Aspects of Spatial Postprocessing for Global Temperature Forecasts

Inauguraldissertation

zur Erlangung des akademischen Grades

eines Doktors der Naturwissenschaften

der Universität Mannheim

vorgelegt von

Kira Feldmann

Mannheim, 2022

Dekan: Dr. Bernd Lübcke, Universität Mannheim
Referent: Prof. Dr. Martin Schlather, Universität Mannheim
Korreferent: Prof. Dr. Tilmann Gneiting, Karlsruher Institut für Technologie und
Heidelberger Institut für Theoretische Studien

Tag der mündlichen Prüfung: 8. März 2022

Abstract

High quality predictions are essential for informed decision-making. This holds especially true in meteorology as weather phenomena can engender high socioeconomic cost. In the past decades, the paradigm in weather prediction has shifted from point forecasts to probabilistic forecasts providing a probability distribution aiming to capture the true uncertainty of the prediction. Operationally, these probabilistic forecasts are generated by ensembles consisting of multiple runs of numerical weather prediction systems that differ in model formulations and/or initial conditions. Despite best efforts, the ensemble forecasts can still be subject to biases and dispersion errors. Statistical postprocessing corrects these systematic shortcomings and releases the full potential of the ensemble. This work focuses on two aspects of statistical postprocessing: incorporating spatial dependency structure into the probabilistic forecast and the choice of an adequate training/verification set for the postprocessing model.

Many real-world applications of statistical postprocessing benefit from modeling of dependencies – e.g. spatial, temporal or inter-variable. The majority of the pioneering postprocessing approaches did not address this need. Here, we extend the well-established postprocessing method Ensemble Model Output Statistics (EMOS) with a Gaussian random field that models global predictive errors. Indicated by the characteristics of the forecast errors, the covariance function of this random field is assumed to be non-stationary, accounting for land-water differences in predictive ability and correlation length. In case studies, we apply this spatial postprocessing methods to 2m temperature forecasts by The Interactive Grand Global Ensemble (TIGGE), as well as the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble, and compare their forecast skill to the reference standards ensemble copula coupling (ECC) and the Schaake shuffle.

When employing statistical postprocessing methods, it is critical to choose appropriate verification data to train and assess the methods. Observation sites are non-homogeneously scattered across the globe, but yield truth data that are independent of the prediction system. Covering the entire Earth on a grid, (re)analyses combine past forecasts and observations and are available on the same spatio-temporal resolution as the forecasting model. Here, we contrast the benefits of postprocessing at observation sites to postprocessing against gridded reanalyses. In a case study, we apply EMOS to 2m ECMWF temperature forecasts, trained and assessed using both verification sets.

Zusammenfassung

Qualitativ hochwertige Vorhersagen sind für eine fundierte Entscheidungsfindung unerlässlich. Dies gilt insbesondere in der Meteorologie, da Wetterphänomene hohe sozioökonomische Kosten verursachen können. In den letzten Jahrzehnten hat sich das Paradigma in der Wettervorhersage von Punktvorhersagen zu probabilistischen Vorhersagen verlagert, welche versuchen die Unsicherheit der Vorhersage mit einer Wahrscheinlichkeitsverteilung zu erfassen. Operativ werden diese probabilistischen Vorhersagen durch Ensembles erzeugt, die sich aus mehreren Läufen von numerischen Wettervorhersagesystemen mit unterschiedlichen Modellformulierungen und/oder Anfangsbedingungen zusammensetzen. Trotz größter Bemühungen können die Ensemble-Prognosen nach wie vor Verzerrungen und Dispersionsfehlern unterliegen. Statistische Nachbearbeitung korrigiert diese systematischen Mängel und setzt so das volle Potenzial des Ensembles frei. Diese Arbeit konzentriert sich auf zwei Aspekte der statistischen Nachbearbeitung: die Einbeziehung räumlicher Abhängigkeitsstrukturen in die probabilistische Vorhersage und die Wahl eines geeigneten Datensatzes zum Trainieren und Verifizieren des Nachbearbeitungsmodells.

In der statistischen Nachbearbeitung profitieren viele Anwendungen von der Modellierung der Abhängigkeitsstrukturen – z.B. räumliche, zeitliche oder inter-variable. Die Mehrheit der grundlegenden Arbeiten auf diesem Gebiet adressierte diesen Bedarf nicht. Hier kombinieren wir die etablierte Nachbearbeitungsmethode Ensemble Model Output Statistics (EMOS) mit einem Gaußsches Zufallsfeld, um globale Vorhersagefehler zu modellieren. Aufgrund der Charakteristiken der Vorhersagefehler wird angenommen, dass die Kovarianzfunktion dieses Zufallsfeldes nicht-stationär ist, um Land-Wasser-Unterschiede in der Vorhersagefähigkeit und Korrelationslänge zu berücksichtigen. In Fallstudien wenden wir diese räumlichen Nachbearbeitungsmethoden auf 2m-Temperaturvorhersagen des Interactive Grand Global Ensemble (TIGGE) sowie des Ensembles vom Europäischen Zentrum für mittelfristige Wettervorhersage (EZMW) an und vergleichen ihre Vorhersagequalität mit den Referenzstandards Ensemble Copula Coupling (ECC) und Schaake Shuffle.

Bei der Anwendung von statistischen Nachbearbeitungsmethoden ist es essenziell, geeignete Verifikationsdaten zu wählen, um die Methoden zu trainieren und auszuwerten. Beobachtungsstationen sind inhomogen über den Globus verteilt, liefern aber Daten, die unabhängig vom Vorhersagesystem sind. Im Vergleich dazu werden durch Kombination von vergangenen Vorhersagen und Beobachtungen (Re-)Analysedaten erzeugt, die in derselben räumlichen Auflösung wie das Vorhersagemodell verfügbar sind. Hier vergleichen wir die Verbesserung durch statistische Nachbearbeitung an Beobachtungsstationen mit der gegenüber Reanalysedaten. In einer Fallstudie wenden wir EMOS auf 2m-Temperaturvorhersagen des EZMW an, die mit beiden Verifikationsdatensätzen trainiert und ausgewertet wurden.

Acknowledgments

My deepest gratitude goes towards Prof. Dr. Martin Schlather and Prof. Dr. Tilmann Gneiting for the supervision of this project. Through their guidance and support I was able to deepen my understanding of the challenges in probabilistic forecasting and far beyond this subject. I am especially indebted to Tilmann Gneiting for his excellent overseeing, quick replies and contagious enthusiasm for research.

In the course of my doctoral studies I had the opportunity to learn from many colleagues. In particular, I thank Mikyoung Jun for sharing her knowledge of statistical processes on spheres during her stay at Heidelberg University. As a visiting scientist, I had the opportunity to stay at European Centre for Medium-Range Weather Forecasts (ECMWF). My gratitude goes to Florian Pappenberger, David Richardson, and Paul Smith for their hospitality and fruitful discussions. Furthermore, I am indebted to former and current members as well as associates of the CST group at Heidelberg Institute for Theoretical Studies – namely Werner Ehm, Timo Dimitriadis, Jochen Fiedler, Stephan Hemri, Alexander Jordan, Fabian Krüger, Johannes Resin, Roman Schefzik, Michael Scheuerer, Patrick Schmidt, Nina Schuhen, Thordis Thorarinsdottir, Peter Vogel. I thank you for helpful ideas, sharing code or just entertaining discussions.

Through Tilmann Gneiting this work was funded by the European Union Seventh Framework Programme as well as ECMWF Fellowship Programme. Furthermore, I acknowledge the support by Research Training Group 1953 “Statistical Modeling of Complex Systems and Processes”, Heidelberg University, Mannheim University, Klaus Tschira Foundation and Heidelberg Institute for Theoretical Studies.

Finally, I am grateful for the enduring support of my family and friends during this time.

Statement on Publications

This thesis contains excerpts from the journal article Feldmann et al. (2019) and a report commissioned by the European Center for Medium-Range Weather Forecasts on spatial calibration. I performed the statistical analyses for both of these myself and was involved in their writing. Chapter 5 displays the report and Chapter 6 is based on the journal paper.

List of Abbreviations

AE	Absolute error
BC	Bias corrected
BMA	Bayesian model averaging
BOM	Bureau of Meteorology
CDF	Cumulative distribution function
CMA	China Meteorological Administration
CMC	Canadian Meteorological Centre
CRPS	Continuous rank probability score
CPTEC	Centro de Previsa de Temp e Estudos
DSS	Dawid-Sebastiani score
DM	Diebold-Mariano
ECC	Ensemble copula coupling
ECCC	Environment and Climate Change
ECMWF	European Centre for Medium-Range Weather Forecasts
EMOS	Ensemble model output statistics
EPS	Ensemble prediction system
ERA	ECMWF's ReAnalyses
ES	Energy score
EW	East-west
EZMW	Europäischen Zentrum für mittelfristige Wettervorhersage
GCA	Gaussian copula approach
GRF	Gaussian random field
IGN	Ignorance score
JMA	Japan Meteorological Agency
KMA	Korea Meteorological Administration
LS	Land/sea
MVN	Multivariate normal
NCEP	National Centers for Environmental Prediction
NCMRWF	National Center for Medium Range Weather Forecast
NN	Nearest neighbor
NS	North-south
NWP	Numerical weather prediction
PIT	Probability integral transform
Q	Equidistant quantiles
R	Random draw
RF	Random field
RMSE	Root mean squared error

UKMO	United Kingdom Meteorological Office
uPIT	Unified PIT
UTC	Universal time coordinated
VRH	Verification rank histogram
VS	Variogram score
WMO	World Meteorological Organization
THORPEX	The Observing System Research and Predictability Experiment
TIGGE	The Interactive Grand Global Ensemble

Contents

Chapter 1

Introduction

Information about the uncertainty of a prediction are essential for informed decision-making. As weather phenomena have a high socioeconomic impact, adequate decisions based on skillful forecasts are essential for humankind. Access to the uncertainty associated with the prediction enables individuals to improve the quality of their decisions, as shown experimentally by Joslyn and LeClerc (2012). However, the predictive uncertainty ought to be precise. Statistical postprocessing provides powerful tools to improve numerical weather prediction (NWP) models by delivering skillful probabilistic forecasts. This thesis focuses on two aspects of statistical postprocessing: modeling spatial correlation of the probabilistic forecast and the selection of a suitable training/verification set for the postprocessing model.

1.1 Temperature forecasts and ensembles

Accurate temperature forecasts are critical for a variety of applications – among others power demand and generation. In an increasingly energy-demanding world, solar power is a sustainable alternative to fossil fuel. In 2020, the share of solar power in German electricity generation amounts to about 9%, making solar energy one of the fastest growing renewable energy sources over the past years (Statistisches Bundesamt, 2021). When collecting solar power, accurate weather predictions are essential to further extend its utilization and improve the competitiveness of this energy source.

Here, we put the emphasis on medium-range predictions with lead times up to 15 days. Usually, forecasts at this scope are based on the output of NWP models, which use physical representations to describe the atmospheric development based on its current state. These models comprise sets of nonlinear partial differential equations for which

no analytical solution exists. Instead, they are solved numerically on a global grid or over certain spatial regions.

At the beginning of the 20th century, NWP originated when Bjerknes (1904) proposed to describe the evolution of the atmosphere by a set of differential equations based on the governing equations of fluid dynamics. The initial conditions for the differential equations are provided by the current state of the atmosphere – namely observations for various weather variables such as pressure, temperature and humidity. Two decades later, Lewis Fry Richardson produced the first forecast in such manner by manually computing a hindcast for 6-hours ahead surface pressure (Richardson, 1922). Although the forecast was highly inaccurate and the calculation time far beyond the prediction horizon (Lynch, 2006), his dream of a forecast factory was born in which a large number of humans would produce weather forecasts by calculating the future state of the atmosphere (Richardson, 1922).

This fantasy became tangible when mathematician John von Neumann proposed to use computers for weather prediction. In 1950, the group led by meteorologist Jule Charney at Princeton University, first successfully implemented dynamical weather prediction models on the multi-purpose digital computer ENIAC and issued a prediction for 500hPa geopotential height over North America (Lynch, 2008). Due to the advances of computing power in the middle of the century, computation time reduced to less than the forecast period itself and the first forecasting system went operational in 1954 (Harper et al., 2007). From here onward, NWP models have underwent tremendous improvements in the following decades, which can be attributed to new atmospheric models, increased computational capacity, and greater availability of observational data for data assimilation.

Historically, weather prediction applied the concept by Bjerknes: A system of differential equations is run forward in time to produce a deterministic forecast of the future state of the atmosphere. This procedure however comprises two sources of uncertainty (Leutbecher and Palmer, 2008): It is humanly impossible to perfectly characterize the current state of the atmosphere required for the initial conditions. Furthermore, the model formulation itself is based on approximations of highly complex physical processes and can be subject to inaccurate numerical schemes. Thus uncertainty quantification in NWP is inevitable.

Lorenz (1963) first challenged the deterministic approach to forecasting by demonstrating that the solutions to a system of nonlinear differential equations are highly sensitive to the initial conditions. Small differences in the current description of the atmosphere, when run forward in time, can lead to strongly deviating solutions. Thus, weather is an example of a deterministic chaotic system, in which despite best efforts perfect predictions are impossible.

To quantify the uncertainty of the initial conditions, Epstein (1969) suggested the use of probability distributions within a stochastic-dynamic approach and generated multiple forecasts through the application of Monte Carlo simulations. Although computationally highly demanding and rather impracticable (Lewis, 2005), Leith (1974)

built upon Epstein's proposition and found that the predictive mean can be accurately approximated with a sample size of 8; but much larger sample sizes are needed for the estimation of higher moments. This approach can be considered an initial ensemble prediction system (EPS).

For the following two decades nevertheless, NWP had still been viewed as a strictly deterministic problem in which the single best model paired with the best input data would issue the best forecast (Gneiting and Raftery, 2005). However, the paradigm in the meteorological community shifted, when the first EPSs were launched in 1992 by the European Centre for Medium-Range Weather Forecasting (ECMWF) and National Centers for Environmental Prediction (NCEP) – see, inter alia, Molteni et al. (1996) and Tracton and Kalnay (1993), respectively. An EPS consists of multiple runs of a NWP model – each with slight variations in the model formulation or specification of initial conditions or both. Instead of a single, deterministic point forecast, an ensemble delivers a set of predictions which can be viewed as a sample from the underlying predictive distribution. Through this setup, an ensemble addresses the two sources of uncertainty in weather forecasting raised earlier.

While ensembles play a major role in the transition from deterministic to probabilistic forecasting, they exhibit some shortcomings. Ultimately, they provide a finite sample and do not deliver a full predictive distribution. Furthermore, they can be subject to biases and tend to underestimate the true uncertainty of the prediction (Hamill and Colucci, 1997). As argued before, an accurate quantification of the uncertainty associated with the forecast is essential in many applications to allow for high quality decisions. Statistical postprocessing releases the full potential of an ensemble forecast by correcting the systematic biases and providing an adequate description of the underlying uncertainty (Gneiting and Raftery, 2005). During the last two decades, a vast variety of statistical postprocessing methods have been developed following the pivotal work of Hamill and Colucci (1997).

State-of-the-art univariate postprocessing methods include Bayesian Model Averaging (BMA; Raftery et al., 2005) and Ensemble Model Output Statistics (EMOS; Jewson et al., 2004 and Gneiting et al., 2005), which are presented in Chapter 2. Both techniques model the future distribution of a weather variable through parametric families of probability distributions. The BMA procedure uses mixture distributions; each ensemble member is assigned a kernel function, which reflects the past skill of that member. Within a regression setting, EMOS fits a parametric distribution function, based on summary statistics of the ensemble. For both models, parameters are statistically estimated over a training period containing past observations and forecasts. To account for global data, we propose different constructions of these training sets – allowing for data from neighboring grid points or grid points on a similar topography only. Employing different distribution families or mixtures thereof, BMA and EMOS are applicable to a variety of weather variables (see for instance Sloughter et al., 2007; Baran and Lerch, 2015 or Scheuerer and Hamill, 2015a).

When evaluating probabilistic forecasts, there are two important concepts to assess their skill: Sharpness describes the concentration of the predictive distribution, while calibration refers to the statistical consistency between the verifying observation and the forecasts. Gneiting et al. (2007) established the objective in probabilistic forecasting to maximize sharpness subject to calibration. We will end the second chapter with a collection of tools that address this notion and thereby assess the quality of univariate probabilistic predictions.

Many of the initial postprocessing methods are univariate, meaning they solely apply to a single weather variable, at a single location or for a single prediction horizon. However, spatial dependence is of crucial importance when producing realistic probabilistic forecast fields. To achieve this, we combine EMOS with a Gaussian random field (GRF) model in Chapter 3. Although conceptually similar to the approach by Feldmann et al. (2015), our goal is to model the correlation structure of EMOS error fields globally. When choosing a covariance for the GRF on the sphere, it is crucial to use an appropriate distance metric to maintain positive definiteness of the covariance function. While great circle distances allow for a more realistic approximation of the Earth's silhouette, the Euclidean norm is often used in applications as it grants access to the rich class of established covariance models on $\mathbb{R}^3$ (Yadrenko, 1983). For the latter, physically unrealistic distortions occur in particular at larger distances. Since the correlation length of the considered predictive errors is rather short, we use the Euclidean norm. As the dependence structure of the error fields changes across planet Earth, we assume a nonstationary covariance function, based on works by Stein (2005). Allowing for land-water differentials in predictive standard deviations and correlation lengths, by virtue of this function we generate physically realistic, calibrated, probabilistic forecasts globally. The chapter ends with a presentation of tools to assess multivariate probabilistic forecasts ranging from histograms to scoring rules.

After the presentation of postprocessing techniques, Chapter 4 displays applications to predictions by the unique, global, multi-model ensemble TIGGE (The Interactive Grand Global Ensemble; Bougeault et al., 2010), which merges forecasts from twelve globally-operating weather centers. In an early experiment, Hagedorn et al. (2012) show first verification results for simply postprocessed TIGGE temperature forecast on the Northern Hemisphere. After successfully reproducing their study, we expand the data set to cover the globe, assess the skill of each of the contributing forecast ensembles individually and refine the forecasts' quality further by the application of different variations of the more sophisticated EMOS postprocessing approach. To account for spatial dependencies, we employ the methods presented in Chapter 3 and compare their predictive performance.

The most skillful contributor to the TIGGE project is the ECMWF medium-range ensemble, also documented in Buizza et al. (2005) or Hagedorn et al. (2012). In a case study in Chapter 5, we explore potential benefits through statistical postprocessing for this EPS individually. Univariate postprocessing methods are applied to ECMWF global temperature forecasts for lead times from 1 up to 15 days. For longer prediction

horizons, the forecast uncertainty increases, the dispersion errors significantly reduce and thus the benefits through postprocessing continuously diminish. Due to conservation of computational resources and this finding, we restrict the data to the subset of 3-days ahead forecasts when applying spatial postprocessing techniques, introduced in Chapter 3. Furthermore, the univariate verification results suggest assuming temperature to be normally distributed may not be ideal and a distribution with heavier tails might be more appropriate for this data. To investigate further, we employ a logistic distribution and the Student's t-distribution, suggested by Gebetsberger et al. (2017), to 5-days ahead forecasts over Europe.

When training and assessing statistical postprocessing methods, it is crucial to choose appropriate verification data: Observation sites are scattered inhomogeneously across planet Earth, whereas (re)analyses cover the entire globe on the same spatio-temporal scale as the forecasting model. In Chapter 6, we provide a systematic comparison of the effects of this choice for both raw and statistically postprocessed temperature predictions from the ECMWF ensemble system. In a study of ECMWF ensemble forecasts for surface wind speed, Pinson and Hagedorn (2012) compare the predictive quality of grid-based and station-based forecasts. They find that the predictive performance is superior in the grid-based data set which can be attributed to the absence of representativeness error and subgrid variability. However, Pinson and Hagedorn (2012) restrict attention to the raw ensemble forecast and do not consider benefits through statistical postprocessing. In Chapter 6, we aim to close this gap in the extant literature.

For the case study, reanalyses by the ECMWF, namely ERA5 (Hersbach et al., 2020), and 9,103 World Meteorological Organization (WMO) stations worldwide constitute the two verification sets. As in the previous chapter, gridded forecasts are provided by the 50-member ECMWF ensemble and bi-linearly interpolated to the observation sites. We apply the EMOS approach to the gridded forecasts paired with the reanalyses and the bi-linearly interpolated predictions paired with the observational data. Analysis-based postprocessing can enhance forecasts at lead times up to twelve days, whereas forecasts at all lead times benefit from postprocessing when trained and verified against observations. Overall, we conclude that more forecast skill can be gained through data collected at observational sites.

The thesis ends with Chapter 7 in which we summarize the findings and point towards future work.

