ again denotes the temporal distance in days between the dates t and t_d, and $A \geq 1$ is a constant. In scheme (8.4), we account for seasonality by assigning low weights to past dates t_d that are close to t in an annual cyclical sense, and high weights to past dates t_d that are far away from t in an annual cyclical sense.

In both schemes (8.3) and (8.4), the weights are non-negative and add up to one. Concerning variant (8.4), the choice of the constant A has to fulfill $A \geq 1$ to ensure the non-negativity of the weights. The specific value assigned to A might depend on the data set at hand, as is discussed with respect to our case study in the next section, where we consider $A = 5$.

Alternatively, monotone transformations of the weights w_{t_d}, such as square root transformations, could also be used to define a modified similarity measure.

It is possible to transform the values of the similarity measures $\Delta_\nu^{t_d}$ for $\nu \in \{1, 2\}$ from $\mathbb{R}_+^0$ to the unit interval $(0, 1]$ by employing the standardization ${}^*\Delta_\nu^{t_d} := \exp(-\Delta_\nu^{t_d})$. With respect to ${}^*\Delta_\nu^{t_d}$, similarity values near to 1 indicate a very high similarity between $\boldsymbol{x}^t$ and $\boldsymbol{x}^{t_d}$, while similarity values near to 0 point at no similarity. Accordingly, if using ${}^*\Delta_\nu^{t_d}$, we then have to choose the dates $\tau_1, \ldots, \tau_N$ corresponding to the highest, and not to the lowest, values of ${}^*\Delta_\nu^{t_d}$ in step 3.

Concerning step 2, we alternatively could search for similar ensembles not among all available dates, but only for the Υ most recent days before t, where $N \leq \Upsilon \leq D$, in a sort of sliding searching window. Then, it is required that both forecast and observation data is available for at least $\max\{\Upsilon, \Lambda\}$ dates in the past of t, where it remains to specify how to choose Υ appropriately.

In step 3, the dates $\tau_1, \ldots, \tau_N$ could alternatively be assigned either chronologically, with τ_1 being the date closest to t, or randomly as in the original Schaake shuffle approach. Exemplary initial case studies with temperature showed that there is no significant impact of the assigning method on the final predictive performance. Hence, we set the procedure described in step 3 as the standard in what follows.

As mentioned, the SimSchaake approach essentially solves two shortcomings appearing with respect to the standard ECC method. First, our new method can also be applied to ensembles consisting of non-exchangeable members, as the reordering is not based on the ensemble forecasts, but on the verifying observations. Second, with our new technique we can principally create ensembles of arbitrary size, as long as there are sufficiently many historical observations in the past. In contrast to ECC, the postprocessed ensemble is thus not restricted to have the same number of members as the raw ensemble, that is, N need not necessarily be equal to M.

As ECC and the Schaake shuffle, the SimSchaake approach can also be interpreted as a discrete copula-based technique, where the relevant empirical copula is derived from the historical observation database in step 4.

8.2 Case study

The SimSchaake approach is tested and evaluated by considering 24 hour ahead forecasts for temperature, pressure, u- and v-wind, respectively, at Berlin, Hamburg and Frankfurt jointly, provided by the $M = 50$-member ECMWF ensemble. Our one-year test period ranges from 1 May 2010 to 30 April 2011. For temperature, we additionally consider the two-year test period from 1 January 2011 to 31 December 2012, as more data is available in this case. Univariate postprocessing is performed via EMOS, using a training period of $\Lambda = 30$ days. We assess and compare the predictive performance of the unprocessed raw, the standard ECC-Q, the random selection and two different SimSchaake ensembles. For the latter three approaches, the past dates from which the corresponding verifying observations are taken, are searched for among all available historical data, where ensemble forecast and observation data is available from 2 February 2010 to 30 April 2011, and in the case of temperature to 31 December 2012. In this context, the random selection ensemble just randomly selects those past dates, whereas the SimSchaake ensemble variants choose them based on the ensemble similarity measures (8.1) and (8.2), respectively. Concerning the time-weighted variant, we use the weighting scheme (8.4) accounting for seasonality in our case study, with $A = 5$. In this context, first studies suggested that $A = 5$ is a suitable choice for the constant A in (8.4) in our setting. While the above choice for A cannot be expected to be optimal in any sense, it has proven to be a reasonable compromise in initial tests.

In Table 8.1, the average energy scores (ES) as overall performance measures are shown,

Table 8.1: Average ES for 24 hour ahead forecasts at Berlin, Hamburg and Frankfurt jointly over the test period from 1 May 2010 to 30 April 2011 (temperature, pressure, u- and v-wind) and from 1 January 2011 to 31 December 2012 (temperature), respectively. The results for the random selection ensemble are averaged over 100 runs.

	1-year period				2-year period
	Temperature (°C)	Pressure (hPa)	u-Wind (m/s)	v-Wind (m/s)	Temperature (°C)
Raw Ensemble	2.278	1.005	1.740	1.839	1.938
ECC-Q Ensemble	1.720	0.829	1.334	1.578	1.548
$N = M = 50$					
Random Selection Ensemble	1.775	0.836	1.345	1.581	1.604
SimSchaake Ensemble	1.728	0.829	1.332	1.576	1.548
Time-Weighted SimSchaake Ensemble	1.728	0.829	1.332	1.576	1.548
$N = 70$					
Random Selection Ensemble	1.773	0.836	1.341	1.578	1.602
SimSchaake Ensemble	1.729	0.829	1.329	1.573	1.546
Time-Weighted SimSchaake Ensemble	1.730	0.828	1.329	1.573	1.546

where the results for the random selection ensemble are averaged over 100 runs. While the raw and the ECC-Q ensemble are constrained to consist of $N = M = 50$ members, the random selection ensemble and the SimSchaake ensemble variants are evaluated for the final ensemble size of both $N = M = 50$ and $N = 70$, respectively. Calibration solely for $N = M = 50$ is assessed via the multivariate, band depth and average rank histograms, respectively, in Figures 8.2 to 8.5, where Figure 8.2 shows results for temperature over the two-year test period, while Figures 8.3 to 8.5 show the histograms for pressure, u- and v-wind, respectively, over the one-year test period.

We first comment on the case of $N = M = 50$, where all postprocessing methods outperform the raw ensemble in terms of calibration and the ES for each weather variable.

There is a benefit to use the SimSchaake ensembles based on a similarity measure instead of the random selection ensemble, as they perform better with respect to calibration and the ES. Except maybe for calibration in the case of pressure, the SimSchaake ensembles can compete with the ECC-Q ensemble, in that they yield similar results and even slightly outperform ECC-Q for u- and v-wind, respectively, concerning the ES.

Furthermore, the differences between two SimSchaake variants are rather minor in terms of the rank histograms and even non-existent with respect to the ES. This might be not that surprising, as the data archive the past dates can be chosen from is very small, which reveals a general shortcoming of this case study to be discussed at the end of this section, such that seasonal issues cannot be expected to have a great impact.

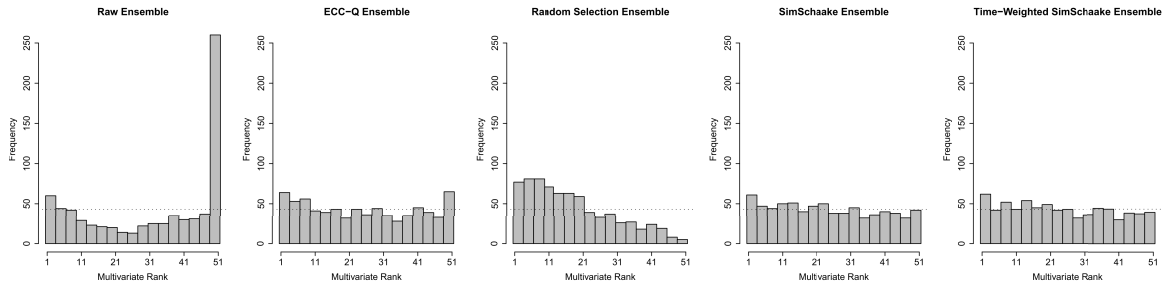
In the case of SimSchaake ensembles comprising $N = 70$ members, the results and conclusions described above essentially continue to hold analogously, with the respective ES values mostly being equally good as or slightly better than those in the case of $N = M = 50$,

except for temperature over the one-year test period. Actually there is no absolute need to extend the ensemble size of $M = 50$ in our case here, although an extension is nonetheless meaningful. The improvements in the ES obtained by the extension from $N = M = 50$ to $N = 70$ are not that pronounced, which is to be expected, as the ensemble size of $M = 50$ already appears to be sufficiently large. However, an extension of the number of members in the SimSchaake ensemble likely has more impact when the raw ensemble is rather small, as is suggested by the corresponding considerations and examinations in the context of the extended ECC ensembles in Section 4.4.

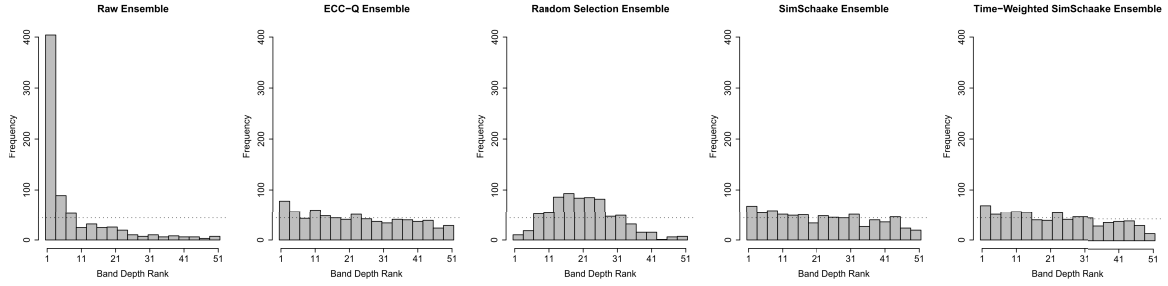
Generally, the above case studies can only be regarded as a proof-of-concept for the SimSchaake approach, as the available data set contains by far too little dates in the past to choose and to learn from. That is why the issue of time-weighting the ensemble similarity measures likely does not come into great effect. It is thus strongly recommended to repeat such studies on the basis on a very much larger data archive in a future work.

Nevertheless, one can guess from our promising initial results here that the SimSchaake scheme provides a reasonable and appealing multivariate postprocessing tool. It might compete with the ECC techniques, while having the benefit of a broader applicabilty, in that it can also be applied to ensembles comprising non-exchangeable members, and the SimSchaake ensemble is not restricted to have the same size as the raw ensemble.

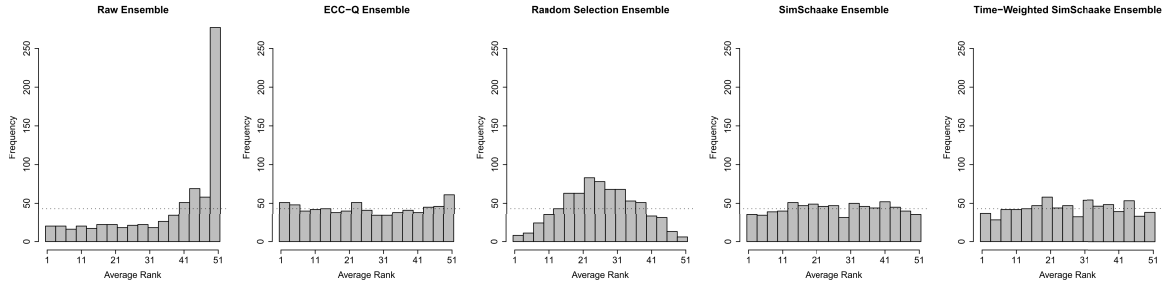
The choice of more sophisticatedly designed similarity measures and/or weighting schemes might improve the results for the SimSchaake technique once more.



(a) Multivariate rank histograms

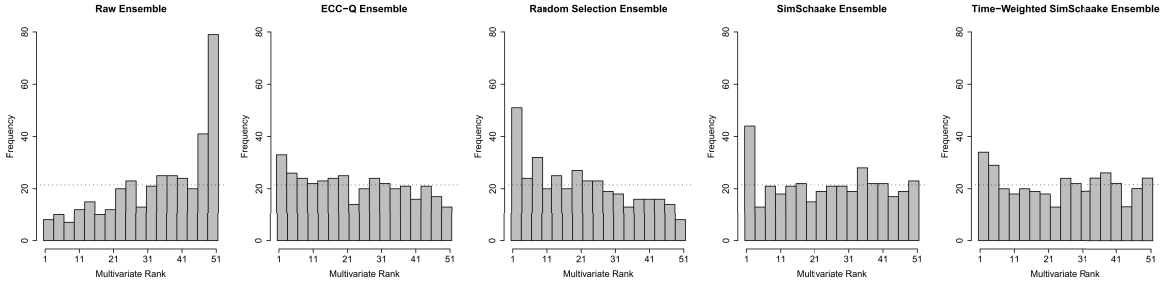


(b) Band depth rank histograms

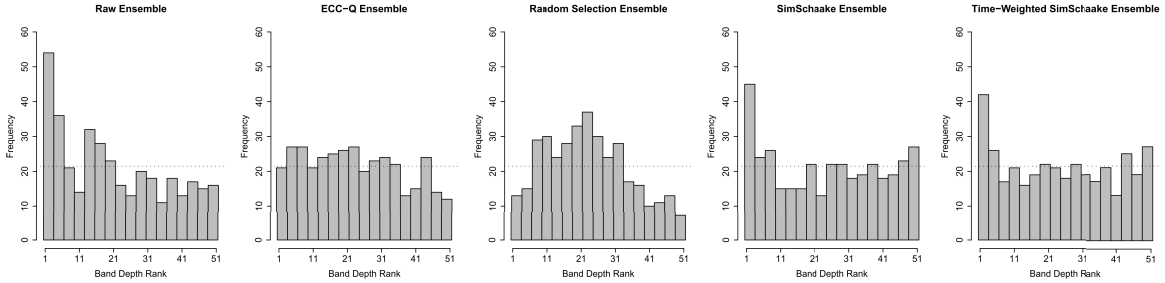


(c) Average rank histograms

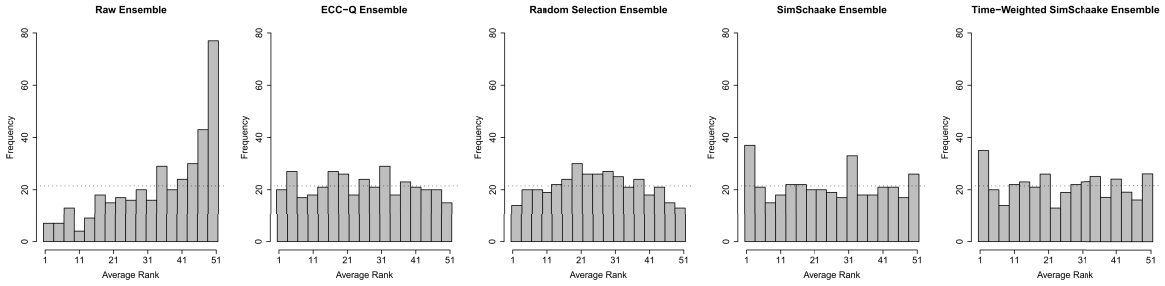
Figure 8.2: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 January 2011 to 31 December 2012. Univariate postprocessing is performed via EMOS.



(a) Multivariate rank histograms

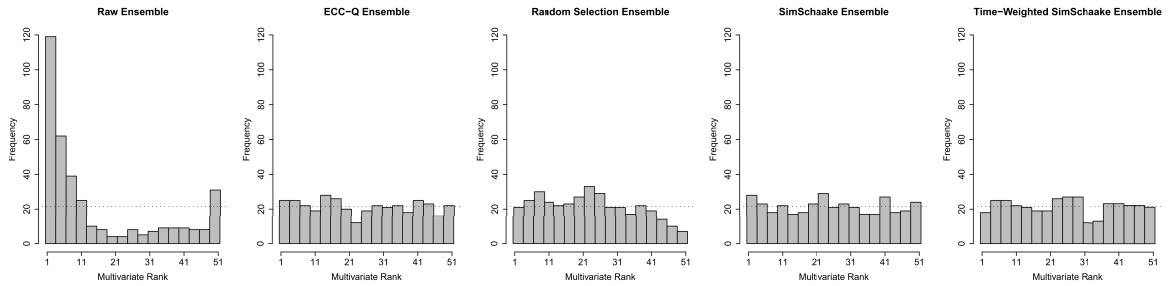


(b) Band depth rank histograms

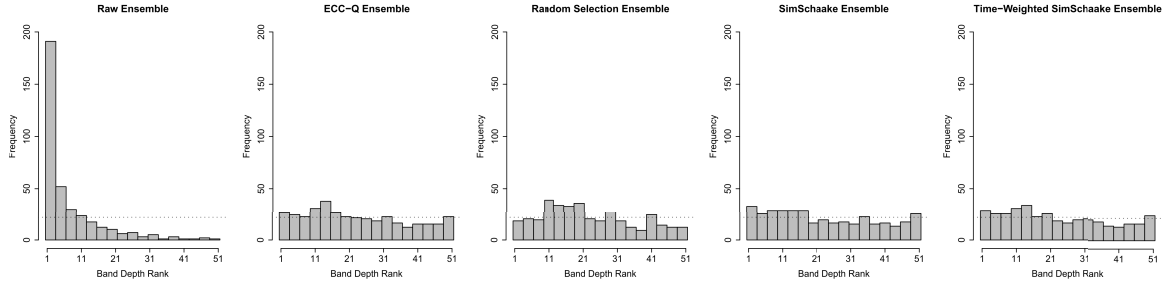


(c) Average rank histograms

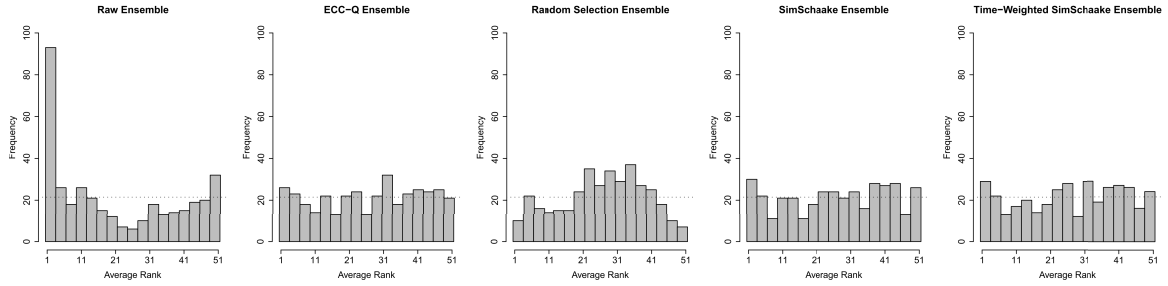
Figure 8.3: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead pressure forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.