1.2 Mathematical framework

Based in measure theory, a *prediction space* is a probability space designed to study probabilistic forecasts. In the framework of point predictions, Murphy and Winkler (1987) first advocated considering the joint distribution of forecast and observation. Following Gneiting and Ranjan (2013), who initially expanded the concept to distributional predictions, a prediction space is a probability space

$$(\Omega, \mathcal{A}, \mathbb{Q}),$$

in which for distributional forecasts $F_1, ..., F_k$ with integer $k \geq 1$ for a real-valued outcome Y, the elements of the sample space Ω can be identified with the tuplets

$$(F_1, ..., F_k, Y).$$

Each probabilistic forecast $F_1, ..., F_k$ is measurable with respect to the sub-σ-fields $\mathcal{A}_1, ...\mathcal{A}_k \subseteq \mathcal{A}$, that contain the information a forecast is built upon. In practice, the distributional predictions $F_1, ..., F_k$ can be issued by different experts, statistical models or institutions. The joint distribution of predictions and observations is described by the probability measure $\mathbb{Q}$ on the measurable space $(\Omega, \mathcal{A})$. All theoretical concepts for probabilistic forecasts in the subsequent chapters are based within this prediction space setting.

Chapter 2

Univariate statistical postprocessing

In numerical weather prediction, the two major sources of uncertainty in the initial conditions and model formulations are not completely addressed by the introduction of ensemble prediction systems, as discussed in Chapter 1. The remaining deficiencies of the ensemble are of the form of biases as well as dispersion errors and call for statistical postprocessing techniques to release the full potential of the EPS. A variety of postprocessing methods have been developed following the pivotal study by Hamill and Colucci (1997). These approaches can be classified in parametric, which assume the predictive distribution to follow a pre-selected distribution family, and a non-parametric, based on non-parametric approximation of the predictive distribution. Here, we put the focus on parametric models, where the parameters of the predictive distribution are linked to ensemble summary statistics. Generally, within a regression framework, the parameters are selected which optimize a loss function over a training period containing past forecasts and observations. Wilks (2018) and Vannitsem et al. (2021) give comprehensive overviews of current procedures, which are commonly used to improve the skill of probabilistic or deterministic forecasts.

The aim of statistical postprocessing is to maximize the sharpness of the forecasts subject to calibration (Murphy and Winkler, 1987; Gneiting et al., 2007). We discuss this notion further in this chapter and present the most commonly used univariate approaches designed to achieve this goal. Univariate in this context refers to postprocessing of weather forecasts at each location, for each lead time and for each weather variable independently. In Chapter 3 we will explore applications to entire weather fields instead of individual locations.

This chapter begins with a review of Bayesian Model Averaging (BMA; Raftery et al., 2005) and Ensemble Model Output Statistics (EMOS; Jewson et al. 2004; Gneiting et al. 2005). Here, we focus on variants applicable to temperature forecasts. Both of the presented methods rely on the idea that characterizations of forecast errors in the past will correct and hence improve future forecasts. In particular, we will explore different approaches on how to construct the training data for these techniques. Subsequently, we will review spatially adaptive EMOS by Hemri et al. (2014) and simple bias correction as reference forecasts which set benchmarks for the models to compete with. The chapter closes with a presentation of verification tools for univariate forecasts.

2.1 Bayesian Model Averaging

BMA is a broadly applied statistical approach for combining competing statistical models, in particular predictive distributions stemming from different forecasting sources. Instead of restricting the predictive distribution to a certain parametric shape, BMA delivers a mixture distribution which comprises of a weighted sum of the distributions from the contributing models. In the context of postprocessing for weather variables, we follow Raftery et al. (2005) to calibrate forecast ensembles.

At location $s \in \mathcal{S}$, we make predictions for a weather variable y_s with the M-member ensemble $f_{1,s}, \ldots, f_{M,s}$, where M is a natural number. Each forecaster is associated with a conditional probability density function $p_m\left(y_s|f_{m,s}\right)$, which can be interpreted as the conditional density of $y_s|f_{m,s}$, given that $f_{m,s}$ is the most skilled forecaster $m \in M$ in the ensemble. Then BMA stipulates a predictive density of the form

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

with $w_1, \ldots, w_M$ non-negative weights summing up to 1. These weights reflect the forecasters' relative performance during the training period. A large weight indicates skillful forecasts; whereas a small weight is assigned to poorly performing members.

For temperature forecasts, a Gaussian or normal distribution is commonly applied, and we write $\mathcal{N}(\mu, \sigma^2)$ to denote a normal distribution with mean μ and variance σ^2. Then for BMA, the conditional probability density functions $p_m\left(y_s|f_{m,s}\right)$ are univariate normal densities

$$y_s|f_{m,s} \sim \mathcal{N}(a_m + b_m f_{m,s}, \sigma^2),$$

where the mean equals each member's bias corrected forecast $a_m + b_m f_{m,s}$ and the variance σ^2 is the same for all members. The coefficients a_m and b_m are real-valued.

The BMA predictive mean is a weighted average of the bias corrected forecasts

$$\mathbb{E}\left(y_{s,}|f_{1,s},...,f_{M,s}\right) = \sum_{m=1}^{M} w_m \left(a_m + b_m f_{m,s}\right),$$

and the predictive variance is

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

The variance can be decomposed into two terms: The first sum represents the between-forecasts spread and σ^2 summarizes the within-forecast variance.

In case of an exchangeable ensemble – meaning that the contributing members stem from the same underlying model and are thus statistically not distinguishable – the weights w_m and the regression parameters a_m and b_m should be the same for every member $m = 1, ..., M$. This procedure is described in more detail in Fraley et al. (2010).

The BMA parameters are estimated in multiple steps. First we define a rolling window training period T consisting of forecasts $f_{1,s,t}, \ldots, f_{M,s,t}$ and verifying observations $y_{s,t}$ for past days $t \in T$. Then each member is bias corrected individually via simple linear regression of $y_{s,t}$ using $f_{k,s,t}$ as predictor and the member-specific parameters a_m and b_m for $m = 1, \ldots, M$ are estimated. Based on the assumption that the forecast errors are independent in space and time, the log-likelihood functions equals

$$l(w_1, \ldots, w_M, \sigma^2) = \sum_{t \in T, s \in \mathcal{S}} \log \left(\sum_{m=1}^{M} w_m p\left(y_{s,t}|f_{m,s,t}\right) \right).$$

By maximizing this function, the variance σ^2 and weights w_m are obtained. Because the maximum cannot be determined analytically, Raftery et al. (2005) use the expectation–maximization algorithm (Dempster et al., 1977). Finally the estimate for σ^2 may be refined by minimizing the continuous rank probability score (CRPS – see Section 2.4.2 for details) over the training period. BMA for ensemble forecasts can be implemented using the R package `ensembleBMA` by Fraley et al. (2011).

The assumption of a normal distribution is not suitable for weather variables that have a e.g. skewed distribution or only take non-negative values. For these cases, BMA has been further refined and many extensions have been published. Sloughter et al. (2010), for example, adjust the model for the needs of wind speed by employing a gamma distribution, while Baran (2014) proposes a truncated Gaussian distribution for this weather variable. Additionally, Bao et al. (2010) put forward a von Mises distribution to model wind direction. Precipitation forecasts consist of two components – a continuous distribution for a positive amount of precipitation and a discrete probability for precipitation being equal to zero. Sloughter et al. (2007) model the point mass

at zero with logistic regression and the rainfall amount with a gamma distribution. Schmeits and Kok (2010) refined this approach and Bentzien and Friederichs (2012) propose a regression mixture distribution in this setting. Furthermore, Roquelaure and Bergot (2008) extend BMA for forecasts of fog and Chmielecki and Raftery (2011) of visibility.

2.2 Ensemble Model Output Statistics

The EMOS method (Jewson et al., 2004; Gneiting et al., 2005) can be interpreted as a distributional regression technique (Gneiting and Katzfuss, 2014) and is thus also referred to as non-homogeneous Gaussian regression, a term that dates back to Wilks (2006). EMOS yields a parametric probability distribution, where the statistical parameters are linked to the ensemble forecast at hand. As we consider forecasts for the entire sphere, the biases and dispersion errors may differ considerably across the globe. So we review a local version of the EMOS approach, where we estimate all parameters at each site individually.

For temperature as for BMA, a Gaussian or normal distribution is typically applied. At any location $s \in S$, and for a generic ensemble with M members, EMOS stipulates a predictive distribution of the form

$$y_s | f_{1,s}, \ldots, f_{M,s} \sim \mathcal{N}(a_s + b_{1,s} f_{1,s} + \cdots + b_{M,s} f_{M,s}, c_s + d_s v_s^2), \tag{2.1}$$

where y_s is the future temperature and $f_{1,s}, \ldots, f_{M,s}$ are the ensemble member forecasts. The predictive mean is a weighted average of the contributing ensemble members and the predictive variance is a linear function of the ensemble variance v_s^2 with spread parameters $c_s > 0$ and $d_s \geq 0$. The coefficients $a_s, b_{1,s}, \ldots, b_{M,s}$ can take any value in $\mathbb{R}$. Since negative weights $b_{1,s}, \ldots, b_{M,s}$ for members are difficult to interpret, we occasionally restrict the estimates to be non-negative and refer to this technique as EMOS$^+$. In case of a forecast ensemble with exchangeable members, the above equation simplifies to

$$y_s | f_{1,s}, \ldots, f_{M,s} \sim \mathcal{N}(a_s + b_s \bar{f}_s, c_s + d_s v_s^2). \tag{2.2}$$

with $\bar{f}_s$ being the mean of the ensemble.

The parameter estimates are based on a rolling training period and obtained by minimizing the CRPS, as proposed in Gneiting et al. (2005). In the case of a normal distribution, the CRPS has a closed form. Let $\mu(a_s, b_{1,s}, \ldots, b_{M,s})$ be the predictive mean, $\sigma(c_s, d_s)$ the predictive standard deviation and $\theta = (a_s, b_{1,s}, \ldots, b_{M,s}, c_s, d_s)$ the parameter vector of interest. Over the training period T of length $|T|$, we write at each site $s \in \mathcal{S}$

$$\Gamma(\theta) = \frac{1}{|T|} \sum_{t \in T} \sigma(c_s, d_s) \Bigg\{ \frac{y_s - \mu(a_s, b_{1,s}, \ldots, b_{M,s})}{\sigma(c_s, d_s)} \left[2\Phi \left(\frac{y_s - \mu(a_s, b_{1,s}, \ldots, b_{M,s})}{\sigma(c_s, d_s)} \right) - 1 \right] + 2\varphi \left(\frac{y_s - \mu(a_s, b_{1,s}, \ldots, b_{M,s})}{\sigma(c_s, d_s)} \right) - \frac{1}{\sqrt{\pi}} \Bigg\}$$

and find the vector θ which minimizes this expression. The predictive probability density and cumulative distribution function (CDF) of the normal distribution are denoted by φ and Φ, respectively.

Alternatively, θ can be determined by maximizing the log-likelihood function

$$l(\theta) = -\frac{1}{2} \left\{ |T| \log(2\pi) + \sum_{t \in T} \frac{(y_s - \mu(a_s, b_{1,s}, \ldots, b_{M,s}))^2}{\sigma(c_s, d_s)} + \sum_{t \in T} \log\left(\sigma(c_s, d_s)\right) \right\},$$

which is equivalent to minimizing the logarithmic (Good, 1952) or ignorance score (Roulston and Smith, 2002). Maximum likelihood is computationally faster and statistically efficient under correct EMOS specifications (Gebetsberger et al., 2018). In this thesis we apply both methods depending on the scenario at hand.

Since its introduction, EMOS has been modified for different weather variables which cannot be modeled by a normal distribution. Some approaches, like Hemri et al. (2015), first transform the predictand and predictor such that they can be considered Gaussian to then apply conventional EMOS. Baran and Lerch (2015; 2016) employ nonhomogeneous lognormal regression to predict wind speed, while Messner et al. (2014) transform wind speed by its square root to then model the transformed weather variable with a logistic distribution.

Some weather variables require non-symmetrical or truncated distributions. Thorarinsdottir and Gneiting (2010), e.g., model future wind speed with a zero-truncated normal predictive distribution, while Scheuerer and Möller (2015) propose a truncated logistic distribution for the same purpose. Lerch and Thorarinsdottir (2013) and Baran and Lerch (2015) investigate regime-switching models, where under alternating conditions different distributions are employed.

In order to fit characteristics of the weather variable at hand, censoring can be used to adapt distributions. Scheuerer (2014) further refined the EMOS model for precipitation forecasts with a zero-censored generalized extreme value distribution. Scheuerer and Hamill (2015a) and Baran and Nemoda (2016) model precipitation amounts with a zero-censored shifted-gamma predictive distribution. EMOS can easily be implemented via the R package `ensembleMOS` by Yuen et al. (2018).

2.2.1 Spatially adaptive EMOS

This spatially adaptive variant of the original EMOS model was put forward by Hemri et al. (2014). Based on the approaches by Scheuerer and Büermann (2014) and Scheuerer and König (2014), the authors employ local anomalies of the observations and forecasts as regression predictors and predictands. In more detail, these anomalies are defined as the difference of the verifying observations or forecasts from their estimated historical trend. So spatially adaptive EMOS yields a local bias correction, while the parameters for the predictive mean remain constant across the domain. While Hemri et al. (2014) incorporate information from the ensemble mean and specific members, we present a simplified version solely for the mean to fit the needs of the data available to us in Chapter 4.

For each location $s \in \mathcal{S}$, Hemri et al. (2014) estimate the parameters $g_{0,s}$, $g_{1,s}$ and $g_{2,s}$ over the training period T of length $|T|$ via least squares regression

$$y_s = g_{0,s} + g_{1,s} \sin\left(\frac{2\pi t}{365}\right) + g_{2,s} \cos\left(\frac{2\pi t}{365}\right) + \varepsilon_s,\ t = 1, \ldots, |T|.$$

This expression models the seasonal variation in y_s at each site and can easily be extrapolated into the future by $t > |T| \in \mathbb{N}$. The same model is fitted to the ensemble mean $\bar{f}_s$, obtaining $\tilde{f}_s$. Let $\tilde{y}_s$ be the local estimated climatology, then

$$\mu_s = \tilde{y}_s + a\left(\bar{f}_s - \tilde{f}_s\right)$$

describes the predictive mean. As in the original EMOS, the predictive variance is a linear combination of the ensemble's variance v_s^2,

$$\sigma_s^2 = c_s + d_s v_s^2.$$

Dabernig et al. (2017) further develop this approach by standardizing the observation and forecast anomalies to remove site-specific and seasonal characteristics.

2.2.2 Spatial augmentation of the training period

In the original approach, EMOS parameters are estimated and kept constant over the entire domain. As the spatial structure of temperature forecasts differs considerably across the globe, this is not an appropriate assumption for the case studies conducted in Chapter 4, 5 and 6. Hence, we mainly apply the so-called local version of EMOS, for which training data and parameters are location-specific. With the combined goals of improving estimation and increasing stability of the estimated statistical parameters across geographic space, we explore simple approaches to augment the training data for one location in space.

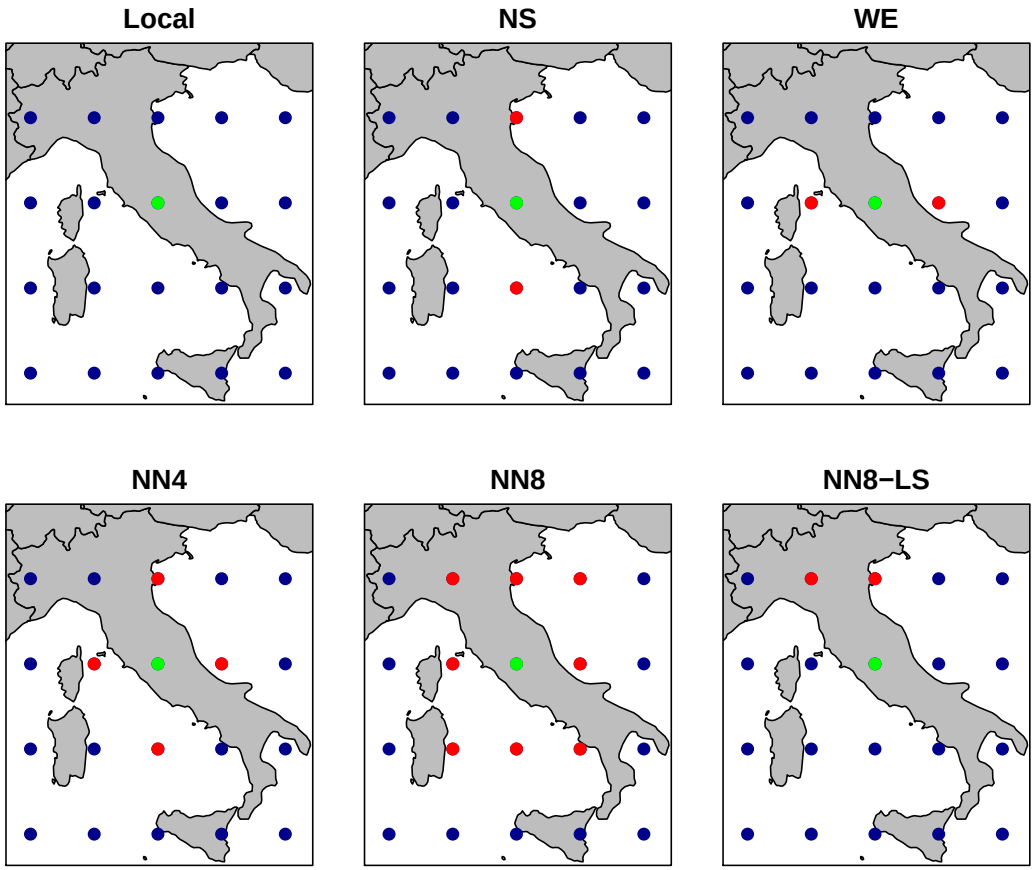

Figure 2.1: Spatial composition of training sets in Local, NS, WE, NN4, and NN8 techniques, with potential further consideration of surface types (land/sea: LS) illustrated over Italy. For estimating the statistical parameters for the EMOS model at the central (green) grid point, recent data from adjacent (red) grid points are included in the training set, in addition to data at the central grid point.

Different techniques have been developed to augment the training set based on geography. Hamill et al. (2008) propose to calculate distances between stations and only add data from stations with small distances to the set. Lerch and Baran (2017) further explore this idea by defining distance functions on different characteristics such as geographical distance and distribution of observations and of forecast errors. As the application of these sophisticated approaches reduces the forecast skill when compared to local EMOS for our data set in Section 4.1.1, we propose a rather simple geographical or grid-based approach.

In contrast to the standard local technique, which uses training data from the grid point at hand only, our neighborhood techniques augment the training sets with data from adjacent grid points, as illustrated in Figure 2.1. Specifically, when estimating parameters for the central (green) grid point, we add data from surrounding (red) grid

points to the training set. For the north-south (NS) and west-east (WE) variant, the two closest grid points in the respective direction are included. Data from all of these four grid points are added to the training set of the four nearest neighbors (NN4) version. The eight nearest neighbors (NN8) variant includes data from all eight surrounding grid points. For each of these neighborhood variants we consider a land/sea (LS) version, where we only augment the training set with data from grid points that are of the same surface type (land or sea) as the point of interest.

2.3 Reference forecast: Bias correction

As a benchmark for the postprocessing methods, we employ a simple bias correction method described in Hagedorn et al. (2012), where at every site $s \in \mathcal{S}$ a correction term is added to the forecast ensemble. Over the training period, the empirical errors $e_s = y_s - \bar{f}_s$ of the ensemble mean $\bar{f}_s$ relative to the verifying observation y_s is calculated. Then the mean error $\bar{e}_s$, averaged over the whole training period, is added as a correction term to each current ensemble member forecast. Therefore the original ensemble spread remains, while the distribution is moved according to the correction term.

2.4 Univariate verification

Verification measures are essential in order to determine the goodness of NWP model output or the effects of statistical postprocessing. These tools deliver information about forecast skill in terms of calibration and sharpness. As concluded in Murphy and Winkler (1987) and Gneiting et al. (2007), probabilistic forecasts should aim to fulfill the principle of maximizing sharpness subject to calibration.

Calibration refers to a joint property of the prediction and the observation. It is a necessary condition for a valuable forecast, requiring the prediction and the observation to be statistically compatible. In detail, the observation ought to be viewed as a random draw from the predictive distribution or indistinguishable from the ensemble forecasts.

Sharpness is a property of the forecast solely and measures the spread of the ensemble or concentration of the predictive distribution. Forecasts with a smaller spread or less uncertainty are preferred, as they provide more information to the user. Multiple tools have been developed to assess predictive performance and we refer to Wilks (2011), Gneiting and Katzfuss (2014) and Thorarinsdottir and Schuhen (2018) for details.