(a) Multivariate rank histograms

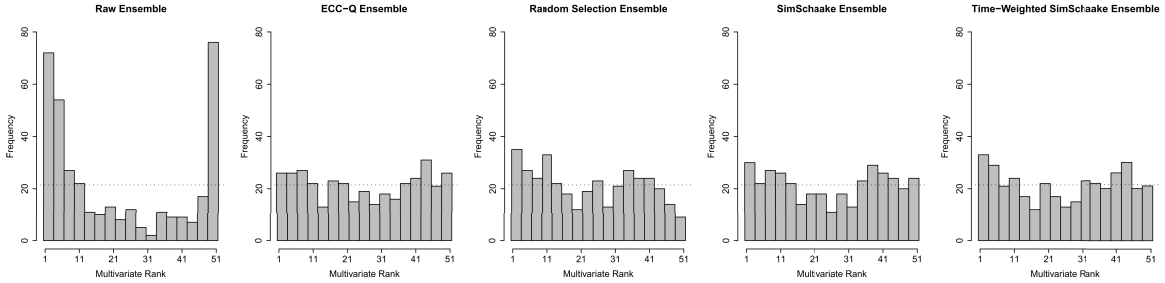


(b) Band depth rank histograms

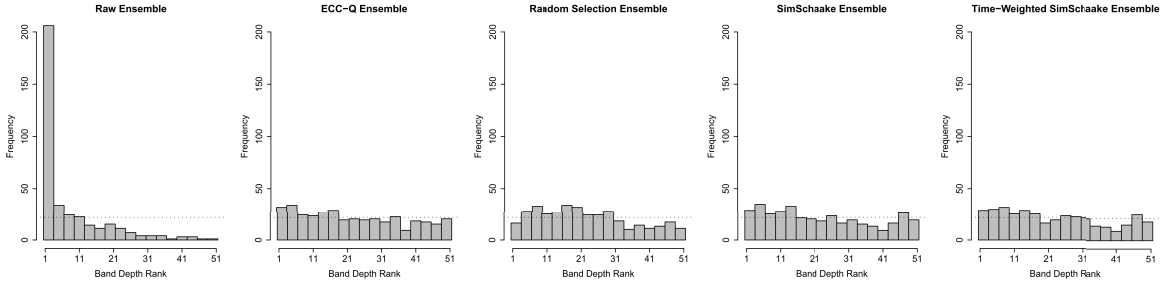


(c) Average rank histograms

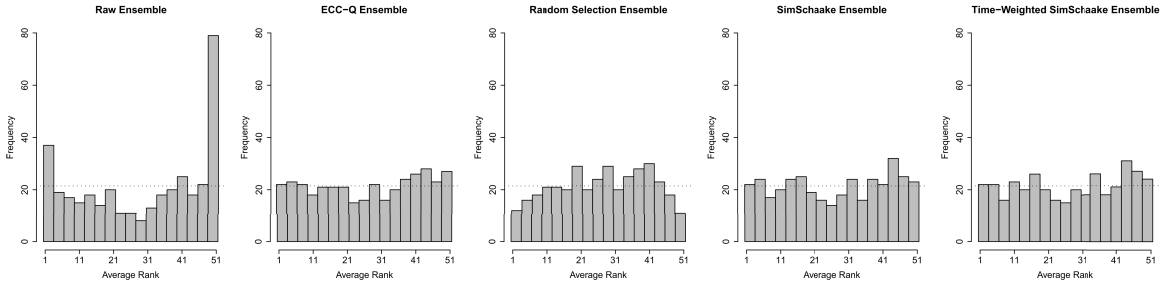
Figure 8.4: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead u-wind forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.



(a) Multivariate rank histograms



(b) Band depth rank histograms



(c) Average rank histograms

Figure 8.5: (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead v-wind forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011. Univariate postprocessing is performed via EMOS.

Chapter 9

Summary and discussion

Finally, we summarize and discuss the motivation and main results of this thesis and provide suggestions for further work in the future.

During the last years, a strong recognition of the need for uncertainty quantification with respect to the potentially high-dimensional output of complex computer models has taken place, with prediction often being a main goal. In this context, the field of probabilistic forecasting has developed steadily, as one nowadays often seeks to issue a forecast in terms of a predictive distribution instead of a single-valued point estimate. In particular, the analysis of uncertainty in weather forecasts has become of great interest, and meteorologists have designed numerical weather prediction (NWP) ensembles to provide forecasts in a probabilistic form. An ensemble comprises multiple runs of NWP models differing in the initial conditions and/or the model formulation in terms of the parameterized numerical representation of the atmosphere, where the 50-member ensemble of the European Centre for Medium-Range Weather Forecasts (ECMWF) employed in this thesis is a prominent and leading representative.

The design of an ensemble to account for uncertainties as discussed in this thesis can be seen as a Monte Carlo approach, in that inputs are generated randomly from a probability distribution, and multiple deterministic simulations are run. So-called stochastic Galerkin methods (Xiu and Karniadakis, 2002; Ghanem and Spanos, 2003; Constantine, 2007) provide an alternative concept, in which the solution of a differential equation with random input parameters is first represented by a truncated polynomial chaos expansion using multidimensional orthogonal basis polynomials. Then, this representation is plugged in the differential equation, and a Galerkin projection of the differential equation onto each basis polynomial is employed. Due to the orthogonality of the basis polynomials, the differential equation with random inputs is reduced to a system of coupled deterministic differential equations for the coefficients of the truncated polynomial chaos expansion (Constantine, 2007). Next, any suitable discretization of the coefficients can be employed to solve the system. Once the expansion coefficients are derived, the PDF of the solution can be approximated, and approximate moments or samples thereof can be calculated. More details are given in the above references. A discussion of advantages and disadvantages of Monte Carlo and stochastic Galerkin procedures to account for and quantify uncertainties is an issue for future work.

Unfortunately, NWP ensembles typically reveal biases and lack calibration. However, cali-

bration, referring to the statistical compatibility between the predictive distribution and the observation, is an essential property of a meaningful and convincing probabilistic forecast. Therefore, ensembles call for statistical postprocessing, and many univariate methods to do so have been developed. Examples include Bayesian model averaging (BMA) and ensemble model output statistics (EMOS), which mostly performed well in our case studies. However, these univariate postprocessing methods typically fail to account for spatial, temporal or inter-variable dependence structures, which are crucial in numerous applications. Hence, multivariate postprocessing being able to handle dependencies is of great importance, and several methods in this spirit have been presented in the literature. Frequently, these are based on the use of copulas, which are well suited for modeling multivariate dependencies whenever univariate margins are given, due to the celebrated Sklar's theorem.

In this thesis, we have introduced ensemble copula coupling (ECC) as a multivariate ensemble postprocessing method. ECC aims at physically coherent probabilistic forecasts of spatio-temporal weather trajectories and is a multi-stage approach. First, we are given a raw ensemble forecast from an NWP model, consisting of exchangeable ensemble members resulting from random perturbations in the inputs or model parameters. Second, univariate methods, such as BMA or EMOS, are employed to postprocess the raw ensemble forecast, leading to calibrated and sharp predictive distributions for each weather variable, location and look-ahead time individually. Then, each postprocessed marginal predictive distribution is represented by a discrete sample thereof, which has the same size M as the raw ensemble. Finally, each sample is reordered according to the rank dependence structure of the raw ensemble, thereby capturing the flow dependence. Thus, ECC returns a multivariate postprocessed forecast ensemble with statistically adjusted marginals in order to correct for systematic biases and/or misrepresentation of the prediction uncertainty, while the multivariate dependence structure is adapted from the original unprocessed NWP ensemble.

Depending on the employed quantization scheme (R), (T) or (Q) at the sampling stage in the ECC approach, we distinguish the variants ECC-R, ECC-T and ECC-Q, referring to the use of M Random samples, Transformations involving fitted CDFs or equidistant Quantiles, respectively. Generally, it is advisable to use the most intuitive and canonical sampling scheme (Q), employing the equally spaced $\frac{1}{M+1}$-, $\ldots$, $\frac{M}{M+1}$-quantiles of the corresponding univariate predictive CDFs, as a default. It turns out to be particularly useful to use equidistant sampling if M is rather small. However, the larger M gets, the smaller the differences between the quantization methods become, with equidistant sampling still performing best. In future research, one could seek optimality results with respect to sampling in multivariate settings.

The ECC approach, which is able to account for spatial, temporal and inter-variable dependence structures, has various benefits, being conceptionally very simple and straightforward to implement, and thus predestined to serve as a natural benchmark. ECC essentially comes computationally for free once the univariate postprocessing is done, which can be performed via any appropriate method. In contrast to purely parametric multivariate methods, ECC can handle model output of any dimensionality. It can be interpreted as an empirical copula-based approach and offers an overarching frame for seemingly unrelated postprocessing techniques scattered in the extant literature. ECC provides a general uncertainty quantification strategy that combines analytical, numerical and statistical modeling. Hence, it is appropriate not only for the key example of weather prediction, but also for a wide

range of other application areas. The ECC scheme can be applied whenever an ensemble of simulation runs is given which is able to appropriately represent multivariate dependence structures, and sufficient training data for the statistical correction of the univariate margins is available.

ECC has been tested and assessed in various case studies in this thesis and has been compared to the raw ensemble, the increasingly ordered quantiles (IOQ) ensemble, assuming a maximally positive dependence structure, and the randomly ordered quantiles (ROQ) ensemble, which assumes no correlation. Essentially, ECC performs well throughout and best in a majority of situations. In cases dealing with very strong dependence structures, the IOQ ensemble might perform best, whereas the ROQ ensemble may reveal the best result when there are very low correlations. However, ECC can compete in these cases of extreme correlation and appears to be at any rate a good and reasonable option, especially in settings in which the correlation structure is a priori unknown. The band depth and average rank histogram (Thorarinsdottir et al., 2014), respectively, as new verification tools for calibration in addition to the established multivariate rank histogram (MRH) (Gneiting et al., 2008), have proven to be very useful, in that they succeed in detecting miscalibration in comparably high dimensions, whereas the MRH might fail in such situations. Moreover, they are often able to distinguish clearer between the reference ensembles than the MRH, especially in high-dimensional settings. Their shapes in the case of over- or underestimation of the correlation structure by the IOQ or ROQ ensemble underline the corresponding simulation results by Thorarinsdottir et al. (2014) in a real-data based illustration. In general, it is advisable to use several verification tools to check calibration in the multivariate case (Thorarinsdottir et al., 2014), maybe also including the minimum spanning tree rank histogram (Smith, 2001; Smith and Hansen, 2004; Wilks, 2004; Gombos et al., 2007). Sometimes, the ROQ and ECC ensembles can hardly be distinguished in terms of the energy score (ES). This effect also emerges in a recent case study of Scheuerer and Hamill (2014) and might be at least partly based on the discrimination inability of the ES as discussed by Pinson and Tastu (2013). While univariate verification tools are already well established, it is important to design further evaluation methods for multivariate settings, both with respect to calibration and in terms of scores. In this context, Scheuerer and Hamill (2014) recently introduced variogram-based proper scoring rules, which are more discriminative with respect to correlation structures.

It has been shown that ECC serves as an overarching frame for various seemingly unrelated postprocessing approaches scattered in the literature. The list of the corresponding methods in this thesis cannot be expected to be complete, and quite surely other techniques also fitting into the ECC frame can be found or will be developed in the future. In particular, we have shown that a member-by-member postprocessing (MBMP) technique which has fused slightly differing formulations in the literature (Doblas-Reyes et al., 2005; Wood and Schaake, 2008; Johnson and Bowler, 2009; Van Schaeybroeck and Vannitsem, 2014) can be interpreted in the ECC-T frame. In this context, we have conducted a comparative case study, revealing that standard EMOS-based ensembles and standard EMOS-based ensembles combined with ECC can outperform the MBMP ensemble in univariate and multivariate settings, respectively.

ECC in its standard implementation also has shortcomings and limitations, which are partly set by its defining feature, that is, the adoption of the rank order pattern of the unprocessed

ensemble. For instance, ECC operates under the perfect model assumption that the raw ensemble is capable of reflecting the actual multivariate rank dependence structure. Although such an assumption appears to be reasonable and can typically be confirmed via diagnostic checks for state-of-the-art NWP models, it cannot be expected to be fulfilled each and every day. In general, it seems adequate to suppose that numerical models might reveal errors in dependence structures, which are to be diagnosed and improved to the extent deemed possible, and further steps in these directions ought to be taken.

Moreover, the standard ECC ensemble is constrained to have the same size M as the raw ensemble, which might be rather small in practice. However, it is often desirable to come up with a postprocessed sample having a larger size $N > M$, in particular for high-dimensional predictions based on a comparably small underlying raw ensemble. To increase the size of the postprocessed ECC ensemble, we have considered three approaches.

The extended ECC method operates with samples of size N from the univariate predictive CDFs instead of size M as in standard ECC. If N is an integer multiple of M, each sample is divided into subsamples consisting of M members each, which can be created by either random or equidistant grouping. Then, standard ECC is applied to each subsample, and the so obtained reordered subsamples are aggregated in an extended ECC ensemble forecast. A slight adaptation for the case that N is not an integer multiple of M has also been proposed. Extended ECC performed well in our case study and slightly improved standard ECC. In particular, using extended ECC makes good sense if the raw ensemble size M is rather small, while the improvements achieved via this method become minor if M is already large.

The recycling approach just takes the standard ECC ensemble and recycles it as often as desired, being restricted to the case that N is an integer multiple of M. A case study comparing the extended ECC-Q with the recycled ECC-R variant, which has also been proposed by Wilks (2014), is an interesting option for future work.

In the lagged ensemble variant, ensemble forecasts with different prediction horizons being valid at the same verification date are combined in order to increase the ensemble size.

Another shortcoming of standard ECC is that it can only be applied to ensembles consisting of exchangeable members. However, many ensemble systems comprise non-exchangeable ensemble members, with the NWP inputs and/or model parameters differing in a systematic rather than a random way. To address this issue, we have proposed an ECC modification which can be applied to groups or clusters of exchangeable ensemble members. However, the thoughts in this context are at their beginnings, and the variant still has to be tested and evaluated in a real-data case study, with the COSMO-DE ensemble prediction system (Gebhardt et al., 2011) of the German Weather Service appearing to be an appropriate candidate for that.

In addition to ECC, which uses information from the raw ensemble, there are various alternative approaches to model dependence structures, with some of them having been recapitulated in this thesis. If the considered model output is low-dimensional or reveals strong structure, carefully designed parametric copula approaches, allowing for the adjustment of any biases in the ensemble's representation of conditional correlation patterns, are

likely to outperform ECC, with Gaussian copula methods providing prominent options (Gel et al., 2004; Berrocal et al., 2007; Schuhen et al., 2012; Pinson, 2012; Möller et al., 2013, for instance). A comparison of the various methods in a real-data case study, with their pros and cons in concrete settings, should be tackled in a future work.

A prominent non-parametric example for a multivariate postprocessing method is the Schaake shuffle (Clark et al., 2004), which is based on the dependence structures in a historical record of weather data and admits an empirical copula interpretation. It performs well and is a useful concept, as is for instance witnessed by a recent paper of Vrac and Friederichs (2014), who combine the notion of the Schaake shuffle and a bias correction approach in a climatological context. The predictive performances of ECC and the Schaake shuffle have recently been compared by Wilks (2014) in a comprehensive case study, with the Schaake shuffle outperforming ECC. However, the results in Wilks (2014) are based on an $M = 11$-member ensemble, which is likely to be too small to get reasonable results with ECC, as the raw ensemble data one can learn from is limited. Thus, it is of great interest to conduct a similar study with the $M = 50$-member ECMWF ensemble employed in this thesis to investigate if ECC is able to perform better in this case.

From a theoretical point of view, we have introduced the notion of multivariate discrete copulas defined on $I_M^L := \left\{0, \frac{1}{M}, \ldots, \frac{M-1}{M}, 1\right\}^L$ and have shown that they are in a one-to-one correspondence to stochastic arrays. In addition, we have proven a multivariate discrete version of Sklar's theorem with the aid of an extension lemma. In future work, one could seek a generalization of the findings to the general case of multivariate discrete copulas defined on $I_{M_1} \times \cdots \times I_{M_L}$ rather than I_M^L, where $M_1, \ldots, M_L \in \mathbb{N}$ might be distinct. It has been shown that ECC and the related Schaake shuffle can be interpreted as discrete copula-based approaches. As mentioned before, many established postprocessing methods scattered in the literature can be regarded as special cases of the overarching ECC scheme. Due to our theoretical results, a unifying interpretation of these approaches in terms of discrete copulas is available. A future goal is to put the ECC variants discussed in this thesis into the discrete copula frame. For instance, it is desirable to develop theory for case of the ECC modification applicable to ensembles comprising non-exchangeable members, raising new questions on the combination of discrete copulas.

The standard ECC approach employs samples from univariate predictive CDFs obtained by postprocessing, which are then reordered according to the rank dependence structure of the raw ensemble, and can deal with high-dimensional settings. As noted before, parametric multivariate postprocessing methods in contrast typically perform well in low-dimensional or highly structured scenarios, where they often only apply to specific cases, such as for particular weather variables. To exploit the benefits of both concepts, we have proposed a generalization of ECC using Low-Dimensional Postprocessing, referred to as the LDP-ECC approach, in which samples from possibly multivariate distributions are aggregated in an ECC-like manner. In this context, we have examined notions of multivariate quantiles, with questions on optimal sampling still being open. Variants of LDP-ECC differing in the underlying multivariate ranking characteristic have basically proven to be promising in a case study dealing with (u, v)-wind vectors.

In addition to ECC and its modifications and generalizations, we have proposed the Sim-Schaake approach, which combines similarity-based ensemble methods with the idea of the

Schaake shuffle. In particular, appropriate similarity measures are employed to identify comparable past forecast cases, and the correlation structure of the corresponding verifying observations is transferred to postprocessed samples to get the final SimSchaake ensemble. In contrast to the case of standard ECC, the size of the SimSchaake ensemble does not need to equal that of the raw ensemble. In a case study, two distinct SimSchaake ensembles basically showed a good predictive performance by outperforming a postprocessed ensemble depending on a random selection of the corresponding past forecast cases in the historical record and being able to compete with the standard ECC-Q ensemble. Likely, the predictive performance of the SimSchaake variants could be further improved if the underlying historical weather database was larger. Hence, the conduction of a similar case study comprising much more historical data in a future work would be appreciated.

The field of probabilistic forecasting in general and the key application of statistical ensemble postprocessing in weather prediction settings in particular have been rapidly developing over the last decade. While the corresponding concepts and approaches are well established in the univariate case, the focus of recent and current research is on multivariate questions dealing with dependence modeling. This thesis has made a contribution to this highly critical subject, but as indicated there are still many open questions, both from a theoretical and an applied perspective. Specifically, the development of further sophisticated multivariate postprocessing methods and the design of new verification tools for multivariate scenarios (Scheuerer and Hamill, 2014) are topics for ongoing work, with a lot of progress to be expected in the coming years.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics & Economics*, 44, 182–198.

Abdous, B., and Theodorescu, R. (1992). Note on the spatial quantile of a random vector. *Statistics & Probability Letters*, 13, 333–336.

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

Bannayan, M., and Hoogenboom, G. (2008). Predicting realizations of daily weather data for climate forecasts using the non-parametric nearest-neighbour re-sampling technique. *International Journal of Climatology*, 28, 1357–1368.

Bao, L., Gneiting, T., Grimit, E. P., Guttorp, P., and Raftery, A. E. (2010). Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Monthly Weather Review*, 138, 1811–1821.

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

Baran, S., Horányi, A., and Nemoda, D. (2014a). Comparison of BMA and EMOS statistical methods in calibrating temperature and wind speed forecast ensembles. *Időjárás – Quarterly Journal of the Hungarian Meteorological Service*, 118, 217–241.

Baran, S., Horányi, A., and Nemoda, D. (2014b). Probabilistic temperature forecasting with statistical calibration in Hungary. *Meteorology and Atmospheric Physics*, 124, 129–142.

Baran, S., and Lerch, S. (2014). Log-normal distribution based EMOS models for probabilistic wind speed forecasting. `arXiv:1407.3252`.

Baran, S., and Möller, A. (2014). Joint probabilistic forecasting of wind speed and temperature using Bayesian model averaging. To appear in *Environmetrics*.

Ben Bouallègue, Z. (2013). Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Weather and Forecasting*, 28, 515–524.

Bentzien, S., and Friederichs, P. (2012). Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Weather and Forecasting*, 27, 988–1002.

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

Berrocal, V. J., Raftery, A. E., and Gneiting, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Annals of Applied Statistics*, 2, 1170–1193.