2.4.1 Sharpness and calibration

Calibration measures the statistical consistency between the observation and the forecasts. Different notions of univariate calibration have been proposed – see Gneiting

et al. (2007), Tsyplakov (2013) or Strähl and Ziegel (2017). Following Dawid (1984), we consider probabilistic calibration in the context of probabilistic weather prediction. Given a probabilistic forecast F and the observation Y, we define the probability integral transform (PIT) as $F(Y) \in [0, 1]$, which equals the value of the predictive CDF at observation Y. If $F(Y)$ is uniformly distributed, then we call F probabilistically calibrated. See Gneiting and Ranjan (2013) for a more sophisticated definition of the PIT in case F has a discrete component.

For CDF-valued probabilistic forecasts, a straightforward tool to check calibration is the PIT histogram (Dawid, 1984; Diebold et al., 1998; Gneiting et al., 2007). Given continuous predictive distributions F_i, $i = 1, ..., n$ and observations y_i, $i = 1, ..., n$, we calculate the PIT values $F_i(y_i)$ over all forecasts cases $i = 1, ..., n$ and plot the corresponding histogram. The shape of the histogram indicates the goodness of the forecasts. A uniform histogram implies calibration, while deviation from this shape suggests miscalibration. Overdispersion is illustrated by a $\cap$-shaped histogram meaning too many observations inhabit the center of the predictive distributions. We call forecasts resulting in a $\cup$-shape underdispersive, implying the forecaster underestimates the uncertainty as many observations lie in the tails of the predictive distribution. Systematic biases manifest in a triangular form. For further details of diagnosing miscalibration see for instance Hamill (2001).

In real life applications, probabilistic forecasts are often provided as an ensemble, which can be interpreted as a random draw from a predictive distribution. To diagnose calibration in this case, the verification rank histogram (VRH) or Talagrand diagram (Anderson, 1996; Hamill and Colucci, 1997; Talagrand et al., 1997) can be used. For an M-member ensemble the rank of the observation within the ensemble is noted, hence taking an integer value between 1 and $M + 1$. These ranks are aggregated; then plotted in a histogram. The interpretation of the VRH histogram coincides with that of the PIT histogram.

When comparing ensembles of different sizes, results via the VRH might be misleading, as the number of bins varies for the corresponding histograms. To evaluate predictive distributions and ensembles with varying sizes, Vogel et al. (2018) introduce the unified PIT (uPIT) histogram. For an M-member ensemble, let the observation take rank i within $1, ..., M + 1$. Then i is mapped to a random draw from the uniform distribution on the interval $\left[\frac{i-1}{M+1}, \frac{i}{M+1}\right]$. Thus the values of the uPIT fall within $[0, 1]$ as for the PIT.

Prediction intervals can be used to assess both sharpness and calibration, via their average width and coverage, respectability. For any real-valued probabilistic forecast, a central $\alpha\%$ prediction interval can be calculated. If a forecast is calibrated the coverage of the prediction interval should be close to the nominal value of $\alpha\%$ with about $\frac{100-\alpha}{2}\%$ of the observations falling to the right and left side of the interval. In case of an M-member ensemble forecast, the nominal value equals $\frac{M-1}{M+1} \cdot 100\%$ for the interval with boarders at the lowest and highest member. Calibrated forecasts with the smallest prediction interval width should be preferred.

2.4.2 Proper scoring rules

Scoring rules are widely used to evaluate probabilistic forecasts by assigning a numerical score to the forecast relative to the observation and thereby providing a summary indicator for skill. These scoring rules should be proper in order to encourage the forecaster to deliver their best prediction and prevent hedging. Proper scoring rules can evaluate sharpness and calibration simultaneously (Gneiting and Raftery, 2007).

Let Ω be a general sample space and $\mathcal{F}$ denote the class of probability measures on Ω. A scoring rule is a function

$$S : \mathcal{F} \times \Omega \to \mathbb{R} \cup \{\infty\},$$

which assigns a score to the predictive distribution $F \in \mathcal{F}$ and observation $y \in \Omega$.

Definition 2.4.2.1. We call a scoring rule *proper* relative to the class $\mathcal{F}$ if

$$\mathbb{E}_G S(G, Y) \leq \mathbb{E}_G S(F, Y)$$

holds for all distributions $F, G \in \mathcal{F}$, where G is the true, but unknown distribution from which $Y \in \Omega$ is sampled. A scoring rule is *strictly proper* if $\mathbb{E}_G S(G, Y) < \mathbb{E}_G S(F, Y)$ for all distributions $F \neq G$.

Propriety ensures a forecaster will deliver their best predictions as only then they will minimize the expected score. Theoretical aspects of proper scoring rules are discussed further in Gneiting and Raftery (2007).

Two commonly used scores are the CRPS and the ignorance or logarithmic score. Both are negatively oriented, meaning that the forecaster should want to minimize the penalty. The ignorance score dates back to Good (1952) and is defined as

$$\mathrm{ign}(F, y) = -\log f(y),$$

where f is the density of F. Hence, it is applicable to continuous distributions only and cannot directly be used for ensemble forecasts. Furthermore the ignorance score is very sensitive to outliers and might rank a forecaster poorly based on a single bad forecast.

For a normal predictive distribution, the value of the ignorance score coincides with the *Dawid-Sebastiani* score (Dawid and Sebastiani, 1999), which is defined as

$$\text{dss}(F, y) = \log \sigma_F^2 + \frac{(y - \mu_F)}{\sigma_F^2}$$

for the predictive distribution F with corresponding mean μ_F and variance σ_F^2.

Matheson and Winkler (1976) introduced the popular CRPS. Further explored in Gneiting and Raftery (2007), Gneiting and Ranjan (2011), Hersbach (2000) and Laio and Tamea (2007), it can be defined in three different, but equivalent ways:

$$\begin{aligned} \text{crps}(F, y) &= \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F|X - X'| && (2.3) \\ &= \int_{-\infty}^{\infty} (F(z)) - \mathbb{I}(z \geq y))^2 dz && (2.4) \\ &= \int_0^1 \left(F^{-1}(x) - y\right) \left(\mathbb{I}\left(y \leq F^{-1}(x)\right) - x\right) dx, && (2.5) \end{aligned}$$

where F is the predictive distribution and F^{-1} its quantile function. $\mathbb{I}$ denotes the indicator function, which equals one if the argument is true and zero otherwise; X and X' are independent random variables with distribution function F and finite first moment. Through Eqs. 2.4 and 2.5, the CRPS is linked to the Brier score (Brier, 1950) and the quantile score (Gneiting and Raftery, 2007; Friederichs and Hense, 2007). Given an ensemble forecast $f_1, ..., f_M$, Eq. (2.3) can be written as (Grimit et al., 2006)

$$\text{crps}(F, y) = \frac{1}{M}\sum_{i=1}^{M}|f_i - y| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}|f_i - f_j|.$$

For various conventional probability distributions, there exist closed forms of the CRPS, which can easily be implemented using the R package `scoringRules` (Jordan et al., 2019).

Deterministic forecasts can be evaluated by a range of different scoring functions $s(x, y)$, which assign a score based on the forecast x and observation y. Among them are the squared error $s(x, y) = (x - y)^2$ and the absolute error $s(x, y) = |x - y|$. When applying these scoring functions to probabilistic forecasts, it is crucial to derive an appropriate point forecast. We rely on consistent scoring functions to avoid misguided inference (Gneiting, 2011).

Technically, a scoring function s is consistent for a functional L relative to a class $\mathcal{F}$ of predictive distributions, if

$$\mathbb{E}_F s(L(F), Y) \leq \mathbb{E}_F s(x, Y)$$

for all x in the sample space Ω and all probability distributions $F \in \mathcal{F}$ (Gneiting, 2011). A scoring function s becomes a proper scoring rule $S(F, y) = s(L(F), y)$ relative to the class $\mathcal{F}$, if s is consistent for the functional L. In case of the above mentioned scoring functions, the absolute error (AE) is consistent for the median,

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

and the squared error for the mean,

$$\text{se}(F, y) = (\text{mean}(F) - y)^2,$$

resulting in proper scoring rules in terms of the median and mean, respectively. In practice we calculate the root mean squared error (RMSE)

$$\text{rmse}(F, y) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \text{se}(F_i, y_i)}$$

where n is the number of available forecasts. All scores reported in the following case studies are averaged over all forecast cases n,

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

Chapter 3

Spatial postprocessing

Statistical postprocessing aims to improve NWP ensemble forecasts by generating calibrated predictive distributions. Many of the early proposed methods are univariate, meaning they apply to a single weather variable at a single prediction horizon and a single site only. However, when forecasting truly multivariate or composite univariate quantities, such as minima, maxima, or averages, the modeling of dependence structures is of great importance. These structures are present in the raw ensemble forecasts, but lost during postprocessing when only the marginal distributions get calibrated. In real world applications, truly multivariate probabilistic forecasts are crucial. For example Hemri et al. (2015) and Scheuerer et al. (2017) model coherent spatio-temporal predictions in hydrology; realistic forecasts for entire wind fields are essential for air traffic control (Chaloulos and Lygeros, 2007); the management of renewable energy requires accurate forecasts of wind and solar resources in terms of location and time (see Pinson et al., 2009; Pinson, 2013; Pinson and Messner, 2018).

Different approaches have been published which aim to capture inter-variable, temporal or spatial dependencies – see Schefzik and Möller (2018) for an overview. In contrast to the methods presented in Chapter 2, these models now yield multivariate predictive distributions, which can either be parametric or non-parametric. Particularly in the case of high-dimensional dependence structures, non-parametric approaches seem suitable as they require low computational efforts. Popular examples comprise the Schaake shuffle (Clark et al., 2004) and ensemble copula coupling (ECC; Schefzik et al., 2013). In some cases, parametric multivariate models can outperform the non-parametric – as e.g. found in Feldmann et al. (2015) for temperature fields and in Schuhen et al. (2012) for wind vectors. Further applications of parametric approaches include forecasts for precipitation fields (Berrocal et al., 2008), for wind vectors (Sloughter et al., 2013; Lang et al., 2019) or inter-variable dependence between wind speed and temperature (Baran and Möller, 2015, 2017).

In our study, we are interested in the generation of physically realistic, calibrated probabilistic forecasts of global temperature, where it is crucial to model spatial dependencies. We utilize Gaussian random fields to represent the EMOS forecast error fields with a nonstationary covariance function accounting for land-water differentials in correlation lengths. Closely related is an approach by Heinrich et al. (2021), who postprocess global sea surface temperature forecasts. Instead of fitting a parametric covariance function to forecast residuals, they model the structure of the error fields with a non-parametric estimate of the sample covariance matrix and apply further restrictions for computational efficiency and interpretability. This allows for a flexible covariance structure which also inspires us to explore our EMOS model paired with an empirical covariance matrix.

In this chapter, we discuss some of the postprocessing techniques mentioned above that account for spatial dependencies to generate calibrated probabilistic forecasts of entire meteorological fields. We begin with the concept of copulas and Sklar's theorem, on which these models are mathematically based. Then we review ECC and the Schaake shuffle as reference standards. Afterwards, we present an example of a nonstationary covariance function for processes on spheres, necessary for the subsequent section, where we discuss a global spatial extension for EMOS, called spatial EMOS. In this more sophisticated postprocessing approach, we utilize Gaussian random fields to represent the EMOS forecast error fields and model spatial dependencies. The chapter ends with a description of numerous tools to assess multivariate probabilistic forecasts.

3.1 Sklar's Theorem

Directly or indirectly, most multivariate postprocessing methods make use of a copula function. A copula $\mathscr{C} : [0,1]^n \to [0,1]$ is an n-variate CDF with $n \in \mathbb{N}$, whose marginal distributions are uniformly distributed on the unit interval $[0,1]$ (Nelsen, 2007). They can be used to model dependence patterns in statistical postprocessing. Sklar (1959) links a copula and marginal distributions to describe a multivariate CDF.

Theorem 3.1.1 (Sklar). Let $F : \mathbb{R} \to [0,1]$ be an n-variate CDF with marginal CDFs $F_1, \ldots, F_n : \mathbb{R} \to [0,1]$. Then there exists a copula $\mathscr{C}$ such that

$$F(y_1, \ldots, y_n) = \mathscr{C}\left(F_1(y_1), \ldots, F_n(y_n)\right). \tag{3.1}$$

If F_i is continuous for $i = 1, \ldots, n$, the copula $\mathscr{C}$ is unique.

Conversely, given marginal CDFs $F_1, \ldots, F_n$ and a copula $\mathscr{C}$, then the function F defined in Eq. (3.1) is an n-variate CDF.

When interested in a multivariate predictive distribution F, Sklar's theorem states that F is characterized by the marginal distributions and the copula. In the context of statistical postprocessing for weather forecasts, we can apply univariate postprocessing to generate predictive distributions $F_1(y_1), \ldots, F_n(y_n)$ for a weather variable y at – in this case – n different locations and then model the dependence structure through the copula $\mathscr{C}$. There are multiple options to specify $\mathscr{C}$, which we will explore further in this chapter.

3.2 Reference forecasts

Due to their simplicity and versatility, the Schaake shuffle and ECC are compelling non-parametric approaches that rely on empirical copulas to construct $\mathscr{C}$. With hardly any computational cost, generating the forecast ensemble under these two methods is rather similar and can be summarized in two steps: First, the forecasts are calibrated in the margins; secondly, the dependence structure is reintroduced via the copula to generate multivariate forecasts. Both models solely differ in the origin of the dependence template: ECC preserves the rank structure of the original ensemble forecasts, while the Schaake shuffle bases the copulas on past observations.

Another multivariate postprocessing technique, closely related to spatial EMOS, is the Gaussian copula approach (GCA), based on works by Pinson and Girard (2012) and Möller et al. (2013). GCA also consists of two separate steps, in which the margins are calibrated individually and then the dependence structure is recovered by a Gaussian copula. In a simulation study, Lerch et al. (2020) conclude that among these three approaches, the Schaake shuffle sets a powerful benchmark for multivariate postprocessing, which is also supported by findings from Wilks (2015) for real-world forecasts (comparing two versions of ECC and the Schaake shuffle). Hence, we focus on the Schaake shuffle and also, due to its low-cost computation, ECC as reference forecasts.

3.2.1 Ensemble copula coupling

Introduced by Schefzik et al. (2013), ECC uses the inherent dependence structure of the ensemble forecasts as a template for the multivariate structure of the prediction. An ECC ensemble that represents spatially coherent multivariate forecasts can be constructed as follows.

1. *Univariate postprocessing*

   In this step, one can apply any univariate postprocessing method to marginally calibrate the forecasts. For coherence with the parametric model in Section 3.4, we employ EMOS (see Section 2.2). Subsequently, at each location $s \in \mathcal{S}$, we draw a random sample $\hat{x}_{1,s}, \ldots, \hat{x}_{M,s}$ from the predictive distribution F_s of the same size $M \in \mathbb{N}$ as the original ensemble. There are several options for sampling from the marginal distributions: the simplest approach is a random draw (ECC-R). After obtaining the sample, it is rearranged in ascending order – for simplicity in notation $\hat{x}_{1,s} \leq \ldots \leq \hat{x}_{M,s}$. Schefzik et al. (2013) recommend generating an ensemble with equidistant quantiles at level $\frac{1}{M+1}, \ldots, \frac{M}{M+1}$, so that the sample comprises of $\hat{x}_{1,s} = F_s^{-1}\left(\frac{1}{M+1}\right), \ldots, \hat{x}_{M,s} = F_s^{-1}\left(\frac{M}{M+1}\right)$, which we refer to as ECC-Q.

2. *Reordering according to dependence template*

   Let $x_{1,s}, \ldots, x_{M,s}$ be the forecast ensemble at location $s \in \mathcal{S}$. We denote the members' ranks by $\omega(1,s), \ldots, \omega(M,s)$; ties within the ranks are resolved at random. To reintroduce the dependence structure of the raw ensemble, we order the sample accordingly, resulting in $\hat{x}_{\omega(1,s)}, \ldots, \hat{x}_{\omega(M,s)}$. Hence, the members of the ECC ensemble equal the vector $\hat{\mathbf{x}}_m = (\hat{x}_{\omega(m,1)}, \ldots, \hat{x}_{\omega(m,|\mathcal{S}|)})$ for $m = 1, \ldots, M$.

By definition, the ECC ensemble is limited to the same size M as the original ensemble. To omit this restriction, the above steps can be repeated multiple times resulting in an ensemble sized nM with $n \in \mathbb{N}$. Wilks (2015) finds that these aggregated ensembles can outperform a smaller ECC ensemble.

The ECC approach can be linked via Sklar's theorem to the empirical copula induced by the raw ensemble. Let $R_1, \ldots, R_{|\mathcal{S}|}$ be the marginal empirical CDFs of the ensemble forecasts at locations in $\mathcal{S}$. These functions take values in the set $I_M = \left\{0, \frac{1}{M}, \ldots, \frac{M-1}{M}, 1\right\}$. The multivariate empirical CDF R of the raw ensemble forecast maps into the same set. According to Sklar's theorem, there exists a copula $\mathscr{C}$ with restriction $E_M : I_M^{|\mathcal{S}|} \to I_M$ such that

$$R\left(y_1, \ldots, y_{|\mathcal{S}|}\right) = E_M\left(R_1\left(y_1\right), \ldots, R_{|\mathcal{S}|}\left(y_{|\mathcal{S}|}\right)\right)$$

for $y_1, ..., y_{|\mathcal{S}|} \in \mathbb{R}$. Hence, the empirical copula E_M connects the univariate distributions $R_1, ..., R_{|\mathcal{S}|}$ to the multivariate distribution R of the ensemble forecast. Schefzik et al. (2013) describe this copula as

$$E_M\left(\frac{i_1}{M}, \ldots, \frac{i_{|\mathcal{S}|}}{M}\right) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}\left(\text{rank}\left(x_m^1\right) \leq i_1, \ldots, \text{rank}\left(x_m^{|\mathcal{S}|}\right) \leq i_{|\mathcal{S}|}\right)$$

with integers $0 \leq i_1, \ldots, i_{|\mathcal{S}|} \leq M$ and rank (x_m^s) denoting the rank of x_m^s within the set $x_1^s, \ldots, x_M^s$.

Now, let $\hat{F}$ be the multivariate empirical CDF of the ECC approach with marginal empirical CDFs $\hat{F}_1, ..., \hat{F}_{|\mathcal{S}|}$ obtained through univariate postprocessing and subsequent sampling as described in step 1. above. Then the marginal distributions are again linked through the same empirical copula E_M to the multivariate forecast distribution:

$$\hat{F}\left(y_1, \ldots, y_{|\mathcal{S}|}\right) = E_M\left(\hat{F}_1\left(y_1\right), \ldots, \hat{F}_{|\mathcal{S}|}\left(y_{|\mathcal{S}|}\right)\right) \tag{3.2}$$

for $y_1, \ldots, y_{|\mathcal{S}|} \in \mathbb{R}$. Hence, ECC can be interpreted as a copula approach. The $|\mathcal{S}|$-dimensional predictive distribution $\hat{F}$ is constructed through the marginal distributions $\hat{F}_1, \ldots, \hat{F}_{|\mathcal{S}|}$ and the empirical copula E_M which reflects the spatial dependency of the raw ensemble. The relationship between ECC, empirical copulas and Sklar's theorem is presented more detailed in Schefzik et al. (2013) and Schefzik (2015).

Since ECC models the dependence pattern based on the direct NWP output, it is of crucial importance for the raw ensemble to deliver physically consistent forecasts. Deficiencies in the NWP model transfer to ECC and might even be amplified. Further extensions to ECC have been published. The lack of flow dependence in the spatio-temporal dependence has been addressed by incorporating the autocorrelation of forecast errors over consecutive lead times (Ben Bouallègue et al., 2016), choosing training data based on similarity criteria (Bellier et al., 2017) or smoothing the temporal trajectories (Bellier et al., 2018). Hu et al. (2016) suggest stratified sampling to construct the sample and improve the quality of the forecasts in the first step of the ECC ensemble assembly.

3.2.2 Schaake shuffle

Similar to ECC, the Schaake shuffle can be considered a copula approach and the technique follows the same steps as ECC. The two methods only differ in the origin of the dependence template which is based on past observations for the Schaake shuffle instead of the ensemble forecasts. This implies that an ensemble of any size can be obtained provided that sufficiently many past observations are available. For coherence with the other methods, we chose the same size as the numerical EPS for the case studies. Then the multivariate predictive distribution of the Schaake shuffle equals Eq. 3.2 replacing E_M with the empirical copula induced by past observations.

In the original approach, the past observations dates comprise of data from all available past years and lie within 7 days of the date of interest. More sophisticated techniques to select the template data have been developed by providing different similarity criteria to match current atmospheric situations to analogues in the observed data – for details see Schefzik (2016), Scheuerer et al. (2017) or Bellier et al. (2017).

3.3 A nonstationary covariance model on spheres

In this section, we introduce the foundation for spatial EMOS, which aims to capture spatial dependencies by modeling the correlation of the EMOS forecast error fields across the globe. These error fields are interpreted as realizations of a Gaussian random field. As the normal distribution is completely characterized by its first and second moment, we can describe the random field by modeling its covariance matrix if we assume the mean to be constant.

Due to more and more availability of spatial data over the past two decades, studying covariance functions is once again a growing research area in spatial statistics – particularly the construction of valid and flexible covariance functions for different spatial scenarios. In many applications nonstationary random fields are necessary, since the dependence structure changes across the domain. As Stein (2007) point out the covariance structure of environmental data often exhibits near stationarity in longitude, while being nonstationary in terms of latitude. The global temperature forecasts we consider in the case study in Chapter 4 exhibit this dependence pattern as well. Hence, we are interested in nonstationary covariance functions describing this characteristic while also being valid on spheres.

For the mathematical background, we introduce geostatistical models, that assign a value to points located across a spatial domain. These models, referred to as *random fields*, are special cases of stochastic processes.

Definition 3.3.1. Let $\mathcal{S} \subseteq \mathbb{R}^p$ for $p \in \mathbb{N}$ be a spatial domain of interest. A *random field (RF)* is a collection of random variables $(Z(\mathbf{s}))_{\mathbf{s}\in\mathcal{S}}$ on a joint probability space $(\Omega, \mathcal{F}, P)$.

The Kolmogorov existence theorem[1] states that under mild conditions stochastic processes are uniquely determined by their finite-dimensional distributions. For every $n \in \mathbb{N}$ and every set of sites $\mathbf{s}_1, ..., \mathbf{s}_n \in \mathcal{S}$, the respective finite-dimensional accumulative distribution function of the process is

$$F(z_1, \ldots, z_n; \boldsymbol{s}_1, \ldots, \boldsymbol{s}_n) = \mathbb{P}\left(Z(\boldsymbol{s}_1) \leq z_1, \ldots, Z(\boldsymbol{s}_n) \leq z_n\right), \tag{3.3}$$

where $z_1, ..., z_n \in \mathbb{R}$. Gaussian processes are very popular in statistical modeling as they approximate many real world phenomena and are easily applicable due to their mathematical properties. For our case studies in particular, we consider error fields which can be assumed to be Gaussian.

[1] For details see e.g. Grimmett and Stirzaker (2020) or Billingsley (2012).

Definition 3.3.2. The RF $(Z(\mathbf{s}))_{\mathbf{s}\in\mathcal{S}}$ is a *Gaussian random field (GRF),* if for all $n \in \mathbb{N}$ and $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathcal{S}$ the distribution of the random vector $(Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^\intercal$ is multivariate normal.

The simplicity of the normal distribution stems from the fact that it can be completely and easily described by its first and second moment. Specifically, given a GRF $(Z(\mathbf{s}))_{\mathbf{s}\in\mathcal{S}}$, its finite-dimensional densities corresponding to Eq. (3.3) equal

$$\frac{1}{(2\pi)^{n/2}\,\det(\Sigma)^{1/2}}\,\exp\left(-\frac{1}{2}(\mathbf{z}-\mu)^\intercal\Sigma^{-1}(\mathbf{z}-\mu)\right),$$

where $\mathbf{z} \in \mathbb{R}^n$, with mean vector $\mu = \mathbb{E}\left(Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\right)^\top$ and covariance matrix $\Sigma = [\mathrm{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j))]_{1\le i,j\le n}$. If we assume $\mathbb{E}[Z(\mathbf{s})]$ to be constant for all sites $\mathbf{s} \in \mathcal{S}$, the whole process can be characterized completely via its *covariance function.*

Definition 3.3.3. Let $(Z(\mathbf{s}))_{\mathbf{s}\in\mathcal{S}}$ be an RF. If its second moments exist, the function

$$C : \mathcal{S}\times\mathcal{S} \to \mathbb{R}, \qquad C(\mathbf{s}_1, \mathbf{s}_2) = \mathrm{Cov}\left(Z(\mathbf{s}_1), Z(\mathbf{s}_2)\right)$$

is the *covariance function* of this field.

There are two particularly interesting classes of RFs considered in the literature. The first one consists of fields whose covariance structure only depends on the lag vector between two points, while the other one consists of fields with covariance functions only depending on the Euclidean distance between two points.

Definition 3.3.4. Let $(Z(\mathbf{s}))_{\mathbf{s}\in\mathcal{S}}$ be an RF for which the second moments exist.

1. The field is called *(second order) stationary*, if $\mathbb{E}\left(Z(\mathbf{s})\right)$ and $\mathrm{Cov}\left(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})\right)$ do not depend on $\mathbf{s} \in \mathcal{S}$. Consequently, the covariance fulfills

   $$\mathrm{Cov}\left(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})\right) = \mathrm{Cov}\left(Z(\mathbf{0}), Z(\mathbf{h})\right) = \tilde{C}(\mathbf{h})$$

   for some function $\tilde{C} : \mathcal{S} \to \mathbb{R}$ and any lag vector $\mathbf{h} \in \mathcal{S}$.

2. The field is called *isotropic*, if it is stationary and the covariance function between two points $\mathbf{s}_1, \mathbf{s}_2$ only depends on the Euclidean distance between these two points, induced by the considered metric. In other words, the covariance reduces to

$$\mathrm{Cov}\left(Z(\mathbf{s}_1), Z(\mathbf{s}_2)\right) = ||\mathbf{s}_1 - \mathbf{s}_2||,$$

where $||\mathbf{s}_1 - \mathbf{s}_2|| = \left((\mathbf{s}_1 - \mathbf{s}_2)^\intercal (\mathbf{s}_1 - \mathbf{s}_2)\right)^{1/2}$ denotes the Euclidean distance between $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{S}$.

Hence, the covariance function can be completely characterized by a function on the domains $\mathcal{S} \times \mathcal{S}$, $\mathcal{S}$ or $\mathbb{R}$. We will call this function covariance function and denote it by C if it is clear on which domain C operates.

Example 3.3.1. An example of a parametric covariance function for an isotropic GRF is the *Matérn* function (Matérn, 1986) of the form

$$C(\delta) = \sigma^2 \mathcal{M}_\nu \left(\sqrt{2\nu}\,\frac{\delta}{\rho}\right) = \sigma^2 \sqrt{2\nu}\,\frac{\delta}{\rho} \mathcal{K}_\nu \left(\sqrt{2\nu}\,\frac{\delta}{\rho}\right), \tag{3.4}$$

where δ is the Euclidean distance, $\mathcal{K}_\nu$ refers to the modified Bessel function of the second kind of order ν, Γ represents the gamma function and σ^2 denotes the variance. The range or scaling parameter ρ and the smoothness parameter ν of the function are both positive. In case of a Matérn model, the covariance between two points solely depends on their Euclidean distance δ. For $\nu = \frac{1}{2}$, the model reduces to the *exponential* covariance function

$$C(\delta) = \sigma^2 \exp\left(-\frac{\delta}{\rho}\right). \tag{3.5}$$

We rely on the Matérn and exponential covariance models in the case studies.

In analogy to covariances, one can define the *correlation* function Cor of a GRF $(Z(\mathbf{s}))_{\mathbf{s}\in\mathcal{S}}$ as

$$\mathrm{Cor}(\mathbf{s}_1, \mathbf{s}_2) = \begin{cases} \frac{C(\mathbf{s}_1, \mathbf{s}_2)}{\sqrt{\mathrm{Var}(Z(\mathbf{s}_1))\mathrm{Var}(Z(\mathbf{s}_2))}} & \text{if } \mathrm{Var}\left(Z(\mathbf{s}_1)\right) \neq 0 \text{ and } \mathrm{Var}\left(Z(\mathbf{s}_2)\right) \neq 0 \\ 0 & \text{otherwise} \end{cases},$$

which is a normalized covariance function.

Determining whether a function is a valid covariance function is equivalent to the question whether it is a positive definite kernel.

Definition 3.3.5. A function $h : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ on $\mathcal{S} \subseteq \mathbb{R}^p$, $p \in \mathbb{N}$ is a *positive definite kernel*, if

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j h(\mathbf{s}_i, \mathbf{s}_j) \geq 0$$

for all $n \in \mathbb{N}$, any $a_1, ..., a_n \in \mathbb{R}$ and points $\mathbf{s}_1, ..., \mathbf{s}_n \in \mathcal{S}$.

The following well-known theorem (see e.g. Yaglom, 1987) describes the relationship between covariance functions and positive definite kernels.

Theorem 3.3.1. For a function $C : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ on $\mathcal{S} \subseteq \mathbb{R}^p$, $p \in \mathbb{N}$ the following statements are equivalent.

(i) The function C is a positive definite kernel.

(ii) On $\mathcal{S}$, there exists a stationary (Gaussian) RF $(Z(\mathbf{s}))_{\mathbf{s}\in\mathcal{S}}$ with C as covariance function.

This theorem ensures the existence of a GRF if we can provide a valid covariance function C. The requirement of positive definiteness follows from the interpretation of a covariance between two points as a scalar product.

Closely related to covariance functions are variograms, which are used widely in the geostatistical community to visualize and model the spatial dependence structure of stationary fields.

Definition 3.3.6. Let $(Z(\mathbf{s}))_{\mathbf{s}\in\mathcal{S}}$ be an RF. If its second moments exist, we define

$$\gamma : \mathcal{S} \times \mathcal{S} \to \mathbb{R}, \qquad \gamma(\mathbf{s}_1, \mathbf{s}_2) = \frac{1}{2}\,\mathbb{E}\left[\left(Z(\mathbf{s}_1) - Z(\mathbf{s}_2)\right)^2\right]$$

to be its *variogram*.

Gneiting et al. (2001) explore in depth analogies and correspondences between variograms and covariances. Assuming the mean vector of the RF to be constant, the relationship between the two is established via

$$\gamma(\mathbf{s}_1, \mathbf{s}_2) = \frac{1}{2}C\left(\mathbf{s}_2, \mathbf{s}_2\right) - C\left(\mathbf{s}_1, \mathbf{s}_2\right) + \frac{1}{2}C\left(\mathbf{s}_1, \mathbf{s}_1\right).$$

For a stationary RF, this expression reduces to $\gamma(\mathbf{h}) = C\left(0\right) - C\left(\mathbf{h}\right)$ for any lag vector $\mathbf{h}$. Scheuerer and Hamill (2015b) employ variograms to evaluate multi-dimensional forecasts – see Section 3.5.

Since the temperature forecasts are provided globally, we are particularly interested in covariance functions for which the domain $\mathcal{S}$ is the surface of a sphere, namely $\mathbb{S}_r^k = \left\{\mathbf{x} \in \mathbb{R}^{k+1} : ||x|| = r\right\}$ for $k \in \mathbb{N}$ and radius $r \in \mathbb{R}^+$. Assuming Earth to be a true sphere, we focus on the two-dimensional spherical surface $\mathbb{S}_r^2$ with radius $r =$ 6371km, embedded in the three-dimensional Euclidean space $\mathbb{R}^3$. A point on $\mathbb{S}_r^2$ can be characterized as (r, lat, lon) with latitude $lat \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ and longitude $lon \in [-\pi, \pi)$. In the Euclidean space, this point is described by (a, b, c), where $a = r\cos(lat)\cos(lon)$, $b = r\sin(lat)\cos(lon)$ and $c = r\sin(lat)$.

For many real-world environmental processes, the covariance decays with distance between points and thus the covariance function should depend on this metric. A useful quantity to calculate distances on a sphere is the central angle between two points (lat_1, lon_1) and (lat_2, lon_2), defined as:

$$\alpha = \arccos\left\{\sin(lat_1)\sin(lat_2) + \cos(lat_1)\cos(lat_2)\cos(lon_1 - lon_2)\right\}. \tag{3.6}$$

Then the canonical distance on $\mathbb{S}_r^2$ – the *great circle* or *geodesic distance* – is defined as $r\alpha$. Also the Euclidean distance in $\mathbb{R}^3$ can be expressed in terms of α by $2r \cdot \sin\left(\frac{\alpha}{2}\right)$. For small ranges, these two metrics almost coincide, but distortion grows with larger distances.

Given a GRF on $\mathbb{S}_r^2$, the property of isotropy directly transfers to the sphere. However, stationarity cannot be defined the same way as there exists no canonical choice of lag vectors $\mathbf{h}$ in Definition 3.3.4 on the sphere. Instead of stationarity, Jones (1963) introduces the term of an axially symmetric process. Here, we follow the definition by Castruccio and Stein (2013):

Definition 3.3.7. Let $(Z(\mathbf{s}))_{\mathbf{s}\in\mathcal{S}}$ be a GRF on the sphere $\mathcal{S} \subset \mathbb{S}_r^2$ with $r \in \mathbb{R}^+$. All points $\mathbf{s} \in \mathcal{S}$ can be described in terms of latitude $lat \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ and longitude $lon \in [-\pi, \pi)$.We call the process *axially symmetric*, if

$$\mathrm{Cov}\left(Z(lat_1, lon_1), Z(lat_2, lon_2)\right) = \check{C}(lat_1, lat_2, lon_1 - lon_2)$$

for some function $\check{C} : \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \times [-2\pi, 2\pi) \to \mathbb{R}$.

Thus the covariance function of the GRF depends on the site-specific latitudes and longitudinal lag of the considered points. Visually, this means that the GRF model is invariant to rotations about the main axis of the Earth. Through this definition, all isotropic processes on the sphere are axially symmetric, but not all axially symmetric processes are isotropic.

Intuitively, the dependence – and hence, covariance structure – of many real world phenomena should somewhat be influenced by the distance – with smaller dependencies for smaller distances and vice versa. As we have seen, the simplest covariance models on $\mathbb{R}^3$ with respect to the Euclidean metric are isotropic ones. Unfortunately, many of these models are no valid covariance functions on the sphere equipped with the geodesic distance – i.e. if we replace the Euclidean with the geodesic metric, the covariance function changes and loses the property of positive definiteness. For instance, Gneiting (2013) points out that the frequently applied Matérn covariance function is only positive definite on a sphere with great circle distance for smoothness parameter $\nu \in \left(0, \frac{1}{2}\right]$. Banerjee (2005) states that careless formulation of distances will lead to false estimation of the range parameter and poor prediction.

Construction of valid and useful covariance functions – like axially symmetric ones – on spheres with geodesic metrics is thus challenging and different approaches have been developed. However, in some applications on the globe it may be advantageous to use the Euclidean rather than the geodesic distance. Jeong and Jun (2015a) and Guinness and Fuentes (2016) compare Matérn covariance models to Matérn-like covariance functions, which are valid on spheres paired with great circle distances. Both conclude that in terms of model fit and prediction, models with Euclidean distance yield better verification results. Only for large correlation lengths, the performance might improve by using a Matérn-like model with great circle distances (Jeong and Jun, 2015b). Since the correlation length of the temperature error fields considered in the case studies does not exceed $1,000$km (see the variogram in Figure 4.9), we choose to employ Euclidean distances. This allows us to use any function of the rich class of established covariance models on $\mathbb{R}^3$ without any changes; since a valid covariance function in $\mathbb{R}^3$ is a covariance function on $\mathbb{S}_r^2 \subset \mathbb{R}^3$ relative to Euclidean metrics (Yadrenko, 1983).

As mentioned before, the covariance structure of the temperature data varies significantly in latitude, while the process may be nearly stationary in terms of longitude. Additionally, the correlation length differs depending on local characteristics. To allow for this flexibility, we chose to combine EMOS with a Matérn model. This model is a member of the following class of covariance functions, described by Stein (2005):

Theorem 3.3.2 (Stein). Suppose Υ is a mapping from $\mathbb{R}^p$ to the class of the positive definite $p \times p$ matrices, λ is a non-negative measure on $[0, \infty)$, and for each $\mathbf{s}_1 \in \mathbb{R}^p$, $g(\cdot, \mathbf{s}_1) \in L^2(\lambda)$, where $L^2(\lambda)$ denotes the space of quadratically integrable functions relative to the measure λ. With $\Upsilon(\mathbf{s}_1, \mathbf{s}_2) = \frac{1}{2}\Upsilon(\mathbf{s}_1) + \frac{1}{2}\Upsilon(\mathbf{s}_2)$ and $Q(\mathbf{s}_1, \mathbf{s}_2) = (\mathbf{s}_1 - \mathbf{s}_2)^\intercal \Upsilon(\mathbf{s}_1, \mathbf{s}_2)^{-1} (\mathbf{s}_1 - \mathbf{s}_2)$,

$$R(\mathbf{s}_1, \mathbf{s}_2) = \frac{|\Upsilon(\mathbf{s}_1)|^{1/4} |\Upsilon(\mathbf{s}_2)|^{1/4}}{|\Upsilon(\mathbf{s}_1, \mathbf{s}_2)|^{1/2}} \int_0^\infty e^{-\omega Q(\mathbf{s}_1, \mathbf{s}_2)} g(\omega, \mathbf{s}_1)\, g(\omega, \mathbf{s}_2)\, \lambda(d\omega),$$

is a covariance function on $\mathbb{R}^p \times \mathbb{R}^p$.

This covariance function is an extension to a nonstationary model proposed by Paciorek (2003). Porcu et al. (2009) prove a more general result and Kleiber and Nychka (2012) extend this approach for multivariate spatial processes. As a special case of Theorem 3.3.2., Stein (2005) finds that

$$R(\mathbf{s}_1, \mathbf{s}_2) = \frac{\sigma(\mathbf{s}_1)\sigma(\mathbf{s}_2)}{|\Upsilon(\mathbf{s}_1, \mathbf{s}_2)|^{1/2}} \mathcal{M}_{\{\nu(\mathbf{s}_1)+\nu(\mathbf{s}_2)\}/2}\left(Q(\mathbf{s}_1, \mathbf{s}_2)^{1/2}\right) \tag{3.7}$$

is a covariance function on $\mathbb{R}^p \times \mathbb{R}^p$ with $\mathcal{M}_\nu$ defined as in Eq. (3.4). In particular, this Matérn-based model allows for locally varying variance $\sigma(\mathbf{s})$, smoothness $\nu(\mathbf{s})$ and distance measures through Υ. If Υ is the unit matrix, then Q describes the squared Euclidean distances between points $\mathbf{s}_1, \mathbf{s}_2$ allowing thus for spatial modulation through matrix Υ.

This very specific model is highly tailored to the requirements of our data set. There exist other options to construct a heterogeneous covariance function – in the sense that it is neither isotropic nor stationary in $\mathbb{R}^3$ or axially symmetric in $\mathbb{S}^2$. Schmidt and Guttorp (2020) discuss current approaches of constructing such functions relative to Euclidean distances. On spheres paired with geodesic distances, defining nonstationary valid processes is more challenging. Das (2000) translates the deformation approach by Sampson and Guttorp (1992) from plane to sphere to describe a new class of parametric anisotropic covariance functions, which at the time was computationally heavy. Instead, Jun and Stein (2008), Jun (2011) and Bolin et al. (2011) use stochastic partial differential equations to construct nonstationary covariance functions in a computationally more tractable way. Heaton et al. (2014) employ the kernel convolution approach based on works by Higdon (1998) for the construction of spatial processes on spheres. Another interesting model is put forward by Castruccio and Guinness (2017): Based on an evolutionary spectrum approach, their covariance structure allows for incorporation of heterogeneous geography. For a more in depth analysis, we refer the reader to Jeong et al. (2017), who give a detailed overview of valid covariance functions in terms of geodesic distances.

3.4 Spatial EMOS

In the spirit of Gel et al. (2004), Berrocal et al. (2007) and Feldmann et al. (2015), we combine the univariate postprocessing method EMOS with a GRF model (see Sections 2.2 and 3.3, respectively) to account for spatial dependencies in the predictive distribution. This yields a multivariate normal (MVN) predictive distribution that models the spatial structure of the forecast field, in that, given an ensemble forecast with $M \in \mathbb{N}$ members at a finite collection of $n \in \mathbb{N}$ spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathcal{S}$,

$$\mathbf{y}|\mathbf{f}_1, \ldots, \mathbf{f}_M \sim MVN(\mathbf{a}+\mathbf{b} \circ \bar{\mathbf{f}}, \boldsymbol{\Sigma}),$$

where the components of vector $\mathbf{y} = (y_{\mathbf{s}_1}, ..., y_{\mathbf{s}_n})^\top \in \mathbb{R}^n$ represent the temperature at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathcal{S}$, $\mathbf{f}_m = (f_{m,\mathbf{s}_1}, \ldots, f_{m,\mathbf{s}_n})^\top \in \mathbb{R}^n$ ´denotes the respective ensemble member forecasts for $m = 1, \ldots, M$, and $\bar{\mathbf{f}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{f}_m$ is the ensemble mean forecast. The vectors $\mathbf{a} = (a_{\mathbf{s}_1}, \ldots, a_{\mathbf{s}_n})^\intercal$ and $\mathbf{b} = (b_{\mathbf{s}_1}, \ldots, b_{\mathbf{s}_n})^\intercal$ are obtained by fitting univariate EMOS models of the form as Eq. (2.2) at every location $\mathbf{s}_1, \ldots, \mathbf{s}_n$, and $\circ$ denotes a component-wise product of two vectors. Given any fixed set of parameter values, the standardized EMOS forecast error at location $\mathbf{s} \in \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ equals

$$\varepsilon_{\mathbf{s}} = \frac{y_{\mathbf{s}} - (a_{\mathbf{s}} + b_{\mathbf{s}}\bar{f}_{\mathbf{s}})}{\sqrt{c_{\mathbf{s}} + d_{\mathbf{s}} v_{\mathbf{s}}^2}}. \tag{3.8}$$

which can be assumed to each follow a standard normal distribution $\epsilon_{\mathbf{s}} \sim \mathcal{N}(0, 1)$. Minimizing these errors independently might result in a misspecified spatial correlation structure of the EMOS forecast field. To compensate for this, we assume a Gaussian random field model for these errors.