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

Bickel, P. J., and Lehmann, E. L. (1979). Descriptive statistics for nonparametric models: IV. Spread. In Jureckova, J., editor, *Contributions to Statistics*, pages 33–40. Academia, Prague.

Bishop, C. H., and Shanley, K. T. (2008). Bayesian model averaging's problematic treatment of extreme weather and a paradigm shift that fixes it. *Monthly Weather Review*, 136, 4641–4652.

Bjerknes, V. (1904). Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik. *Meteorologische Zeitschrift*, 21, 1–7.

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., Ebert, B., Fuentes, M., Hamill, T. M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y.-Y., Parsons, D., Raoult, B., Schuster, D., Silva Dias, P., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L., and Worley, S. (2010). The THORPEX interactive grand global ensemble. *Bulletin of the American Meteorological Society*, 91, 1059–1072.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society Series A*, 143, 383–425.

Bremnes, J. B. (2007). Improved calibration of precipitation forecasts using ensemble techniques. Part 2: Statistical calibration methods. Technical Report 04/2007, Norwegian Meteorological Institute. Available online at `http://met.no/Forskning/Publikasjoner/Publikasjoner_2007/filestore/report04_2007.pdf`.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.

Bröcker, J. (2012). Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, 138, 1611–1617.

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

Buizza, R. (2006). The ECMWF ensemble prediction system. In Palmer, T. N., and Hagedorn, R., editors, *Predictability of Weather and Climate*, pages 459–489. Cambridge University Press, Cambridge.

Buizza, R., Miller, M., and Palmer, T. N. (1999). Stochastic representations of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125, 2887–2908.

Chaloulos, G., and Lygeros, J. (2007). Effect of wind correlation on aircraft conflict probability. *Journal of Guidance, Control, and Dynamics*, 30, 1742–1752.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91, 862–872.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula Methods in Finance*. John Wiley & Sons, Chichester.

Chilès, J. P., and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. Wiley, Hoboken, 2nd edition.

Chmielecki, R. M., and Raftery, A. E. (2011). Probabilistic visibility forecasting using Bayesian model averaging. *Monthly Weather Review*, 139, 1626–1636.

Clark, M. P., Gangopadhyay, S., Hay, L. E., Rajagopalan, B., and Wilby, R. L. (2004). The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5, 243–262.

Constantine, P. G. (2007). A primer on stochastic Galerkin methods. Technical report, Stanford UQ. Available online at `http://inside.mines.edu/~pconstan/docs/constantine-primer.pdf`.

Cooke, E. (1906). Forecasts and verifications in Western Australia. *Monthly Weather Review*, 34, 23–24.

Courtney, J. F., Lynch, P., and Sweeney, C. (2013). High resolution forecasting for wind energy applications using Bayesian model averaging. *Tellus A*, 65, 19669.

Csima, J. (1970). Multidimensional stochastic matrices and patterns. *Journal of Algebra*, 14, 194–202.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65, 1254–1261.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012). Statistial modelling of spatial extremes. *Statistical Science*, 27, 161–186.

Dawid, A. P. (1984). Statistical theory: The prequential approach (with discussion and rejoinder). *Journal of the Royal Statistical Society Series A*, 147, 278–292.

Dawid, A. P., and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *The Annals of Statistics*, 27, 65–81.

Deheuvels, P. (1979). La fonction dépendence empirique et ses propriétés. Un test non-paramétrique d'independance. *Bulletin de la Classe des Sciences, V. Série, Académie Royale de Belgique*, 65, 274–292.

Delle Monache, L., Nipen, T., Liu, Y., Roux, G., and Stull, R. (2011). Kalman filter and analog schemes to postprocess numerical weather predictions. *Monthly Weather Review*, 139, 3554–3570.

Demarta, S., and McNeil, A. J. (2005). The t copula and related copulas. *International Statistical Review*, 73, 111–129.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39, 863–883.

Doblas-Reyes, F. J., Hagedorn, R., and Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting–II. Calibration and combination. *Tellus*, 57A, 234–252.

Eckel, F. A., and Mass, C. F. (2005). Aspects of effective mesoscale, short-range ensemble forecasting. *Weather and Forecasting*, 20, 328–350.

ECMWF Directorate (2012). Describing ECMWF's forecasts and forecasting system. *ECMWF Newsletter*, 133, 11–13.

Ehm, W., and Gneiting, T. (2012). Local proper scoring rules of order two. *The Annals of Statistics*, 40, 609–637.

Embrechts, P., Lindskog, F., and McNeil, A. (2003). Modelling dependence with copulas and applications to risk management. In Rachev, S. T., editor, *Handbook of Heavy Tailed Distributions in Finance*, pages 329–384. Elsevier, Amsterdam.

Epstein, E. S. (1969). Stochastic dynamic prediction. *Tellus*, 21, 739–759.

Erhardt, T. M., Czado, C., and Schepsmeier, U. (2014). R-vine models for spatial time series with an application to daily mean temperature. `arXiv:1403.3500`.

Evensen, G. (1994). Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99, 10143–10162.

Feldmann, K., Scheuerer, M., and Thorarinsdottir, T. L. (2014). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. `arXiv:1407.0058`.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.

Fermanian, J.-D., Radulović, D., and Wegkamp, M. (2004). Weak convergence of empirical copula approaches. *Bernoulli*, 10, 847–860.

Flowerdew, J. (2012). Calibration and combination of medium-range ensemble precipitation forecasts. Technical Report 567, United Kingdom MetOffice. Available online at `http://www.metoffice.gov.uk/media/pdf/h/6/FRTR567.pdf`.

Flowerdew, J. (2014). Calibrating ensemble reliability whilst preserving spatial structure. *Tellus A*, 66, 22662.

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

Fraley, C., Raftery, A. E., Gneiting, T., Sloughter, J. M., and Berrocal, V. J. (2011). Probabilistic weather forecasting in R. *R Journal*, 3, 55–63.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Annales de l' Université de Lyon, Sciences*, 14, 53–77.

Fricker, T. E., Ferro, C. A. T., and Stephenson, D. B. (2013). Three recommendations for evaluating climate predictions. *Meteorological Applications*, 20, 246–255.

Friederichs, P., and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23, 579–594.

Fritz, H., Filzmoser, P., and Croux, C. (2012). A comparison of algorithms for the multivariate L_1-median. *Computational Statistics*, 27, 393–410.

Gebhardt, C., Theis, S. E., Paulat, M., and Ben Bouallègue, Z. (2011). Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variations of lateral boundaries. *Atmospheric Research*, 100, 168–177.

Gel, Y., Raftery, A. E., and Gneiting, T. (2004). Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (GOP) method (with discussion and rejoinder). *Journal of the American Statistical Association*, 99, 575–590.

Genest, C., and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12, 347–368.

Genest, C., Gendron, M., and Bourdeau-Brien, M. (2009). The advent of copulas in finance. *European Journal of Finance*, 15, 609–618.

Genest, C., and Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin*, 37, 475–515.

Genest, C., Nešlehová, J., and Rémillard, B. (2014). On the empirical multilinear copula process for count data. *Bernoulli*, 20, 1344–1371.

Ghanem, R. G., and Spanos, P. D. (2003). *Stochastic Finite Elements: A Spectral Approach*. Dover Publications, New York.

Gigerenzer, G., Hertwig, R., van den Broek, E., Fasolo, B., and Katsikopoulos, K. V. (2005). "A 30 % chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Analysis*, 25, 623–629.

Gneiting, T. (2011a). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 746–762.

Gneiting, T. (2011b). Quantiles as optimal point forecasts. *International Journal of Forecasting*, 27, 197–207.

Gneiting, T. (2014). Calibration of medium-range weather forecasts. Technical Memorandum 719, ECMWF. Available online at `http://old.ecmwf.int/publications/library/ecpublications/_pdf/tm/701-800/tm719.pdf`.

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

Gneiting, T., and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and its Application*, 1, 125–151.

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

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

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

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

Gneiting, T., and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7, 1747–1782.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds (with discussion and rejoinder). *Test*, 17, 211–264.

Gombos, D., Hansen, J. A., Du, J., and McQueen, J. (2007). Theory and applications of the minimum spanning tree rank histogram. *Monthly Weather Review*, 135, 1490–1505.

Graf, S., and Luschgy, H. (2000). *Foundations of Quantization for Probability Distributions*. Lecture Notes in Mathematics 1730, Springer, Berlin.

Gupta, H. (1974). On permutation cubes and Latin cubes. *Indian Journal of Pure and Applied Mathematics*, 5, 1003–1021.

Hagedorn, R. (2010). On the relative benefits of TIGGE multi-model forecasts and reforecast-calibrated EPS forecasts. *ECMWF Newsletter*, 124, 17–23.

Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M., and Palmer, T. N. (2012). Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138, 1814–1827.

Hagedorn, R., Hamill, T. M., and Whitaker, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperature. *Monthly Weather Review*, 136, 2608–2619.

Hall, T. J., Thessin, R. N., Bloy, G. J., and Mutchler, C. N. (2010). Analog sky condition forecasting based on a k-nn algorithm. *Weather and Forecasting*, 25, 1463–1478.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129, 550–560.

Hamill, T. M. (2006). Ensemble-based atmospheric data assimilation: A tutorial. In Palmer, T. N., and Hagedorn, R., editors, *Predictability of Weather and Climate*, pages 124–156. Cambridge University Press.

Hamill, T. M. (2007). Comments on Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Monthly Weather Review*, 135, 4226–4230.

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

Hamill, T. M., Hagedorn, R., and Whitaker, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Monthly Weather Review*, 136, 2620–2632.

Hamill, T. M., and Whitaker, J. S. (2006). Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review*, 134, 3209–3229.

Harper, K., Uccellini, L. W., Morone, L., Kalnay, E., and Carey, K. (2007). 50th anniversary of operational numerical weather prediction. *Bulletin of the American Meteorological Society*, 88, 639–650.

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

Hoffman, R. N., and Kalnay, E. (1983). Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus A*, 35, 100–118.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.

Johnson, C., and Bowler, N. (2009). On the reliability and calibration of ensemble forecasts. *Monthly Weather Review*, 137, 1717–1720.

Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press.

Kann, A., Haiden, T., and Wittmann, C. (2011). Combining 2-m temperature nowcasting and short range ensemble forecasting. *Nonlinear Processes in Geophysics*, 18, 903–910.

Kann, A., Wittmann, C., Wang, Y., and Ma, X. (2009). Calibrating 2-m temperature of limited-area ensemble forecasts using high-resolution analysis. *Monthly Weather Review*, 137, 3373–3387.

Kemperman, J. H. B. (1987). The median of a finite measure on a Banach space. In Dodge, Y., editor, *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, pages 217–230. North-Holland, Amsterdam.

Klausner, Z., Kaplan, H., and Fattal, E. (2009). The similar days method for predicting near surface wind vectors. *Meteorological Applications*, 16, 569–579.

Kleiber, W., Raftery, A. E., Baars, J., Gneiting, T., Mass, C., and Grimit, E. P. (2011a). Locally calibrated probabilistic temperature foreasting using geostatistical model averaging and local Bayesian model averaging. *Monthly Weather Review*, 139, 2630–2649.

Kleiber, W., Raftery, A. E., and Gneiting, T. (2011b). Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting. *Journal of the American Statistical Association*, 106, 1291–1303.

Kolesárová, A., Mesiar, R., Mordelová, J., and Sempi, C. (2006). Discrete copulas. *IEEE Transactions on Fuzzy Systems*, 14, 698–705.

Krzysztofowicz, R., and Toth, Z. (2008). Bayesian processor of ensemble (BPE): Concept and implementation. Workshop slides 4th NCEP/NWS Ensemble User Workshop, Laurel, Maryland, available online at `http://www.emc.ncep.noaa.gov/gmb/ens/ens2008/Krzysztofowicz_Presentation_Web.pdf`.

Lawrence, A. R., and Hansen, J. A. (2007). A transformed lagged ensemble forecasting technique for increasing ensemble size. *Monthly Weather Review*, 135, 1424–1438.

Lerch, S., and Thorarinsdottir, T. L. (2013). Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A*, 65, 21206.

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

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28, 129–137.

López-Pintado, S., and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104, 718–734.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20, 130–141.

Lorenz, E. N. (1993). *The Essence of Chaos*. University of Washington Press, Seattle.

Lu, C., Yuan, H., Schwartz, B. E., and Benjamin, S. G. (2007). Short-range numerical weather prediction using time-lagged ensembles. *Weather and Forecasting*, 22, 580–595.

Lynch, P. (2008). The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, 227, 3431–3444.

Marchi, E., and Tarazaga, P. (1979). About (k, n) stochastic matrices. *Linear Algebra and its Applications*, 26, 15–30.

Mass, C., Joslyn, S., Pyle, J., Tewson, P., Gneiting, T., Raftery, A. E., Baars, J., Sloughter, J. M., Jones, D., and Fraley, C. (2009). PROBCAST: A web-based portal to mesoscale probabilistic forecasts. *Bulletin of the American Meteorological Society*, 90, 1009–1014.

Matheson, J. E., and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22, 1087–1096.

Max, J. (1960). Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6, 7–12.

Mayor, G., Suñer, J., and Torrens, J. (2005). Copula-like operations on finite settings. *IEEE Transactions on Fuzzy Systems*, 13, 468–477.

Mayor, G., Suñer, J., and Torrens, J. (2007). Sklar's theorem in finite settings. *IEEE Transactions on Fuzzy Systems*, 15, 410–416.

McNeil, A., and Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and ℓ_1-norm symmetric distributions. *Annals of Statistics*, 37, 3059–3097.

Mesiar, R. (2005). Discrete copulas — what they are. In Montseny, E., and Sobrevilla, P., editors, *Joint EUSFLAT-LFA 2005*, pages 927–930. Universitat Politècnica de Catalunya, Barcelona.

Messner, J. W., and Mayr, G. J. (2011). Probabilistic forecasts using analogs in the idealized Lorenz96 setting. *Monthly Weather Review*, 139, 1960–1971.

Mikosch, T. (2006). Copulas: Tales and facts. *Extremes*, 9, 3–20.

Mittermaier, M. P. (2007). Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Quarterly Journal of the Royal Meteorological Society*, 133, 1487–1500.

Möller, A. (2014). *Multivariate and Spatial Ensemble Postprocessing Methods*. PhD thesis, Faculty of Mathematics and Computer Science, Heidelberg University.

Möller, A., Lenkoski, A., and Thorarinsdottir, T. L. (2013). Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, 139, 982–991.

Möller, D. (2013). *Postprocessing of Ensemble Forecasts for Wind Speed over Germany*. Diploma thesis, Faculty of Mathematics and Computer Science, Heidelberg University. Available online at `http://www.rzuser.uni-heidelberg.de/~kd4/files/Moeller2013.pdf`.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. (1996). The new ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122, 73–119.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, New York, 2nd edition.

O'Hagan, A. (2003). HSSS model criticism. In Green, P. J., Hjort, N. L., and Richardson, S., editors, *Highly Structured Stochastic Systems*, pages 423–444. Oxford University Press, Oxford.

Paez, M., and Glisson, T. H. (1972). Minimum mean-squared-error quantization in speech PCM and DPCM systems. *IEEE Transactions on Communications*, 20, 225–230.

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

Palmer, T. N., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G. J., Steinheimer, M., and Weisheimer, A. (2009). Stochastic parameterization and model uncertainty. Technical Report ECMWF Technical Memorandum 598, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2-9AX, UK. Available online at `http://www2.physics.ox.ac.uk/sites/default/files/2011-08-15/techmemo598_stochphys_2009_pdf_50419.pdf`.

Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107, 1063–1072.

Parry, M., Dawid, A. P., and Lauritzen, S. (2012). Proper local scoring rules. *The Annals of Statistics*, 40, 561–592.

Pfeifer, D., and Nešlehová, J. (2003). Modeling dependence in finance and insurance: The copula approach. *Blätter der deutschen Gesellschaft für Versicherungs- und Finanzmathematik*, XXVI/2, 177–191.

Pinson, P. (2012). Adaptive calibration of (u, v)-wind ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138, 1273–1284.

Pinson, P. (2013). Wind energy: Forecasting challenges for its operational management. *Statistical Science*, 28, 564–585.

Pinson, P., and Girard, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, 96, 12–20.

Pinson, P., Madsen, H., Nielsen, H. A., Papaefthymiou, G., and Klöckl, B. (2009). From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy*, 12, 51–62.

Pinson, P., and Tastu, J. (2013). Discrimination ability of the Energy score. Technical report, Technical University of Denmark. Available online at `http://orbit.dtu.dk/fedora/objects/orbit:122326/datastreams/file_b919613a-9043-4240-bb6c-160c88270881/content`.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Raftery, A. E., Kárný, M., and Ettler, M. (2010). Online prediction under model uncertainty via dynamic model averaging. *Technometrics*, 52, 52–66.

Richardson, L. F. (1922). *Weather Prediction by Numerical Process*. Cambridge University Press, Cambridge (reprinted by Dover Publications, New York, 1965).

Roquelaure, S., and Bergot, T. (2008). A local ensemble prediction system for fog and low clouds: Construction, Bayesian model averaging calibration, and validation. *Journal of Applied Meteorology and Climatology*, 47, 3072–3088.

Roulin, E., and Vannitsem, S. (2012). Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Monthly Weather Review*, 140, 874–888.

Rüschendorf, L. (1976). Asymptotic distributions of multivariate rank order statistics. *Annals of Statistics*, 4, 912–923.

Rüschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139, 3921–3927.

Schaake, J., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X., and Seo, D. J. (2007). Precipitation and temperature ensemble forecasts from single-valued forecasts. *Hydrology and Earth Systems Sciences Discussions*, 4, 655–717.

Schaake, J., Pailleux, J., Thielen, J., Arritt, R., Hamill, T. M., Luo, L., Martin, E., McCollor, D., and Pappenberger, F. (2010). Summary of recommendations of the first Workshop on Postprocessing and Downscaling Atmospheric Forecasts for Hydrologic Applications held at Météo-France, Toulouse, France, 15-18 June 2009. *Atmospheric Science Letters*, 11, 59–63.

Schefzik, R. (2011). *Ensemble Copula Coupling*. Diploma thesis, Faculty of Mathematics and Computer Science, Heidelberg University. Available online at `http://www.nr.no/~thordis/files/Schefzik2011.pdf`.

Schefzik, R. (2013). Ensemble copula coupling as a multivariate discrete copula approach. `arXiv:1305.3445`.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28, 616–640.

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140, 1086–1096.

Scheuerer, M., and Büermann, L. (2014). Spatially adaptive post-processing of ensemble forecasts for temperature. *Journal of the Royal Statistical Society, Series C*, 63, 405–422.

Scheuerer, M., and Gneiting, T. (2011). Spatial Bayesian Model Averaging for Precipitation Forecasts from COSMO-DE-EPS. Workshop slides 11th Annual Meeting of the European Meteorological Society and 10th European Conference on Applications of Meteorology, September 2011, Berlin, Germany, available online at `http://www.presentations.copernicus.org/EMS2011-101_presentation.pdf`.

Scheuerer, M., and Hamill, T. M. (2014). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. Preprint, available online at `http://www.esrl.noaa.gov/psd/people/michael.scheuerer/variogram-score.pdf`.

Scheuerer, M., and König, G. (2014). Gridded locally calibrated, probabilistic temperature forecasts based on ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*. `doi:10.1002/qj.2323`.

Scheufele, K., Kober, K., Craig, G. C., and Keil, C. (2014). Combining probabilistic precipitation forecasts from a nowcasting technique with a time-lagged ensemble. *Meteorological Applications*, 21, 230–240.

Schoelzel, C., and Friederichs, P. (2008). Multivariate non-normally distributed random variables in climate research — Introduction to the copula approach. *Nonlinear Processes in Geophysics*, 15, 761–772.

Schoelzel, C., and Hense, A. (2011). Probabilistic assessment of regional climate change in Southwest Germany by ensemble dressing. *Climate Dynamics*, 36, 2003–2014.