Feldmann et al. (2015) fit a GRF model to the respective standardized error fields with an exponential correlation function $\mathrm{Cor}_{\theta\rho}$ (see Eq. 3.5) with nugget effect θ given by

$$\mathrm{Cor}_{\theta,\rho}(\mathbf{s}_i, \mathbf{s}_j) = (1 - \theta) \exp\left(-\frac{||\mathbf{s}_i - \mathbf{s}_j||}{\rho}\right) + \theta\delta_{ij}. \tag{3.9}$$

where the range parameter is denoted by ρ and δ_{ij} indicates the Kronecker delta. This model is stationary and isotropic. As in Berrocal et al. (2007), Feldmann et al. (2015) estimate the parameters via variograms and weighted least squares (Cressie, 1985) over a sliding window training set with the same length as for the univariate EMOS parameters. The estimated correlation matrix is then converted to a covariance matrix, by multiplying it with a diagonal matrix with entries corresponding to the predictive variances $c + dv_{\mathbf{s}}^2$ from the univariate EMOS models.

For comparison in forecast performance, Feldmann et al. (2015) fit a Matérn correlation function $\mathrm{Cor}_{\theta,\nu,\rho}$ (Guttorp, 2006) to the Gaussian random field:

$$\mathrm{Cor}_{\theta,\nu,\rho}(\mathbf{s}_i, \mathbf{s}_j) = (1 - \theta)\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{||\mathbf{s}_i - \mathbf{s}_j||}{\rho}\right)^{\nu} \mathcal{K}_\nu\left(\sqrt{2\nu}\frac{||\mathbf{s}_i - \mathbf{s}_j||}{\rho}\right) + \theta\delta_{ij}, \tag{3.10}$$

where Γ denotes the gamma function, $||\cdot||$ represents the Euclidean distance and $\mathcal{K}_\nu$ is the modified Bessel function of the second kind with smoothness parameter $\nu > 0$ and δ_{ij} indicates the Kronecker delta. This is the normalized case of the covariance function in Eq. (3.4) with nugget effect θ. However the performance results did not improve over the application of an exponential model.

As the area of interest was limited to Germany in Feldmann et al. (2015), a rather simple stationary and isotropic correlation model is suitable for this specific data set. The forecasts we consider in the following chapters cover the entire globe, however. Since the covariance structure varies significantly over this grand domain, we propose to use a nonstationary correlation function instead rooted in Eq. (3.7). This model allows for more flexibility in the Matérn-based parameters.

For 2m temperature, the type of surface in terms of land or sea has a huge impact on the statistical behavior of the data. Multiple publications have shown that incorporating such geographical covariates can improve the covariance models – e.g. see Jun (2014) or Castruccio and Guinness (2017). So, we chose to also account for these two regimes by multiple range and smoothness parameters in our model.

However when we fit a simple Matérn correlation function to different sets of forecasts residuals over land and sea separately, the estimated smoothness parameters almost coincide (estimate around 1 for both regimes), but the range parameters differ. Hence, we only account for the type of surface in two different range parameters, resulting in a covariance of the form

$$\mathrm{Cov}_{\nu,\rho_1,\rho_2,\sigma}(\mathbf{s}_1,\mathbf{s}_2) = \frac{\sigma(\mathbf{s}_1)\sigma(s_2)}{|\Upsilon\left(s_1,s_2\right)|^{\frac{1}{2}}}\mathcal{M}_\nu\left(\left[(\mathbf{s}_1-\mathbf{s}_2)^\top\Upsilon(\mathbf{s}_1,\mathbf{s}_2)^{-1}(\mathbf{s}_1-\mathbf{s}_2)\right]^{\frac{1}{2}}\right).$$

Here, the term $\sigma(s)$ corresponds to the local variance, $\mathcal{M}_\nu(x) = x^\nu K_\nu(x)$ and we define a distance function

$$\Upsilon(\mathbf{s}_1,\mathbf{s}_2) = \frac{1}{2}\left(\Upsilon(\mathbf{s}_1)+\Upsilon(\mathbf{s}_2)\right); \quad \Upsilon(\mathbf{s}) = \begin{cases} \rho_1\mathbf{E}_3 & \text{if } \mathbf{s}\in \text{land} \\ \rho_2\mathbf{E}_3 & \text{if } \mathbf{s}\in \text{sea} \end{cases}$$

with the unit matrix $\mathbf{E}_3$ of dimension 3, smoothness parameter ν, different range parameters ρ_1 and ρ_2 over land and sea, respectively. For the local variance, our model relies on the EMOS predictive variance, hence we are interested in the corresponding correlation structure:

$$\mathrm{Cor}_{\nu,\rho_1,\rho_2}(\mathbf{s}_1,\mathbf{s}_2) = \frac{2^{1-\nu}}{\Gamma(\nu)}\frac{|\Upsilon(\mathbf{s}_1)\Upsilon(\mathbf{s}_2)|^{\frac{1}{4}}}{|\Upsilon(\mathbf{s}_1,\mathbf{s}_2)|^{\frac{1}{2}}}\mathcal{M}_\nu\left(\left[(\mathbf{s}_1-\mathbf{s}_2)^\top\Upsilon(\mathbf{s}_1,\mathbf{s}_2)^{-1}(\mathbf{s}_1-\mathbf{s}_2)\right]^{\frac{1}{2}}\right). \tag{3.11}$$

which is based on the model in Eq. (3.7). When estimating the parameters ν, ρ_1 and ρ_2, we use maximum likelihood. Assuming the forecast errors are independent over different days, the log-likelihood stipulates

$$l(\nu, \rho_1, \rho_2) = \sum_{t \in T} \left\{ -\frac{|\mathcal{S}|}{2} \log(2\pi) - \frac{1}{2} \log\det \Sigma - \frac{1}{2} \mathbf{e}_t^\top \Sigma^{-1} \mathbf{e}_t \right\},$$

where $|\mathcal{S}|$ denotes the number of sites and $\mathbf{e}_t$ the standardized EMOS error field (see Eq. (3.8)) on the training day $t \in T$ and $\Sigma = [\mathrm{Cor}_{\nu,\rho_1,\rho_2}(\mathbf{s}_i, \mathbf{s}_j)]_{1 \leq i,j \leq n}$ the correlation matrix.

For a non-parametric variant of spatial EMOS (spatial EMOS emp cor), we use the empirical correlation function of the standardized error field in lieu of fitting a Matérn correlation model. Then we proceed as described before.

3.5 Multivariate verification

Similarly to Section 2.4, we now present diagnostic tools applicable to multivariate forecasts. First we review different techniques to assess multivariate calibration, followed by a description of multivariate scoring rules.

3.5.1 Multivariate calibration

Gneiting et al. (2008), Ziegel and Gneiting (2014) and Wilks (2017) discuss different tools to evaluate multivariate calibration. In the spirit of Thorarinsdottir and Schuhen (2018), we focus on four histogram techniques, namely the multivariate, the minimum spanning tree, the average and the band depth rank histogram, which are all based on the following two-steps approach. Let $\mathcal{G} = \{\mathbf{f}_0, \mathbf{f}_1, \ldots, \mathbf{f}_M \in \mathbb{R}^n\}$ be the set of an M-member forecast ensemble and the corresponding observation vector, denoted here by $\mathbf{f}_0$. To calculate the rank of the observation within the ensemble, we proceed as follows:

1. *Pre-rank*

   To each ensemble member and the observation, we apply a pre-rank function $\psi : \mathbb{R}^n \to \mathbb{R}$ to find the related pre-ranks $\psi(\mathbf{f}_m)$ for $m = 0, \ldots, M$.

2. *Multivariate rank*

   Then the rank of the observation vector $\mathbf{f}_0$ equals the rank of the corresponding pre-rank $\psi(\mathbf{f}_0)$ within the set of pre-ranks $\{\psi(\mathbf{f}_0), \ldots, \psi(\mathbf{f}_M)\}$, in particular

   $$\mathrm{rank}(\mathbf{f}_0) = \sum_{m=0}^{M} \mathbb{I}\left\{\psi(\mathbf{f}_m) \leq \psi(\mathbf{f}_0)\right\}$$

   with indicator function $\mathbb{I}$ and ties resolved at random.

After accumulating a large number of forecast cases, multivariate calibration can be assessed via the rank histogram similarly to the univariate setting.

The four methods only differ in the definition of the pre-rank function in the first step. For the *multivariate rank* histogram (Gneiting et al., 2008), the pre-rank function equals

$$\psi^{\text{mul}}(\mathbf{f}_i) = \sum_{m=0}^{M} \mathbb{I}\left\{\mathbf{f}_m \preceq \mathbf{f}_i\right\}$$

where $\mathbf{f}_m \preceq \mathbf{f}_i$ holds if all components fulfill $f_m^{(j)} \leq f_i^{(j)}$, $j = 1, \ldots, n$. Ties are resolved at random and the pre-ranks take integers between 1 and $M+1$. Especially for dimensions greater than 3, the multivariate rank histogram can falsify signal calibration, because seldom all components of two vectors follow the same order structure (Pinson and Girard, 2012). The *average rank* histogram (Thorarinsdottir et al., 2016) counteracts this flaw by calculating the univariate ranks of each component instead

$$\text{rank}(\mathbf{f}_i, j) = \sum_{m=0}^{M} \mathbb{I}\left\{f_m^{(j)} \leq f_i^{(j)})\right\}$$

and then averaging them in the pre-rank function

$$\psi^{\text{av}}(\mathbf{f}_i) = \frac{1}{n} \sum_{j=1}^{n} \text{rank}(\mathbf{f}_i, j).$$

Ties are again resolved at random and pre-ranks take real values between 1 and $M+1$. Furthermore, Thorarinsdottir et al. (2016) propose the *band depth* ranking, which again incorporates the component-wise ranks

$$\psi^{\text{bd}}(\mathbf{f}_i) = \frac{1}{n} \sum_{j=1}^{n} \Big[\text{rank}(\mathbf{f}_i, j) \left[(M+1) - \text{rank}(\mathbf{f}_i, j)\right] + \left[\text{rank}(\mathbf{f}_i, j) - 1\right] \sum_{m=0}^{M} \mathbb{I}\left\{f_m^{(j)} = f_i^{(j)}\right\} \Big].$$

If all components of all elements in the set $\mathcal{G}$ differ, the equation simplifies to

$$\psi^{\text{bd}}(\mathbf{f}_i) = \frac{1}{n} \sum_{j=1}^{n} \left[\left[(M+1) - \text{rank}(\mathbf{f}_i, j)\right] \left[\text{rank}(\mathbf{f}_i, j) - 1\right]\right].$$

Ties within the pre-ranks are again resolved at random. This ranking assesses the centrality of the observation within the ensemble, meaning $\psi^{\text{bd}}(\mathbf{f}_0)$ gets assigned small values when $\mathbf{f}_0$ falls to the extreme end of the- ensemble and large values when $\mathbf{f}_0$ is closed to the middle of the ensemble.

Smith (2001), Smith and Hansen (2004) and Wilks (2004) introduce the minimum spanning tree ranking. A minimum spanning tree is a connected graph without any loops, where each edge between two points gets assigned a weight. Then the minimum spanning tree is the set of edges for which the sum of these weights, also called length, is minimized. More details on minimum spanning tree can be found in Kruskal (1956). For the minimum spanning tree ranking, the pre-rank function $\psi^{\text{mst}}(\mathbf{f}_i)$ calculates the length of the minimum spanning tree of the set $\mathcal{G}\backslash\mathbf{f}_i$, meaning the set $\mathcal{G}$ without element $\mathbf{f}_i$:

$$\psi^{\text{mst}}(\mathbf{f}_i) = ||\text{MST}\left[\{\mathbf{f}_0, \ldots, \mathbf{f}_M\} \setminus \{\mathbf{f}_i\}\right]||$$

with Euclidean distances $||\cdot||$. Smith and Hansen (2004) state that other distance metrics might be more suitable in certain settings. Like the band depth ranking, the minimum spanning tree rank histogram evaluates the centrality of the observation within the forecast ensemble.

All of these ranking methods can easily be used to plot the rank histogram of the verifying observations. While calibration always results in a uniform shape, the interpretation of the histogram varies for misspecified forecasts. In a simulation study, Wilks (2017) found that all of the presented methods lack the capacity to identify miscalibration in certain settings and should always be used in conjunction.

Gneiting et al. (2008) state that the interpretation of multivariate rank histogram is the same as for the univariate rank histogram (Section 2.4.1); for $n = 1$ both collapse. The implications of the average ranking also coincides with the univariate rank histogram in terms of overdispersion, underdispersion and biases. Additionally, underestimation of correlation can result in a $\cap$-shaped histogram and overestimation in a $\cup$-shaped histogram. For the minimum spanning tree ranking, too weak correlation, underdispersion or biases in the ensemble result in a triangular-shaped histogram with many low ranks. In contrast, overestimation of correlation or overdispersion in the forecasts display in a triangle with many higher ranks. In case of the band depth ranking, a skewed histogram with many low ranks indicates underdispersion or a biased ensemble, while many high ranks imply overdispersion. Too low or too high correlation within the forecasts results in a $\cup$-shaped or $\cap$-shaped histogram, respectively.

As we have seen, interpretation of the resulting histograms is not straightforward and further discussion on this topic can be found in Thorarinsdottir et al. (2016) and Wilks (2017). Assessing multivariate calibration is an ongoing research question, as for instance Jacobson et al. (2020) proposed a new diagnostic tool applicable to forecast fields, which aims to detect correlation length and thus facilitates interpretability of the histogram.

3.5.2 Scoring rules

Applying appropriate multivariate scores is a straightforward method to assess the overall skill of multivariate forecasts. Another option is to reduce the dimension of the forecasts by considering aggregated univariate quantities instead to enable the application of univariate scores (see Section 2.4.2). Dependence structures are of critical importance when constructing aggregated univariate quantities, such as minima, maxima, totals, or averages. The selection of a suitable univariate quantity depends on the data, context and user's needs. For example, see Berrocal et al. (2007) who evaluate temperature forecast fields using this approach.

There exists a variate of applicable scoring rules to assess the multivariate forecasts directly. Some univariate scores have a multivariate equivalent, as the CRPS can be generalized to the *energy score* (ES), which Gneiting et al. (2007) define as

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

where $||\cdot||$ denotes the Euclidean norm, F is the predictive distribution, $\mathbf{y} \in \mathbb{R}^n$ represents the observation and $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{X}' \in \mathbb{R}^n$ are independent random vectors distributed according to F. For $n = 1$, the ES coincides with the CRPS. In case of an ensemble forecast, the ES can be calculated like the CRPS via

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

for ensemble members $\mathbf{f}_1, \ldots, \mathbf{f}_M \in \mathbb{R}^n$.

As an alternative to the ES, the *Dawid-Sebastiani score* (DSS, Dawid and Sebastiani, 1999) can be applied to multivariate forecasts

$$\mathrm{DSS}(F, \mathbf{y}) = \log\det\Sigma_F + (\mathbf{y} - \mu)^\intercal \Sigma_F^{-1}(\mathbf{y} - \mu),$$

with μ being the mean predictive vector and the covariance matrix Σ_F of the predictive distribution. Since not all multivariate forecasts have an explicit predictive covariance matrix, in certain scenarios it has to be estimated. To then avoid sampling errors, the ensemble size needs to be larger than the dimension of the multivariate forecasts - see e.g. Feldmann et al. (2015). If the predictive distribution is Gaussian, the DSS coincides with the multivariate ignorance score (IGN, Roulston and Smith, 2002)

$$\mathrm{IGN}(F, \mathbf{y}) = \log(f(\mathbf{y})),$$

with predictive density f. This score is limited in its applications, as a predictive density f is seldom issued by an EPS. To resolve this constriction, Lerch et al. (2017) propose to estimate such density.

Motivated by the variogram, a popular tool in geostatistics, Scheuerer and Hamill (2015b) introduce the *variogram score* (VS) of the form

$$\mathrm{VS}(F, \mathbf{y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \left(|y^{(i)} - y^{(j)}|^p - \frac{1}{M} \sum_{m=1}^{M} |f_m^{(i)} - f_m^{(j)}|^p \right)^2,$$

where $w_{i,j}$, $i, j = 1, \ldots, n$ are non-negative weights and $p > 0$ is the order of the score. By comparing the pairwise differences in the components of the verification and forecast vectors, the VS captures differences in the correlation structure. As suggested by Scheuerer and Hamill (2015b), we use constant weights $w_{i,j} = 1$ for $i, j = 1, \ldots, n$ and order p equal to 0.5 as well as 1 and 2 which they also considered. The VS and ES can be implemented using the R package `scoringRules` (Jordan et al., 2019).

The multivariate generalization of the absolute error is the *Euclidean error* (EE), defined as

$$\mathrm{EE}(F, \mathbf{y}) = ||\mathrm{smed}_F - \mathbf{y}||,$$

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

$$\mathrm{smed}_F = \arg\min_{\mathbf{X} \in \mathbb{R}^n} \mathbb{E}_F ||\mathbf{X} - \mathbf{X}'||,$$

with $\mathbf{X}'$ a random vector with distribution F.

Apart from the EE, all of these presented scores fulfill the requirement of being proper (see Section 2.4.2); particularly the ES and IGN are even strictly proper. Evaluation of multi-dimensional probabilistic forecasts is an ongoing research topic. In simulation studies and a real-world application to wind speed forecasts, Scheuerer and Hamill (2015b) compare the discrimination ability of the VS, DSS and ES. They recommend using different scores to evaluate multivariate forecasts, as each score shows benefits and limitations in certain settings. Ziel and Berk (2019) find in multiple simulation studies that only the ES can select the true model from all alternatives and should thus be the preferred tool for evaluation. To properly assess the discrimination ability of multivariate scores, they strongly recommend pairing it with the Diebold-Mariano test (Diebold and Mariano, 1995) for calculation of the significance level. Lerch et al. (2020) find that the ranking of different multivariate postprocessing methods is highly sensitive to the applied score and demand further research on this topic.

Statistical tests can be applied to evaluate the statistical significance of score differences for competing forecasting methods. For probabilistic predictions, Ziel and Berk (2019) recommend the Diebold-Mariano (DM) test, which is commonly used in economic studies. In the context of statistical postprocessing for weather forecasts, for example Baran and Lerch (2016) and Lerch et al. (2020) apply this test.

Given two competing prediction models delivering distributions F_i and G_i, their corresponding scores averaged over the test set $i = 1, ..., k$ are denoted by $\bar{s}(F, \mathbf{y}) = \frac{1}{k}\sum_{i=1}^{k} s(F_i, \mathbf{y}_i)$ and $\bar{s}(G, \mathbf{y}) = \frac{1}{k}\sum_{i=1}^{k} s(G_i, \mathbf{y}_i)$, respectively. Then the DM test statistic equals

$$T_k = \sqrt{k}\frac{\bar{s}\left(F, \mathbf{y}\right) - \bar{s}\left(G, \mathbf{y}\right)}{\hat{\sigma}},$$

where $\hat{\sigma}$ is the estimated asymptotic standard deviation of the score differences $s\left(F_i, \mathbf{y}_i\right) - s\left(G_i, \mathbf{y}_i\right)$ for $i = 1, ..., k$. Values greater zero imply superiority of forecasts F_i, while negative values show that forecasts G_i are preferred.

The null hypothesis states equal predictive performance of both models. Under this hypothesis and some regularity assumptions, the test statistic T_k asymptotically follows a standard normal distribution. To assess the statistical significance of the differences in scores, we calculate the test statistic and the corresponding p-values. The DM test can easily be implemented using the R package `forecast` (Hyndman and Khandakar, 2008).

Chapter 4

Postprocessing for TIGGE forecasts

Embedded in the World Meteorological Organization (WMO), The Observing System Research and Predictability Experiment (THORPEX) was an international research program, lasting from 2005 to 2014, with the aim to improve skill of weather forecasts for 1- up to 16-days ahead. Part of its objectives was the development of The Interactive Grand Global Ensemble (TIGGE), first known as THORPEX Interactive Grand Global Ensemble. As Ebert (2001) notes, combining forecasts from different operational NWP centers generates an ensemble prediction system which can outperform the contributing individual NWP models. In this spirit, TIGGE collects ensemble forecasts from twelve internationally operating NWP centers to construct a multi-model ensemble, providing forecasts on a global grid since the beginning of the project in 2006. Due to the research objective of the task, these predictions are not available in real time, but with a delay of 48 hours to avoid commercial exploitation.

The TIGGE archive provides 6-hourly forecasts for a wide range of weather quantities (also called parameters or variables) such as surface pressure, total precipitation, surface temperature and total cloud cover. A complete list of parameters can be found in Bougeault et al. (2010), who discuss the objectives of the project and give detailed information on the contributing ensemble prediction systems (EPSs). Furthermore, Park et al. (2008) describe the characteristics of the contributing EPSs and analyze their strengths and weaknesses, while also providing first verification results for forecasts combined from different sources. Additional information on TIGGE can be found in Table 4.1 or on the TIGGE website at ECMWF `https://confluence.ecmwf.int/display/TIGGE`.

After a decade in existence, Swinbank et al. (2016) review TIGGE's achievements and conclude that the data base has supported the development of new forecast products for extreme weather events and inspired a wide range of scientific studies. In particular, they highlight research results by Hagedorn et al. (2012), who find that reforecast-calibrated ECMWF forecasts are of comparable or even superior quality than the solely bias corrected TIGGE ensemble for 2m temperature over the Northern Hemisphere.

Table 4.1: Information on TIGGE's contributing sub-ensembles evaluated in the case studies. The table displays forecast horizon, the archived resolution and number of ensemble members with +1 indicating the control run. Two centers marked by asterisks only contribute to the updated TIGGE data considered in Section 4.2; the remaining nine centers are used in the reproduction of the study by Hagedorn et al. (2012).

Center	Acronym	No.	Resolution archived	Forecast horizon (days)
Bureau of Meteorology, Australia	BOM	32+1	1.5°×1.5°	10
China Meteorological Administration	CMA	14+1	0.56°×0.56°	16
Canadian Meteorological Centre	CMC	20+1	1.0°×1.0°	16
Centro de Previsa de Temp e Estudos Climáticos, Brazil	CPTEC	14+1	∼1.0°×1.0°	15
Environment and Climate Change Canada*	ECCC	20+1	1.0°×1.0°	16
European Centre for Medium-Range Weather Forecasts	ECMWF	50+1	∼0.5°×0.5°	15
Japan Meteorological Agency	JMA	50+1	1.25°×1.25°	9
Korea Meteorological Administration	KMA	16+1	1.25°×1.25°	10
National Centers for Environmental Prediction, USA	NCEP	20+1	1.0°×1.0°	16
National Center for Medium Range Weather Forecasting, India*	NCMRWF	11+1	0.18°×0.12°	13
United Kingdom Meteorological Office	UKMO	23+1	∼1.25°×0.83°	15

In the context of statistical postprocessing for TIGGE, different techniques have been applied to various contributing sub-ensembles individually – for the univariate case see e.g. Vogel et al. (2018) or Tao et al. (2014), while Aminyavari and Saghafian (2019) focus on spatial postprocessing by applying the non-parametric ECC method (see Section 3.2.1) to precipitation forecasts. Incorporation of the multi-model structure can be found in Barnes et al. (2019), who combine forecasts by three of the sub-ensembles within a Bayesian framework. We aim to utilize the full TIGGE ensemble in postprocessing while simultaneously accounting for the spatial dependencies.

At the beginning of this chapter, we reproduce parts of the verification results for TIGGE by Hagedorn et al. (2012) as a reference and then apply more advanced postprocessing techniques. The objective is to determine one univariate version of the EMOS method (see Section 2.2) which yields the best performance results for the full TIGGE ensemble. After successfully calibrating the marginal distributions, we apply multivariate postprocessing methods (see Chapter 3) to generate calibrated forecasts fields on the entire globe. During this process, we update to a more current and larger data set from the TIGGE archive. To this set, we apply the findings from before and conduct further analyses with a focus on spatial postprocessing.

4.1 Replicating Hagedorn et al. (2012) and beyond

4.1.1 Data set

Hagedorn et al. (2012) compare TIGGE multi-model forecasts with reforecast-calibrated ECMWF ensemble forecasts. Their analysis includes evaluation of 2-days ahead predictions for 2m temperature issued by TIGGE from December 1, 2008 until February 28, 2009. They assess forecasts initialized at 12 as well as 00 universal time coordinated (UTC) with a focus on the latter. For predictions started at 12 UTC, nine out of the contributing ensembles are available (see models without asterisk in Table 4.1), whereas at 00 UTC a smaller subset is given – namely CMC, CMA, CPTEC, ECMWF, KMA, NCEP and UKMO. Hagedorn et al. (2012) furthermore evaluated the performance of a sub-multi-model ensemble, called TIGGE-4, which consists of the ensemble prediction systems run by CMC, ECMWF, NCEP and UKMO. Throughout the chapter, the unit used is degrees Celsius.

Choosing a data set to train and verify the models is of crucial importance and discussed more extensively in Chapter 6. Hagedorn et al. (2012) contrast the use of the NCEP reanalysis (Kanamitsu et al., 2002) and ECMWF's ReAnalyses-Interim (ERA-Interim; Simmons et al., 2007; Dee et al., 2011); here we only employ the latter. While Hagedorn et al. (2012) evaluate forecast for lead times from 1 to 16 days, we restrict our analyses to 2-days ahead predictions. In their study, they consider the extra-tropical region over the Northern Hemisphere (20° - 90°N), interpolated to a

Table 4.2: Averaged scores and assessment of the prediction interval for bias corrected 2-days ahead temperature forecasts initialized at 00 UTC over the Northern Hemisphere from December 1, 2008 until February 28, 2009. The predictions are verified against ERA-Interim reanalyses as in Figure 9a in Hagedorn et al. (2012). The last column shows the ratio of the empirical coverage and nominal prediction interval.

				Prediction Intervals			
Model	CRPS [°C]	AE [°C]	RMSE [°C]	Width [°C]	Coverage [%]	Level [%]	Ratio [%]
TIGGE	0.90	1.21	1.83	9.07	97.00	98.79	**98.19**
TIGGE-4	**0.85**	**1.16**	**1.70**	7.75	95.44	98.31	97.08
CMA	1.41	1.74	2.58	2.68	42.63	87.50	48.72
CMC	1.20	1.66	2.37	5.59	78.72	90.91	**86.60**
CPETC	2.00	2.08	3.00	0.54	6.84	87.50	7.82
ECMWF	**0.96**	**1.23**	**1.80**	3.58	66.32	96.15	68.98
KMA	1.97	2.24	3.51	2.27	35.40	88.89	39.82
NCEP	1.33	1.62	2.59	2.72	51.93	90.91	57.12
UKMO	1.23	1.57	2.32	3.38	58.99	92.00	64.12

$2.5° \times 2.5°$ grid by the ECMWF TIGGE data portal, resulting in 4.176 grid points. When calculating the verification measures over areas, Hagedorn et al. (2012) propose to weight the score at a certain grid point by the corresponding cosine latitude before averaging them to properly represent verification results on the grid across a sphere; we follow this recommendation.

4.1.2 Performance on the Northern Hemisphere

In Figure 9 (a) of Hagedorn et al. (2012), the authors evaluate the performance of forecasts with a lead time from 1- up to 16-days ahead by selected TIGGE models and the reforecast-calibrated ECMWF EPS. The TIGGE set consists of predictions by four sub-ensembles – namely CMC, ECMWF, NCEP and UKMO – and the combination of them, resulting in TIGGE-4; all of these ensembles are individually bias corrected with a training period of 30 days (see Section 2.3). We partially reproduce the study from 2012 by only considering 2-days ahead predictions from TIGGE.

Hagedorn et al. (2012) state that ECMWF provides the most skillful contributing EPS and in terms of CRPS, TIGGE-4 yields a lower mean score than the entire TIGGE, which our findings confirm as shown in Table 4.2. Apart from the CRPS, we evaluate the predictive performance with mean AE and RMSE – again supporting this ranking. To solely assess calibration, for each M-member ensemble the table shows the empirical coverage and width of the nominal $\frac{M-1}{M+1} \cdot 100\%$ prediction interval, which corresponds to the ensemble range. For comparable results between ensembles of different sizes, the table presents the ratio between the real and nominal coverage as a percentage. For

Table 4.3: Scores for raw 2-days ahead temperature forecasts by TIGGE and sub-ensembles initialized at 00 UTC over the Northern Hemisphere verified against ERA-Interim reanalyses averaged over December 1, 2008 until February 28, 2009. The last column shows the ratio of the empirical coverage and nominal prediction interval.

				Prediction Intervals			
Model	CRPS [°C]	AE [°C]	RMSE [°C]	Width [°C]	Coverage [%]	Level [%]	Ratio [%]
TIGGE	1.07	1.45	2.25	9.07	95.57	98.79	**96.74**
TIGGE-4	**1.00**	**1.37**	**2.00**	7.75	93.79	98.31	95.40
CMA	1.92	2.28	3.30	2.68	36.05	87.50	41.20
CMC	1.37	1.90	2.68	5.59	74.92	90.91	**82.41**
CPTEC	2.47	2.55	3.68	0.54	5.23	87.50	5.98
ECMWF	**1.18**	**1.48**	**2.13**	3.58	60.18	96.15	62.59
KMA	2.92	3.21	5.19	2.27	30.38	88.89	34.18
NCEP	1.63	1.94	2.96	2.72	47.52	90.91	52.27
UKMO	1.55	1.91	2.82	3.38	54.58	92.00	59.32

both TIGGE ensembles the coverage of the empirical predictive interval is closest to its nominal value highlighting the benefits of the multi-model system while most (single-model) ensembles underestimate the uncertainty in the forecasts (as in for example Park et al., 2008; Bougeault et al., 2010). Among the sub-ensembles, the empirical coverage of CMC and ECMWF demonstrates the smallest deviation from the nominal value.

The overall quality of the predictions by TIGGE and the contributing ensembles is remarkable, as also presented in Table 4.3 which displays the performance of the direct model output. The TIGGE-4 ensemble shows averaged CRPS, AE and RMSE of 1.00, 1.37 and 2.00, respectively. Similar to the bias corrected forecasts, TIGGE-4 is the most skillful ensemble and both multi-models outperform every individual contributing EPS. Among the sub-ensembles, ECMWF yields the best scores, followed by CMC and UKMO. Comparing Table 4.2 and Table 4.3, the average width of the predictions intervals is the same, while there is a minimal gain in empirical coverage for the simply postprocessed ensembles because of the interval shift. Bias correction in particular improves forecasts in terms of AE and RMSE, which in turn results in slight improvements in CRPS.

The comparison of uPIT histograms for the direct model output and the bias corrected forecasts in Figure 4.1 again displays minor benefits through simple postprocessing, as the highest bars of the histograms are lowered. Since the predictive variance is not manipulated by the bias correction, underdispersion cannot profoundly be corrected resulting in solely small changes to the shape of the histograms. Especially Figure 4.1 highlights the benefits of the multi-model ensemble, as histograms of both TIGGE variants demonstrate that this ensemble is closer to uniformity and thus calibration, while all contributing sub-ensembles show some sort of dispersion error. For CMC and ECMWF, this underdispersion is less pronounced, which is also supported by the smaller

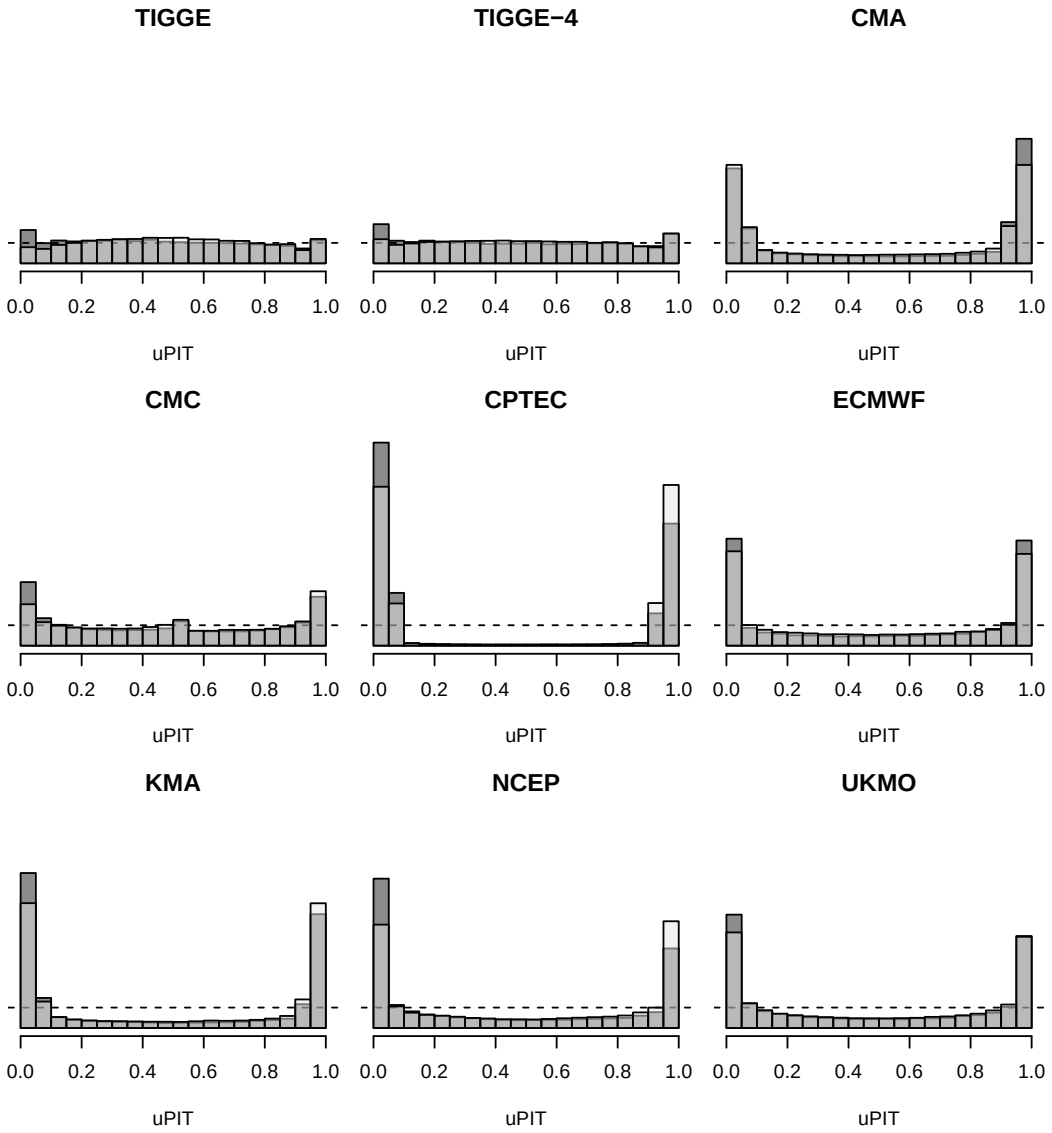

Figure 4.1: uPIT histograms for raw and bias corrected 2-days ahead temperature forecasts initialized at 00 UTC over the Northern Hemisphere and verified against ERA-Interim for December 1, 2008 to February 28, 2009. The dark gray bars coincide with values for the raw predictions, while the light gray bars indicate results of the bias corrected ensembles; medium gray denotes the overlapping area.

deviation of the averaged empirical coverage from the nominal predication intervals in Tables 4.2 and 4.3. The forecasts by CPTEC strongly underestimate the inherent uncertainty, already foreshadowed by the low mean width of the prediction interval of 0.54°C. According to their individual histograms, each of these sub-ensembles would benefit from sophisticated postprocessing to correct the predictive variances and biases.

Table 4.4: Global scores for unprocessed 2-days ahead temperature forecasts by TIGGE and individual sub-ensembles initialized at 00 UTC verified against ERA-Interim reanalysis averaged over December 1, 2008 until February 28, 2009.

Model	CRPS [°C]	AE [°C]
TIGGE	0.700	0.955
TIGGE-4	**0.672**	**0.917**
CMA	1.291	1.523
CMC	0.917	1.234
CPTEC	1.524	1.599
ECMWF	**0.818**	**1.019**
KMA	1.782	1.886
NCEP	1.094	1.284
UKMO	1.057	1.262

The combination of these independent EPS widens the prediction intervals and thus the uncertainty associated with the forecasts, which the individual ensembles lack. As Hagedorn et al. (2005) state the prerequisite for success of a multi-model concept is based on the combination of skillful and independent EPSs with their own strengths and weakness. The TIGGE project might have achieved this aim, when considering the histograms in Figure 4.1 and performance results in Table 4.2 and 4.3.

4.1.3 Global analysis

Since the objective of this work is to model spatial dependence of forecast error fields across the entire sphere, we expand the TIGGE data set from Section 4.1.1 with the remaining 6,192 grid points on the globe (90°S – 20°N). Hence, we now evaluate predictions for December 1, 2008 until February 28, 2009 with a lead time of 2 days, still initialized at 00 UTC and interpolated to a $2.5° \times 2.5°$ grid, resulting in forecasts at a total of 10,224 sites across planet Earth (without North and South Pole) on each day. For a first glimpse at the global predictive performance, we apply different variants of EMOS and analyze how the predictions by TIGGE benefit from univariate postprocessing.

As a reference, Table 4.4 summarizes the verification results of the direct model output. Compared to Table 4.3, we see an overall decrease in the averaged scores. Through expanding the data set by the Southern Hemisphere, the proportion of grid points over the sea increases. As we will argue in Chapter 6, when verifying forecasts against reanalyses, especially for sites over water, where surface observations are spread sparse, the reanalysis and the model output produce similar values. This results in low (good) performance scores for forecasts. The ranking of the EPSs is the same as seen in Table 4.3, hence no ensemble is specifically adapted to either the Northern or the Southern Hemisphere.

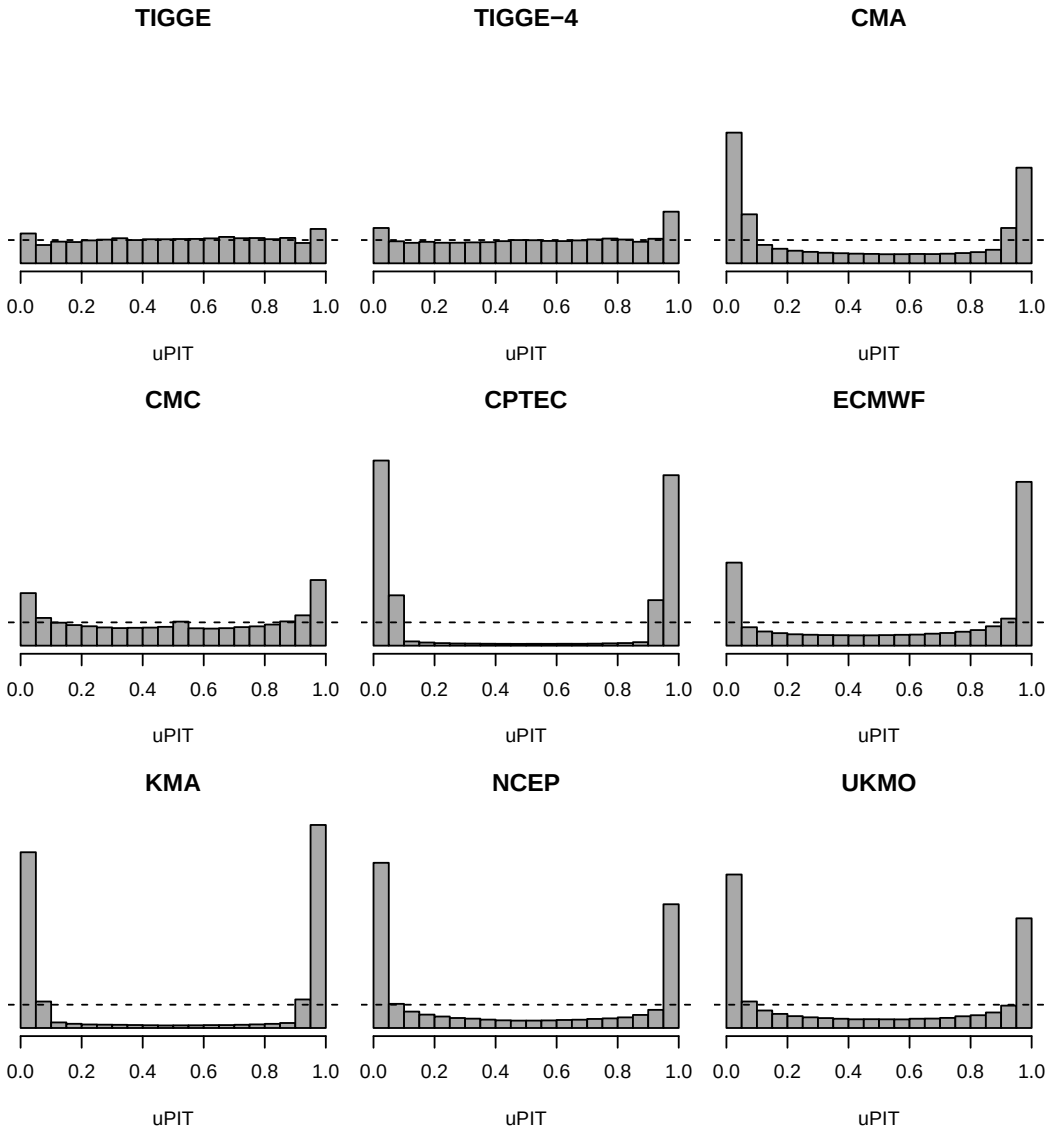

Figure 4.2: uPIT histograms for global raw TIGGE 2-days ahead temperature forecasts and its sub-ensembles initialized at 00 UTC. The predictions are verified against ERA-Interim reanalysis and accumulated globally during December 1, 2008 until February 28, 2009.