Schuhen, N., Thorarinsdottir, T. L., and Gneiting, T. (2012). Ensemble model output statistics for wind vectors. *Monthly Weather Review*, 140, 3204–3219.

Sempi, C. (2011). Copulae: Some mathematical aspects. *Applied Stochastic Models in Business and Industry*, 27, 37–50.

Serfling, R. (2002a). A depth function and a scale curve based on spatial quantiles. In Dodge, Y., editor, *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, pages 25–38. Birkhäuser, Basel.

Serfling, R. (2002b). Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, 56, 214–232.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

Sloughter, J. M., Gneiting, T., and Raftery, A. E. (2010). Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, 105, 25–35.

Sloughter, J. M., Gneiting, T., and Raftery, A. E. (2013). Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Monthly Weather Review*, 141, 2107–2119.

Sloughter, J. M., Raftery, A. E., Gneiting, T., and Fraley, C. (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 Mees, A. I., editor, *Nonlinear Dynamics and Statistics*, pages 31–64. Birkhäuser, Boston.

Smith, L. A., and Hansen, J. A. (2004). Extending the limits of ensemble forecast verification with the minimum spanning tree histogram. *Monthly Weather Review*, 132, 1522–1528.

Soltanzadeh, I., Azadi, M., and Vakili, G. A. (2011). Using Bayesian model averaging (BMA) to calibrate probabilistic surface temperature forecasts over Iran. *Annales Geophysicae*, 29, 1295–1303.

Stephenson, D. B., and Doblas-Reyes, F. J. (2000). Statistical methods for interpreting Monte Carlo forecasts. *Tellus A*, 52, 300–322.

Talagrand, O., Vautard, R., and Strauss, B. (1997). Evaluation of probabilistic prediction systems. In *Proceeding of ECMWF Workshop on Predictability*, pages 1–25. Reading, UK, European Centre for Medium-Range Weather Forecasts.

Thorarinsdottir, T. L., and Gneiting, T. (2010). Probabilistic forecasts of wind speed: Ensemble model output statistics using heteroskedastic censored regression. *Journal of the Royal Statistical Society, Series A*, 173, 371–388.

Thorarinsdottir, T. L., and Johnson, M. S. (2012). Probabilistic wind gust forecasting using non-homogeneous Gaussian regression. *Monthly Weather Review*, 140, 889–897.

Thorarinsdottir, T. L., Scheuerer, M., and Feldmann, K. (2012). Statistische Nachbereitung von Ensemblevorhersagen. *Promet*, 37, 43–52.

Thorarinsdottir, T. L., Scheuerer, M., and Heinz, C. (2014). Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. To appear in *Journal of Computational and Graphical Statistics*.

Todini, E. (2008). A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management*, 6, 123–137.

Toth, Z., and Kalnay, E. (1997). Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review*, 125, 3297–3319.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.

van der Vaart, A. W., and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Van Schaeybroeck, B., and Vannitsem, S. (2014). Ensemble post-processing using member-by-member approaches: theoretical aspects. *Quarterly Journal of the Royal Meteorological Society*. doi:10.1002/qj.2397.

Vardi, Y., and Zhang, C. H. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of the United States of America*, 97, 1423–1426.

Voisin, N., Pappenberger, F., Lettenmeier, D. P., Buizza, R., and Schaake, J. C. (2011). Application of a medium-range global hydrologic probabilistic forecast scheme to the Ohio River basin. *Weather and Forecasting*, 26, 425–446.

von Storch, H. (1999). On the use of "inflation" in statistical downscaling. *Journal of Climate*, 12, 3505–3506.

Vrac, M., and Friederichs, P. (2014). Multivariate – inter-variable, spatial and temporal – bias correction. To appear in *Journal of Climate*.

Wilks, D. S. (2002). Smoothing forecast ensembles with fitted probability distributions. *Quarterly Journal of the Royal Meteorological Society*, 128, 2821–2836.

Wilks, D. S. (2004). The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Monthly Weather Review*, 132, 1329–1340.

Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications*, 16, 361–368.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. Academic Press, Amsterdam, 3rd edition.

Wilks, D. S. (2014). Multivariate ensemble Model Output Statistics using empirical copulas. *Quarterly Journal of the Royal Meteorological Society*. doi:10.1002/qj.2414.

Wilks, D. S., and Hamill, T. M. (2007). Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, 135, 2379–2390.

Williams, R. M., Ferro, C. A. T., and Kwasniok, F. (2014). A comparison of ensemble postprocessing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140, 1112–1120.

Wilson, L. J., Beauregard, S., Raftery, A. E., and Verret, R. (2007). Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Monthly Weather Review*, 135, 1364–1385.

Wood, A. W., and Schaake, J. C. (2008). Correcting errors in streamflow forecast ensemble mean and spread. *Journal of Hydrometeorology*, 9, 132–148.

Woods, A. (2006). *Medium-Range Weather Prediction. The European Approach. The Story of the European Centre for Medium-Range Weather Forecasts*. Springer, New York.

Xiu, D., and Karniadakis, G. E. (2002). The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM Journal of Scientific Computing*, 24, 619–644.

Xu, J. J. (1996). *Statistical Modelling and Inference for Multivariate and Longitudinal Discrete Response Data*. PhD thesis, The University of British Columbia.

Zhang, Q., Li, J., and Singh, V. P. (2012). Application of Archimedean copulas in the analysis of the precipitation extremes: Effects of precipitation changes. *Theoretical and Applied Climatology*, 107, 255–264.

Ziegel, J. F., and Gneiting, T. (2013). Copula calibration. `arXiv:1307.7650`.

List of Figures

1.1 Illustration of the goal of probabilistic forecasting 2
1.2 24 hour ahead ensemble forecast for temperature over Germany, valid at 2:00 am on 3 July 2010 5
1.3 24 hour ahead 50-member ECMWF ensemble forecasts for temperature at Hamburg and corresponding verifying observations, valid 2:00 am for the period from 1 April 2011 to 14 April 2011 6
1.4 24 hour ahead EMOS predictive PDF for temperature at Hamburg, valid 2:00 am on 1 April 2011 7

3.1 24 hour ahead BMA predictive PDF for temperature at Berlin, valid 2:00 am on (a) 21 April 2011 and (b) 25 April 2011, respectively 23
3.2 24 hour ahead BMA predictive PDF for six hour precipitation accumulation at Hamburg, valid 1:00 am on 11 December 2010 24
3.3 24 hour ahead EMOS predictive normal PDF for pressure at Frankfurt, valid 2:00 am on 20 June 2010 27
3.4 24 hour ahead ensemble forecasts of temperature at Hamburg, valid 2:00 am for the test period from 1 April 2011 to 14 April 2011, and characteristic quantities of the corresponding postprocessed (a) BMA and (b) EMOS predictive distribution, respectively 29
3.5 24 hour ahead ensemble forecasts of u-wind at Frankfurt, valid 1:00 am for the test period from 1 January 2011 to 14 January 2011, and characteristic quantities of the corresponding postprocessed (a) BMA and (b) EMOS predictive distribution, respectively 29
3.6 Calibration check for 24 hour ahead forecasts of temperature, pressure, u- and v-wind and precipitation at Berlin over the test period from 1 May 2010 to 30 April 2011 30
3.7 Calibration check for 24 hour ahead forecasts of temperature, pressure, u- and v-wind and precipitation at Hamburg over the test period from 1 May 2010 to 30 April 2011 31
3.8 Calibration check for 24 hour ahead forecasts of temperature, pressure, u- and v-wind and precipitation at Frankfurt over the test period from 1 May 2010 to 30 April 2011 32

4.1 24 hour ahead ensemble forecasts of temperature and pressure at Berlin and Hamburg, valid 1:00 am on 8 November 2010, with univariate postprocessing performed via BMA 41
4.2 24 hour ahead ensemble forecasts of temperature and pressure at Berlin and Hamburg, valid 1:00 am on 8 November 2010, with univariate postprocessing performed via EMOS 42

4.3 24 hour ahead ensemble forecasts of temperature on a grid over Germany, valid 2:00 am on 23 April 2011 43
4.4 24 hour ahead 50-member ECMWF raw ensemble forecast for temperature on a grid over Germany, valid 2:00 am on 23 April 2011 45
4.5 24 hour ahead 50-member individual BMA ensemble forecast for temperature on a grid over Germany, valid 2:00 am on 23 April 2011 46
4.6 24 hour ahead 50-member ECC ensemble forecast for temperature on a grid over Germany, valid 2:00 am on 23 April 2011 47
4.7 CRPS and ES, respectively, for 24 hour ahead temperature forecasts according to different quantization schemes at (a) Berlin, (b) Hamburg and (c) Berlin and Hamburg jointly, as a function of the ensemble size and averaged over the test period from 1 May 2010 to 30 April 2011 52
4.8 CRPS and ES, respectively, for 24 hour ahead pressure forecasts according to different quantization schemes at (a) Berlin, (b) Hamburg and (c) Berlin and Hamburg jointly, as a function of the ensemble size and averaged over the test period from 1 May 2010 to 30 April 2011 54
4.9 CRPS and ES, respectively, for 24 hour ahead u-wind forecasts according to different quantization schemes at (a) Berlin, (b) Hamburg and (c) Berlin and Hamburg jointly, as a function of the ensemble size and averaged over the test period from 1 May 2010 to 30 April 2011 56
4.10 CRPS and ES, respectively, for 24 hour ahead v-wind forecasts according to different quantization schemes at (a) Berlin, (b) Hamburg and (c) Berlin and Hamburg jointly, as a function of the ensemble size and averaged over the test period from 1 May 2010 to 30 April 2011 57
4.11 Verification rank histograms for 24 hour ahead pressure forecasts at Berlin according to different quantization schemes and for different ensemble sizes over the test period from 1 May 2010 to 30 April 2011 58
4.12 Verification rank histograms for 24 hour ahead pressure forecasts at Hamburg according to different quantization schemes and for different ensemble sizes over the test period from 1 May 2010 to 30 April 2011 59
4.13 Multivariate rank histograms for 24 hour ahead pressure forecasts at Berlin and Hamburg jointly according to different quantization schemes and for different ensemble sizes over the test period from 1 May 2010 to 30 April 2011 . 60
4.14 Verification rank histograms for 24 hour ahead u-wind forecasts at Berlin according to different quantization schemes and for different ensemble sizes over the test period from 1 May 2010 to 30 April 2011 61
4.15 Verification rank histograms for 24 hour ahead u-wind forecasts at Hamburg according to different quantization schemes and for different ensemble sizes over the test period from 1 May 2010 to 30 April 2011 62
4.16 Multivariate rank histograms for 24 hour ahead u-wind forecasts at Berlin and Hamburg jointly according to different quantization schemes and for different ensemble sizes over the test period from 1 May 2010 to 30 April 2011 63
4.17 24 hour ahead ensemble forecasts of temperature at Berlin and Hamburg, valid 2:00 am on 27 June 2010 65
4.18 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011 69