The uPIT histograms for the raw forecasts are displayed in Figure 4.2 and indicate the need for statistical postprocessing for all sub-ensembles, as their predictive spreads are too narrow, leading to uncalibrated forecasts. When uniting these truly independent ensembles, both combinations, TIGGE and TIGGE-4, represent the uncertainty rather well in comparison to its individual contributors. Because TIGGE and TIGGE-4 provide almost perfectly calibrated predictions, the improvements achieved through postprocessing might be smaller than for the single models.

To explore the benefits of more sophisticated postprocessing for TIGGE, firstly we apply EMOS with the ensemble mean as predictor. In the context of multi-models, numerous studies have demonstrated that the average of all ensemble members can outperform the most skillful contributing single model (e.g. Hagedorn et al., 2005 and Doblas-Reyes et al., 2005). With this in mind, we reduce TIGGE to its mean in order to make the application of EMOS more computationally feasible, as the entire original ensemble contains 164 individual members. When selecting the site-specific training data sets, we rely on local estimation and the neighborhood variants as described in Section 2.2.2. The performance results in terms of averaged CPRS and AE are shown in the first row of Figure 4.3 for different lengths of the training period from 20 up to 50 days. Although TIGGE is not as underdispersive as most EPSs, postprocessing still increases the quality of the forecasts. In this setting, including data from neighboring grid points for the estimation of the EMOS parameters does not improve the performance when compared to the local approach. Restricting neighborhoods to only contain grid points from the same surface type (land/sea) as the grid point at hand improves the mean scores marginally. However, the comparison is almost obsolete because the score difference between best and worst performing model is quite small ($<0.034°$C). For local EMOS, the skill of the forecasts improves with increasing training period length. Due to the small data set, we cannot extend the training period beyond 50 days to find a local minimum, but will eventually analyze this issue further in Section 4.2.

In order to incorporate the multi-model structure of the TIGGE forecasts, we calculate the mean of each TIGGE sub-ensemble and unite them, thus creating a small ensemble of in this scenario seven members. To these combined predictive means, we apply EMOS with the neighborhood approach for training data selection. The forecast skill further improves when compared to the simple mean version – see the second row of Figure 4.3. In this scenario, most neighborhood versions outperform local EMOS. Compared to the mean-only approach from before, the number of estimated parameters for EMOS has increased from 4 in Eq. (2.2) to 10 in Eq (2.1). Hence, the local version might not provide enough data for a stable estimation. The variant NN4 with a differentiation for land/sea yields the lowest scores in terms of both averaged CRPS and averaged AE.

To not only explore variants of EMOS for TIGGE focusing on the predictive mean, but also the predictive variance, we incorporated the ensemble spread in three different appreoaches within local EMOS applied to TIGGE's mean. These results are shown in the bottom row of Figure 4.3. For "Var 1", the parameter v of the EMOS Eq. (2.2) denotes the empirical standard deviation over the seven means of the contributing ensembles. The parameter v represents the standard deviation of all 164 individual members in "Var 2" as in the original approach. For "Var 3" the predictive distribution equals $\mathcal{N}(a_s + b_s\bar{f}_s, c_s + d_{1,s}v_{1,s}^2 + \cdots + d_{7,s}v_{7,s}^2)$, where v_i, $i \in \{1, .., 7\}$ describes the empirical standard deviation of each sub-ensemble and the d's are estimated within the regular EMOS fit. As a reference, "Var 0" is a simple EMOS version, that uses a fixed, estimated variance which does not depend on the ensemble spread; the

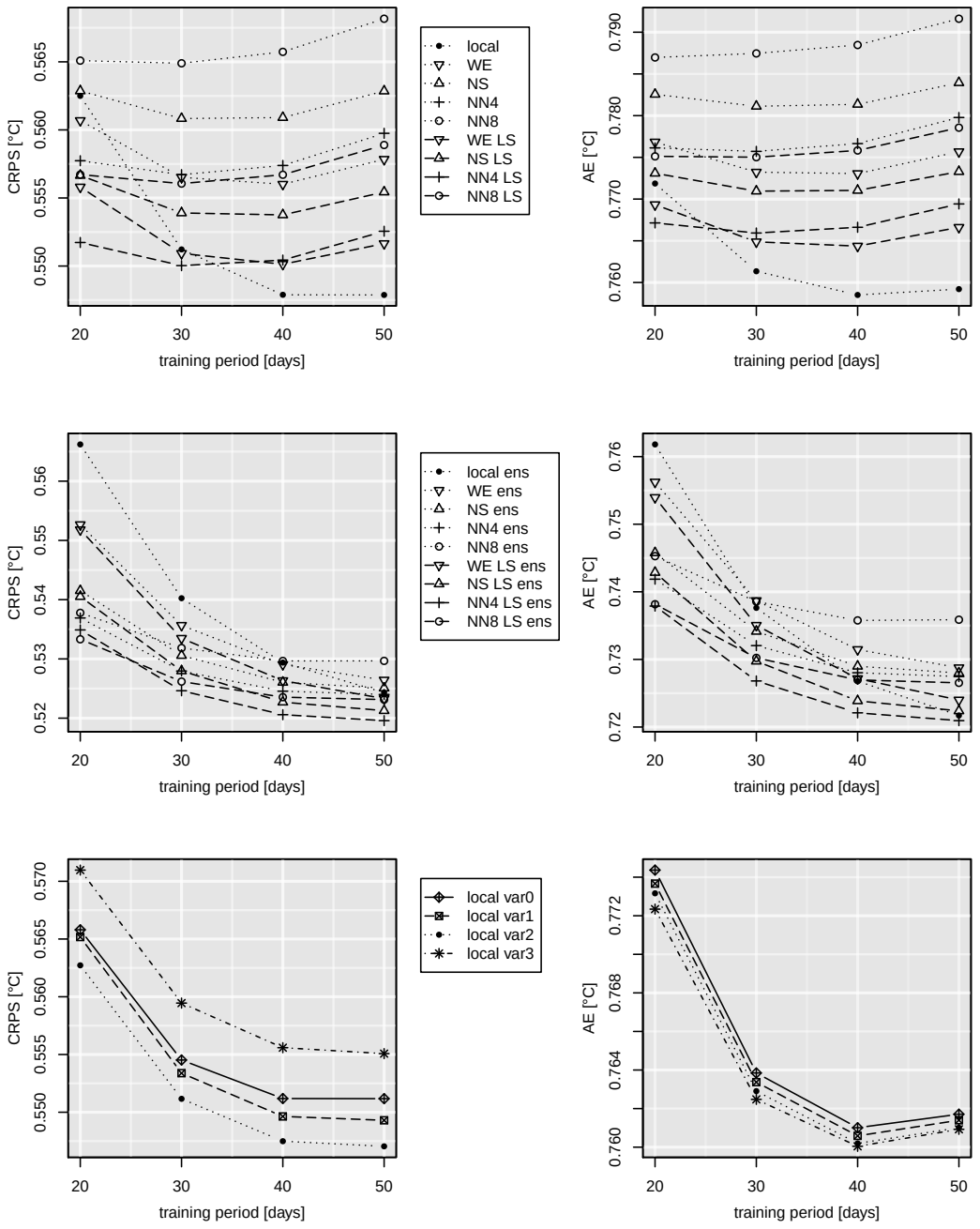

Figure 4.3: Averaged CRPS and AE for EMOS applied to the mean of the TIGGE and combined means of TIGGE's sub-ensembles with different neighborhood training sets and multiple formulations for the estimation of the predictive variance. The TIGGE forecasts for temperature are initialized at 00 UTC for a prediction horizon of 2 days, verified against ERA-Interim reanalysis.

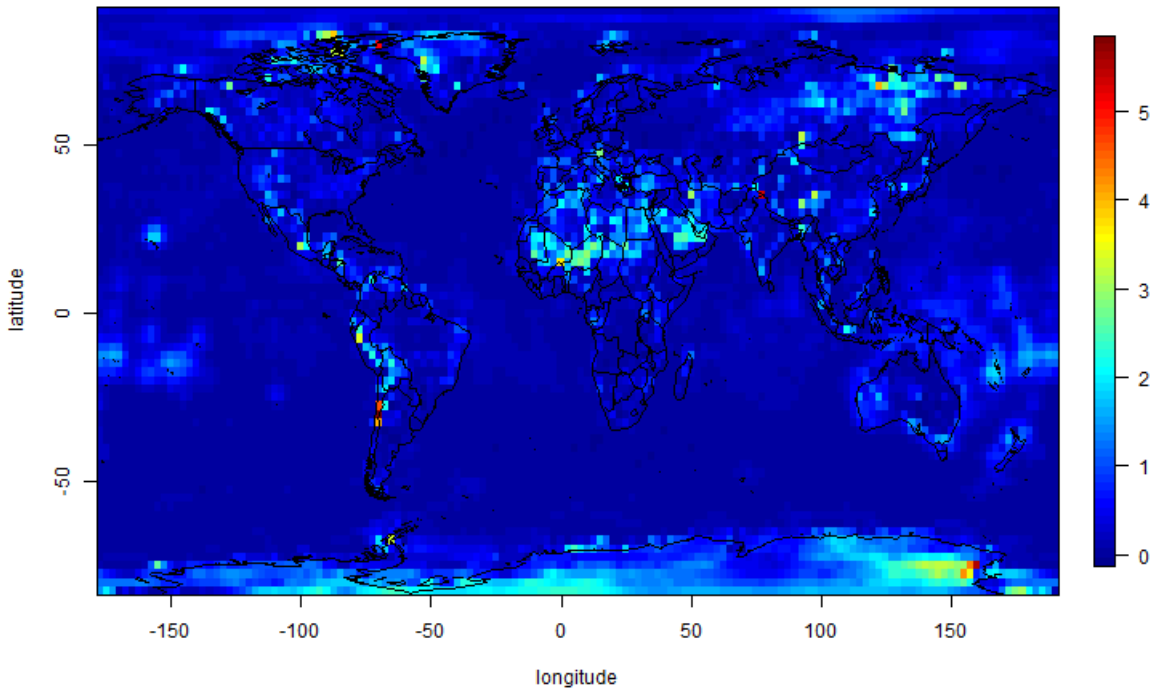

Figure 4.4: Difference in CRPS in degrees Celsius of local EMOS with 50 training days compared to the raw ensemble for 2-days ahead TIGGE temperature forecasts initialized at 00 UTC, verified against ERA-Interim reanalysis averaged over December 1, 2008 until February 28, 2009.

predictive distribution is thus $\mathcal{N}(a_s + b_s\bar{f}_s, c_s)$. In terms of performance, all models yield very similar results, with "Var 2" achieving the lowest scores, but again the overall differences in averaged scores are negligible, as they are smaller than 0.02°C. Due to these findings, we will continue to employ the variance according to "Var 2", namely $y_s|f_{1,s}, \ldots, f_{M,s} \sim \mathcal{N}(a_s + b_s\bar{f}_s, c_s + d_s v_s^2)$ as suggested in the original EMOS approach by Gneiting et al. (2005).

To analyze the spatial patterns in the predictive performance, Figure 4.4 shows the difference in terms of CRPS for the direct model output relative to postprocessed forecasts, namely application of local EMOS with 50 training days. In general, post-processing has a slight positive effect, with some areas showing larger improvements – e.g. northern Africa, Australia, eastern South America. Over ocean areas, there is almost no change to the direct model output. Again, this might be explained by the fact that observational sites over the ocean are sparsely distributed, so the reanalyses and the forecasting model produce similar values and the benefits from postprocessing diminish.

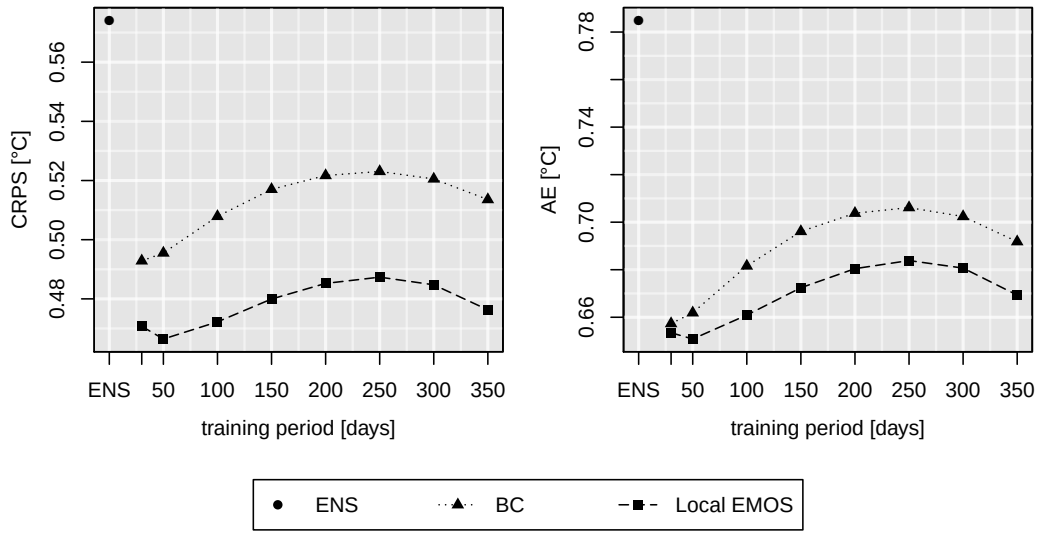

Figure 4.5: Globally averaged CRPS for EMOS postprocessed, bias corrected (BC) and raw TIGGE 2-days ahead temperature forecasts initialized at 12 UTC. The predictions are verified against ERA5 and accumulated over January 1, 2010 through December 31, 2018.

4.2 Updated TIGGE

4.2.1 Data set

After recreating part of the work by Hagedorn et al. (2012) and applying EMOS to the same data set which they analyzed in their case study, we enlarge the TIGGE data to cover the time period of 1.01.2009 until 31.12.2018 and update the reference data set to the more recently developed ERA5 reanalyses (Hersbach et al., 2020). The evaluation time frame ranges from 1.01.2010 to 31.12.2018 – for some analyses starting on 1.01.2012 or 1.12.2018. As before, the 2-days ahead temperature forecasts are spread on a $2.5° \times 2.5°$ grid across the globe, resulting in 10,224 grid points without the North and South Pole. To realize the full potential of the multi-model ensemble, we focus on 12 UTC, because ten of the contributing sub-ensembles are available for this initialization time – namely BOM, CMA, CPTEC, ECCC, ECMWF, JMA, KMA, NCEP, NCMRWF and UKMO. Predictions from some sub-ensembles are missing for certain days. When no forecasts to train or verify are given, we exclude the respective sub-ensemble from the EMOS fit for this specific time.

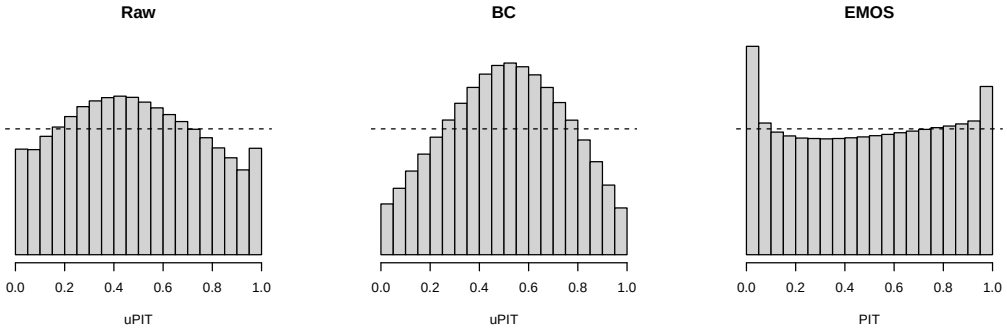

Figure 4.6: (Unified) PIT histograms for 2-days ahead global temperature forecasts initialized at 12 UTC and verified against ERA5 from January 1, 2010 until December 31, 2018 for the raw, bias corrected (BC) and EMOS postprocessed TIGGE ensemble. The length of training period is based on the best performance in terms of CRPS, so 30 days of data for BC and 50 days for EMOS.

4.2.2 Univariate postprocessing

As in Section 4.1.2, we begin postprocessing with application of the bias correction used by Hagedorn et al. (2012). The performance in terms of averaged CRPS and AE is displayed in Figure 4.5 for different lengths of training periods from 30, 50, 100, 150 up to 350 days. We compare the bias correction to simple local EMOS, which we apply to the TIGGE mean forecasts for the benefit of quick computation and stable parameter estimation. Yielding an averaged CRPS and AE of 0.566°C and 0.782°C, respectively, the raw TIGGE ensemble already delivers highly skillful predictions when evaluated against ERA5 reanalyses. The direct model output benefits from the application of bias correction as well as EMOS, but improvements in CRPS are more pronounced for EMOS. At short training periods, results of the mean AE are similar for both methods. These findings can be explained by the fact that the bias correction shifts the entire ensemble by the correcting term, but retains the original spread. Meanwhile EMOS can adjust the predictive variance to improve calibration of the forecasts which is reflected in the CRPS.

To solely evaluate calibration, Figure 4.6 displays uPIT histograms for each of the three forecasters. Contrary to most EPSs, TIGGE suffers from minimal overdispersion as the uPIT histogram shows a slight hump shape. This feature becomes more pronounced after the bias correction, when the ensemble forecasts are shifted and more observations fall close to the center of the predictive interval. The application of EMOS generally counteracts this flaw as the histogram almost flattens, while retaining one predominant bar on each side. We further explore possible reasons for this shape in Chapter 5.

For the previous and smaller data set, no optimal length of the training period could be determined (see Figure 4.3). The local EMOS curve shows a local minimum for a training period of 50 days in Figure 4.5. Beyond 250 days, the curve starts to decline and we are interested if there exists an even lower minimum at longer training periods.

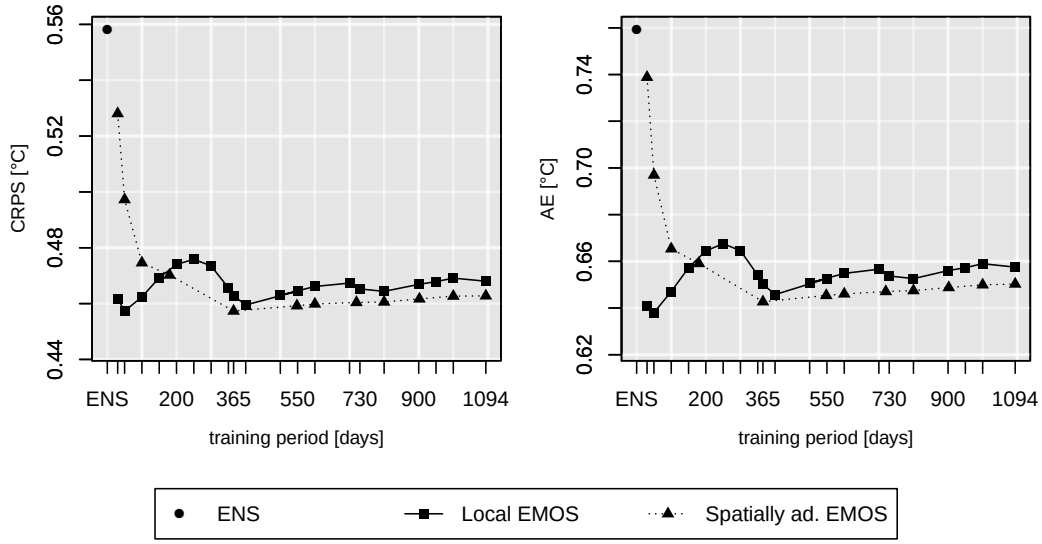

Figure 4.7: Globally averaged CRPS and AE for raw TIGGE temperature forecasts and two EMOS versions applied to TIGGE mean temperature forecasts initialized at 12 UTC. The 2-days ahead predictions are verified against ERA5 for January 1, 2012 to December 31, 2018.