4.19 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead pressure forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011 70
4.20 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead u-wind forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011 71
4.21 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead v-wind forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011 72
4.22 Test locations and areas over Germany and surrounding regions, including (a) the three observation sites Berlin, Hamburg and Frankfurt, (b) the four contiguous test regions and (c) the scattered test set 75
4.23 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature forecasts for test area II, that is, $L = 45$, over the test period from 1 May 2010 to 30 April 2011 76
4.24 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead pressure forecasts for test area II, that is, $L = 45$, over the test period from 1 May 2010 to 30 April 2011 77
4.25 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead u-wind forecasts for test area II, that is, $L = 45$, over the test period from 1 May 2010 to 30 April 2011 78
4.26 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead v-wind forecasts for test area II, that is, $L = 45$, over the test period from 1 May 2010 to 30 April 2011 79
4.27 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead joint temperature and pressure forecasts at Berlin, that is, $L = 2$, over the test period from 1 May 2010 to 30 April 2011 81
4.28 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead joint u- and v-wind forecasts at Frankfurt, that is, $L = 2$, over the test period from 1 May 2010 to 30 April 2011 82
4.29 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature and pressure forecasts jointly at Berlin, Hamburg and Frankfurt, that is, $L = 6$, over the test period from 1 May 2010 to 30 April 2011 84
4.30 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature, pressure, u- and v-wind forecasts jointly at Berlin, Hamburg and Frankfurt, that is, $L = 12$, over the test period from 1 May 2010 to 30 April 2011 85
4.31 (a) Multivariate, (b) band depth and (c) average rank histograms for joint 24, 48, 72 and 96 hour ahead temperature forecasts at Hamburg, that is, $L = 4$, over the 1000 initialization days from 1 April 2010 to 25 December 2012 . . . 87
4.32 Scenario (A): Scatterplots for 24 hour ahead temperature forecasts at Hamburg and Berlin, valid 1:00 am on 24 November 2010, of a 10-member raw ensemble, the corresponding 10-member standard ECC-Q ensemble and the 30-member extended ECC-Q random and extended ECC-Q equidistant ensembles, respectively 93

4.33 Scenario (B): Scatterplots for 24 hour ahead temperature forecasts at Hamburg and Berlin, valid 1:00 am on 24 November 2010, of a 10-member raw ensemble, the corresponding 10-member standard ECC-Q ensemble and the 25-member extended ECC-Q random ensemble 95
4.34 Average ES for 24 hour ahead temperature forecasts at Berlin and Hamburg jointly over the test period from 1 January 2011 to 31 December 2012, based on an $M = 10$-member raw ensemble and as a function of the desired extended ensemble size $N \geq M$ 96
4.35 Average ES for 24 hour ahead temperature forecasts at Berlin and Hamburg jointly over the test period from 1 January 2011 to 31 December 2012, based on an $M = 30$-member raw ensemble and as a function of the desired extended ensemble size $N \geq M$ 96
4.36 Average ES for 24 hour ahead temperature forecasts at Berlin and Hamburg jointly over the test period from 1 January 2011 to 31 December 2012, based on an $M = 50$-member raw ensemble and as a function of the desired extended ensemble size $N \geq M$ 97
4.37 Scenario (A): (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature forecasts at Berlin and Hamburg jointly over the test period from 1 January 2011 to 31 December 2012, with an $M = 10$-member raw ensemble extended to an $N = 50$-member ensemble . 98
4.38 Scenario (B): (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature forecasts at Berlin and Hamburg jointly over the test period from 1 January 2011 to 31 December 2012, where an $M = 10$-member raw ensemble is extended to different $N = 35$-member ensembles 99
4.39 Illustration of the connection between initialization dates, verification date and prediction horizons to construct an enlarged lagged raw ensemble 101

5.1 Verification rank histograms for 24 hour ahead temperature forecasts at (a) Berlin and (b) Hamburg for the test period from 1 January 2011 to 31 December 2012 109
5.2 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature forecasts at Berlin and Hamburg jointly for the test period from 1 January 2011 to 31 December 2012 111

6.1 Different 24 hour ahead ensemble prediction approaches for temperature at Berlin and Hamburg comprising the (a) ECMWF raw, (b) IOQ, (c) ROQ and (d) ECC-Q ensemble, valid 2:00 am on 27 June 2010 133
6.2 Pressure at Berlin and Hamburg: (a) 24 hour ahead ECMWF raw ensemble forecast valid 1:00 am on 9 March 2011, (b) past observations from 7 February 2011 to 8 March 2011, (c) ROQ ensemble forecast valid 1:00 am on 9 March 2011 and (d) Schaake shuffle ensemble valid 1:00 am on 9 March 2011 136

7.1 Illustration of the LDP-ECC approach based on Example 7.2 and Tables 7.1 to 7.3 147
7.2 (a) Multivariate, (b) band depth and (c) average rank histograms for 48 hour ahead joint u- and v-wind forecasts at Berlin and Hamburg simultaneously, that is, $L = 4$, over the test period from 1 May 2010 to 30 April 2011 150

8.1 Scheme of the SimSchaake approach 155
8.2 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead temperature forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 January 2011 to 31 December 2012 . . . 160
8.3 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead pressure forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011 161
8.4 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead u-wind forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011 162
8.5 (a) Multivariate, (b) band depth and (c) average rank histograms for 24 hour ahead v-wind forecasts at Berlin, Hamburg and Frankfurt jointly, that is, $L = 3$, over the test period from 1 May 2010 to 30 April 2011 163

List of Tables

3.1 BMA settings for univariate weather quantities 22
3.2 EMOS settings for univariate weather quantities 26
3.3 Average CRPS and AE over the test period from 1 May 2010 to 30 April 2011 for univariate 24 hour ahead forecasts of temperature, pressure, u- and v-wind and precipitation at Berlin, Hamburg and Frankfurt 30

4.1 Average ES over the test period from 1 May 2010 to 30 April 2011 for 24 hour ahead ensemble forecasts of temperature, pressure, u- and v-wind, respectively, at Berlin, Hamburg and Frankfurt jointly ($L = 3$) and at Berlin and Hamburg jointly ($L = 2$), respectively 66
4.2 Average EE over the test period from 1 May 2010 to 30 April 2011 for 24 hour ahead ensemble forecasts of temperature, pressure, u- and v-wind, respectively, at Berlin, Hamburg and Frankfurt jointly ($L = 3$) and at Berlin and Hamburg jointly ($L = 2$), respectively 67
4.3 Average DS over the test period from 1 May 2010 to 30 April 2011 for 24 hour ahead ensemble forecasts of temperature, pressure, u- and v-wind, respectively, at Berlin, Hamburg and Frankfurt jointly ($L = 3$) and at Berlin and Hamburg jointly ($L = 2$), respectively 68
4.4 Test areas over Germany and surrounding regions 74
4.5 Average ES over the test period from 1 May 2010 to 30 April 2011 for 24 hour ahead ensemble forecasts of temperature, pressure, u- and v-wind separately over the test areas I to V according to Table 4.4 and Figure 4.22 75
4.6 Average ES over the test period from 1 May 2010 to 30 April 2011 for 24 hour ahead standardized joint forecasts of temperature and pressure and joint forecasts of u- and v-wind, respectively, at the individual stations of Berlin, Hamburg and Frankfurt 83
4.7 Average ES over the test period from 1 May 2010 to 30 April 2011 for different combinations of 24 hour ahead standardized joint forecasts of temperature, pressure, u- and v-wind at Berlin, Hamburg and Frankfurt 86
4.8 Average ES and DSS over 1000 initialization days from 1 April 2010 to 25 December 2012 for joint 24, 48, 72 and 96 hour ahead forecasts of temperature at Berlin, Hamburg and Frankfurt, respectively 88

5.1 Average CRPS and DSS over the test period from 1 January 2011 to 31 December 2012 for univariate 24 hour ahead forecasts of temperature at Berlin and Hamburg 108
5.2 Average ES and DSS over the test period from 1 January 2011 to 31 December 2012 for 24 hour ahead temperature forecasts at Berlin and Hamburg jointly 110

6.1 Example of a discrete extension copula D for the scenario in Example 6.14 . 131

7.1 Raw ensemble values, corresponding multivariate ranking characteristics according to (7.6) and permutations in the setting of Example 7.2 145
7.2 Bivariate samples from $F_{\text{Ber }48}$ and $F_{\text{Ham }48}$, respectively, corresponding multivariate ranking characteristics according to (7.6) and permutations for the setting of Example 7.2 145
7.3 The LDP-ECC ensemble: Reordered bivariate samples from Table 7.2 146
7.4 Average ES over the test period from 1 May 2010 to 30 April 2011 for 48 hour ahead joint u- and v-wind ensemble forecasts at Berlin and Hamburg simultaneously 149

8.1 Average ES for 24 hour ahead forecasts at Berlin, Hamburg and Frankfurt jointly over the test period from 1 May 2010 to 30 April 2011 (temperature, pressure, u- and v-wind) and from 1 January 2011 to 31 December 2012 (temperature), respectively 158