Hence, we restrict the evaluation period from January 1, 2012 to December 31, 2018, which allows for a training set containing up to three years of data. Figure 4.7 displays the verification results for local EMOS and spatially adaptive EMOS (Hemri et al., 2014), described in Section 2.2.1. Due to the construction of spatially adaptive EMOS, we choose certain training periods relative to the length of a year: half a year, one, two and three years. The remaining majority of training periods lengths are multiples of 10. For local EMOS, there exists a second local minimum at a training period of 400 days with a CRPS of 0.460°C; however, this minimum is greater than the CRPS of 0.457°C at 50 days. Spatially adaptive EMOS outperforms local EMOS for training periods greater than 150 days and exhibits hardly any sensitivity to the length of the training period beyond 365 days, as the curve becomes almost flat. The local minimum at a year's worth of training data is associated with a CRPS of 0.457°C and thus nearly identical to the local minimum of local EMOS (first difference in the sixth decimal place). Because the latter is computationally faster due to the simplicity of the model and smaller training data set, we continue to focus on local EMOS.

At the beginning of this chapter, we noted that the forecast skill can be greatly improved by applying EMOS not to the overall ensemble mean, but to the ensemble of sub-ensemble means (see Figure 4.3). Paired with the NN4 LS training data selection, this EMOS variant generated the most skillful predictions. We apply these findings to the updated TIGGE data set and summarize the results in Figure 4.8, where we compare the performance of the direct model output to local EMOS fitted to the mean

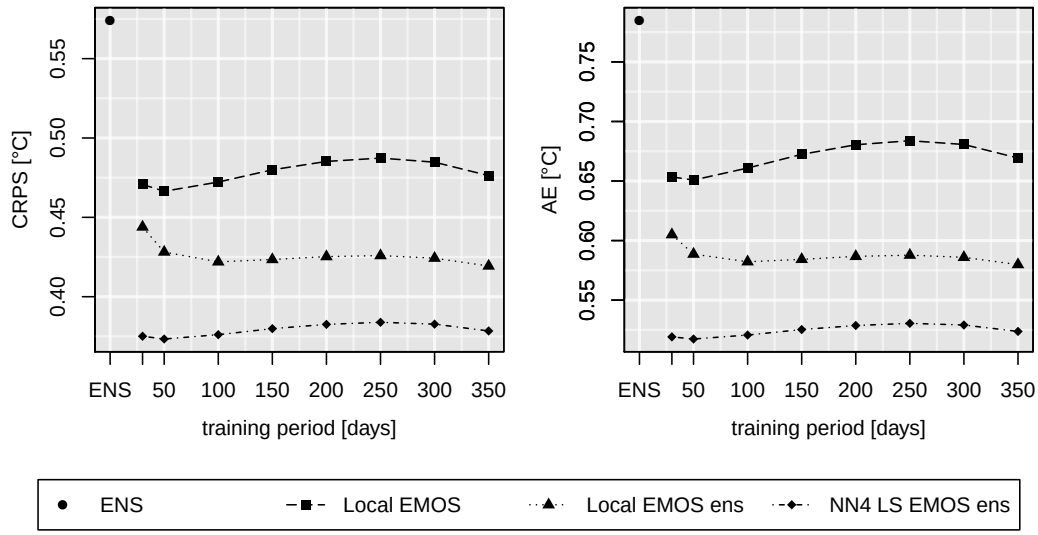

Figure 4.8: Globally averaged CRPS and AE for 2-days ahead temperature forecasts initialized at 12 UTC of raw TIGGE predictions and two EMOS variants, one applied to the TIGGE mean only and one applied to the combination of the means of the contributing sub-ensemble (ens). The forecasts are verified against ERA5 for January 1, 2010 through December 31, 2018.

as well as local EMOS and NN4 LS EMOS which are both applied to the sub-ensemble means. The results from the previous section are confirmed as local EMOS of the sub-ensembles outperforms local EMOS applied to the overall TIGGE mean, while NN4 LS EMOS based on the sub-ensembles yields the lowest scores. Especially, local EMOS of the sub-ensembles benefits from longer training periods as more data stabilizes the estimation of the enlarged set of EMOS parameters (now up to 13 instead of four). Similar to the mean local EMOS version, NN4 LS exhibits a minimum averaged CRPS at a training period of 50 days.

For the spatial applications, we will therefore base the univariate postprocessing on NN4 LS EMOS applied to the means of the contributing sub-ensembles. Considering the collection of means as an ensemble, we apply EMOS accordingly, which yields the distribution $y_s|f_{1,s},\dots,f_{M,s} \sim \mathcal{N}\left(a_s + b_{1,s}\bar{f}_{1,s} + \cdots + b_{k,s}\bar{f}_{k,s}, c_s + d_s v_s^2\right)$ for temperature variable y_s at location $s \in \mathcal{S}$ given forecasts $f_{1,s},\dots,f_{M,s}$ by an M-member ensemble with variance v_s^2. The mean forecast of each of the k sub-ensembles is denoted by $\bar{f}_{l,s},\dots,\bar{f}_{k,s}$ and $a_s, b_{1,s},\dots,b_{k,s}c_s, d_s$ are the EMOS parameters. The training set for their estimation consists of data over the past 50 days from the four neighboring grid points when the grid point is located on the same surface type (land or sea) as the point of interest. From here onward we refer to this specification as EMOS.

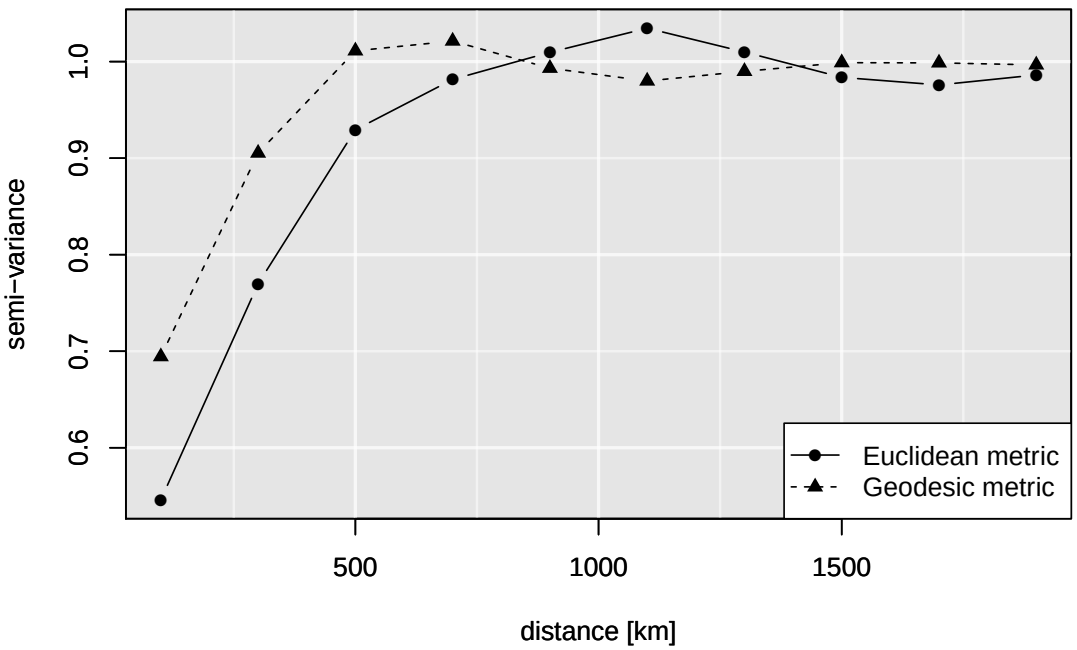

Figure 4.9: Binned variogram for standardized forecast errors by (NN4 LS sub-ensemble) EMOS postprocessed TIGGE temperature forecasts at a prediction horizon of 2 days initialized at 12 UTC and verified against ERA5 on the first day of the training period on 1.01.2017. The continuous and dotted lines show the averaged semi-variance of error pairs relative to the Euclidean and geodesic metric, respectively.

4.2.3 Spatial postprocessing

After determining the most suitable univariate EMOS version for the data at hand, we now combine spatial extensions with this postprocessing method. These approaches are discussed in detail in Chapter 3. Specifically, spatial EMOS is computationally highly demanding, so we restrict the verification period to the most current available year 2018, starting January 1 and ending December 31. Instead of a rolling window training period, we estimate the parameters of the correlation functions (Eq. 3.11) with maximum likelihood over forecast errors on all days of the entire year 2017.

In Section 3.3, we discussed the use of geodesic and Euclidean distances for covariance functions. Figure 4.9 displays the variogram of the forecast errors on the first day of the training data set, 1 January 2017, relative to these metrics. Depending on the distance function, different pairs of forecast errors fall into each of the ten bins of the variogram. Nevertheless, both variograms display a similar shape and indicate a correlation length of no greater than 1000km.

When verifying the spatial methods, the models are fitted simultaneously across the globe before a sample from the predictive distribution is drawn. To obtain a

Table 4.5: Globally averaged ES and VS for direct model output by TIGGE and postprocessed 2-days ahead forecasts of temperature field. Model runs were initialized at 12 UTC and forecasts are verified against ERA5 in 5×5 grid boxes (containing 25 grid points) starting January, 1, 2018 until December 31, 2018. The predictive sample is either generated by a random draw (R) or the calculation of equidistant quantiles (Q).

	ES [°C]	VS		
Forecast		$p = 0.5$	$p = 1$	$p = 2$
Raw ensemble	3.762	83.8	1057	515,015
EMOS-R	2.967	67.4	749	297,678
ECC-R	2.965	67.9	748	297,145
Schaake-R	2.960	67.4	747	297,056
EMOS-Q	2.884	68.5	684	264,524
ECC-Q	2.796	62.6	665	263,400
Schaake-Q	**2.791**	**61.0**	**657**	**263,066**
Spatial EMOS	2.955	66.1	735	294,681
Spatial EMOS emp cor	2.958	65.9	735	295,681
Spatial EMOS emp cor 1 year	2.976	66.4	780	322,774

fair comparison, the sample size each day is the same for all methods and based on the contributing sub-ensembles available in the training set for the verification day. When assessing the predictive performance, we restrict the forecasts fields to boxes of $5 \times 5 = 25$ neighboring grid points over the surface of planet Earth to ensure computational feasibility. We consider all possible boxes and assign the calculated score to the grid point at the center. The scores reported are weighted by their corresponding cosine latitude as in Hagedorn et al. (2012) and averaged across the respective boxes over space and time. Besides fluctuations due to sampling, the marginal predictive performance for all spatial methods coincides with that of univariate EMOS and we therefore refrain from reporting these results as Figure 4.8 shows this performance over the entire verification period.

Table 4.5 summarizes the spatial performance in terms of averaged ES and VS for the direct model output, EMOS and various spatial variants thereof. When obtaining a sample for EMOS, ECC and the Schaake shuffle, two different methods are used. We either draw a random sample (EMOS-R, ECC-R and Schaake-R) from the predictive distribution F or take the equidistant quantiles $F^{-1}\left(\frac{1}{M+1}\right), ..., F^{-1}\left(\frac{M}{M+1}\right)$ of the available M-member ensemble; the latter referred to as EMOS-Q, ECC-Q and Schaake-Q respectively. As mentioned before, we fit the parametric correlation function for spatial EMOS to the EMOS error fields of the entire previous year 2017 in order to save computational time. The non-parametric version of spatial EMOS relies on the empirical correlations either fitted over the same 50-days rolling window training period as the univariate EMOS or a single fit over the past year as for the parametric version.

All postprocessing methods improve the direct model output in terms of mean ES and VS. Most conclusions drawn from the scores coincide, but the VS is more sensitive

Table 4.6: Values of the test statistic for the two-sided DM test of equal predictive performance for the ES in Table 4.5. The comparison focuses on spatial EMOS with a parametric correlation function and Schaake-Q as benchmarks. Positive values indicate a superior predictive performance by the method in the top row, while negative values indicate a better performance of the model in the left column. All values are significant at the 0.1% level under the null hypothesis of equal predictive performance.

	Schaake-Q	Spatial EMOS
Raw ensemble	266	225
EMOS-R	210	11
ECC-R	201	8
Schaake-R	202	4
EMOS-Q	405	−71
ECC-Q	36	−160
Schaake-Q	—	−165
Spatial EMOS	165	—
Spatial EMOS emp cor	154	2
Spatial EMOS emp cor 1 year	157	30

to misspecification in the forecasted correlation structure, whereas the ES puts more emphasis on the predictive mean vector according to early works by Scheuerer and Hamill (2015b). The Schaake shuffle based on quantiles yields the lowest values for both scores, which confirms findings in the simulation studies by Lerch et al. (2020), who describe this method as a "powerful benchmark [...] that proves difficult to outperform". Hence, the spatial dependence structure of past verification sets contains more valuable information about the true dependence structure than the template based on the raw ensemble, since ECC-Q performs worse than Schaake-Q.

As Bröcker (2012) argues quantile forecasts are most suitable when evaluated by the CRPS, Schefzik et al. (2013) recommend using ECC-Q, which is supported by inference from Lerch et al. (2020). We find that within each sampling scheme (Q and R), the reference approaches are ordered as expected in terms of performance: most skillful forecasts by the Schaake shuffle, second place for ECC and lastly independent EMOS.

However, the sampling method impacts the ranking significantly. In Table 4 of Schefzik et al. (2013), the authors show a similar pattern for the verification results of their case study. ECC-Q outperforms ECC-R, but the univariate approach yields lower scores than ECC-R[1]. They argue variations in spatial dependence of temperature are a small scale phenomena and negligible at 400km distance, thus the benefit of ECC diminishes. Minimal distances between grid centers within the considered $5 \times 5 = 25$ neighboring grid boxes vary across the globe due to the segmentation into latitude and

[1]In terms of ES, the same ranking pattern for these ECC variants and the univariate postprocessing can be found in Table 3 of the case study by Aminyavari and Saghafian (2019). Additionally in some applications by Schefzik (2011), ECC-Q is outperformed by quantiles derived solely via univariate postprocessing.

Table 4.7: Globally averaged CRPS for 2-days ahead forecasts of the minimum (min), maximum (max) and average (ave) temperature initialized at 12 UTC verified against ERA5 in 5×5 grid boxes from January, 1, 2018 until December 31, 2018.

	CRPS [°C]			AE [°C]		
Forecast	min	max	ave	min	max	ave
Raw ensemble	0.698	0.474	0.344	0.936	0.634	0.475
EMOS-R	0.497	0.365	0.206	0.652	0.483	0.252
ECC-R	0.488	0.354	0.198	0.652	0.472	0.264
Schaake-R	0.485	0.354	0.193	0.646	0.474	0.260
EMOS-Q	0.484	0.353	0.197	0.653	0.483	**0.251**
ECC-Q	0.470	0.334	0.192	0.629	0.447	0.257
Schaake-Q	**0.461**	**0.326**	**0.188**	**0.613**	**0.439**	0.253
Spatial EMOS	0.491	0.359	0.201	0.650	0.478	0.265
Spatial EMOS emp cor	0.490	0.358	0.201	0.650	0.478	0.268
Spatial EMOS emp cor 1 year	0.498	0.365	0.210	0.662	0.488	0.284

longitude. Near the poles, centers can be located as close as 25km, whereas at the equator the minimal distance between grid centers measures 278km. Thus especially for grid boxes around the equator, we notice a similar pattern as Schefzik et al. (2013). Furthermore, the weighing according to cosine latitude puts more emphasis on scores at the equator, where grid centers are located further apart, and the corresponding scores have a higher impact on the overall results.

For all versions of spatial EMOS, we draw a random sample from the multivariate normal predictive distribution instead of quantiles. Spatial EMOS with parametric correlation functions outperforms the raw ensemble, and all random-sample-based approaches, but performs worse than forecasts derived from quantiles. The application with a parametric correlation function yields slightly better performance than the non-parametric, empirical correlation. This suggests that the structure we assumed for the correlation functions (presented in Sections 3.4) is reasonable, supported by the improved performance. Spatial EMOS with non-parametric correlation functions benefits when using the same sliding window training period as for univariate EMOS instead of one fixed training set which covers the previous year.

Besides the methods based on the random sample scheme, all postprocessing which accounts for spatial dependency outperforms the univariate approaches in terms of VS. The spatial EMOS versions deliver more skillful predictions than the sample variants of ECC and the Schaake shuffle, but based on quantiles the reference methods yield better scores. In terms of VS, spatial EMOS paired with an empirical correlation fitted on a sliding window training period outperforms the same model paired with a constant training set and the model using a parametric correlation function. This indicates that the spatial correlation structure changes over the course of a year and that spatial

Table 4.8: Values of the test statistic for the two-sided DM test of equal predictive performance for the CRPS minimum temperature forecasts in Table 4.7. The comparison focuses on spatial EMOS with a parametric correlation function and Schaake-Q as benchmarks. Positive values indicate a superior predictive performance by the method in the top row, while negative values indicate a better performance of the model in the left column. All values are significant at the 0.1% level under the null hypothesis of equal predictive performance.

	Schaake-Q	Spatial EMOS
Raw ensemble	149	123
EMOS-R	44	8
ECC-R	40	−3
Schaake-R	35	−7
EMOS-Q	76	−8
ECC-Q	37	−28
Schaake-Q	—	−40
Spatial EMOS	40	—
Spatial EMOS emp cor	41	0
Spatial EMOS emp cor 1 year	44	7

EMOS with the parametric correlation function might benefit from also using a sliding window training set.

To evaluate the statistical significance of the differences in scores, we apply the DM test, as described in Section 3.5.2. Because the application of the DM test requires independence of the scores, we only consider scores from non-overlapping blocks of $5 \times 5 = 25$ grid boxes. So instead of collecting 9360 scores daily, we evaluate a subset of size 377. In particular, we are interested in the performance in terms of averaged ES of the best model, Schaake-Q, and the newly proposed spatial EMOS with parametric correlation functions. The test statistics for both of these approaches compared to all other methods are shown in Table 4.6. Although the score differences in Table 4.5 are rather small, they are highly significant being based on 137,605 forecast cases and all p-values are smaller than 0.1%. The test statistics support the performance ranks according to the ES as presented in Table 4.5.

Spatial dependence structure is of critical importance when predicting aggregated univariate quantities, such as minima, maxima, totals, or averages over a region. In Table 4.7, we consider forecasts for these quantities over the sets of 5×5 grid points evaluated by the univariate scores CRPS and AE. Some results are similar to the conclusions from VS and ES, in that all postprocessing improves the direct model output and Schaake-Q yields the best results in most settings. Again, the sampling method has a large impact on the scores and for the majority of cases the forecasts based on quantiles outperform the random sampling. Spatial EMOS with a parametric (but constant over time) correlation function shows a similar predictive skill as spatial EMOS with an empirical correlation estimated over a sliding window training period. This

again supports the idea of pairing spatial EMOS with a parametric correlation fitted on a sliding window training period. Similar to findings by Schefzik (2011), Table 4.7 shows that in certain settings EMOS can outperform ECC, e.g. when there is small spatial dependence across the domain.

In order to ensure statistical significance of the differences in scores in Table 4.7, again we apply the DM test, as described in the paragraph before. Table 4.8 displays representationally the test statistics in terms of CRPS for minimum temperature predictions for the best performing model, Schaake-Q, and the newly proposed spatial EMOS with parametric correlation functions. Compared to Table 4.7 results for the test statistics are smaller, but as in the previous performance evaluation, Schaake-Q delivers the most skillful predictions. All postprocessing improves upon the raw ensemble forecasts, but compared to spatial EMOS, other postprocessing techniques perform better. However, these improvements are small (see Table 4.7), but again highly significant because the evaluation is based on 137,605 forecast cases and all p-values are smaller than 0.1%. The test statistics support the performance ranks according to the CRPS presented in Table 4.7.

To evaluate multivariate calibration, we consider the band depth, average and multivariate rank histograms, presented in Section 3.5.1, as well as the univariate uPIT for the aggregated forecast quantities. Because the number of contributing sub-ensembles changes over the time range considered, we calculate unified versions of the multivariate histograms in the spirit of the uPIT proposed by Vogel et al. (2018). The smallest number of contributing sub-ensembles available on a specific day is five, so we choose to plot histograms with six bins for a sound comparison. Spatial EMOS based on a non-parametric correlation function fitted over a sliding window training period outperforms the version with a fixed training period for all considered scores. From here onward, we will only report multivariate calibration results for the more skillfull variant.

Figure 4.10 displays the unified band depth rank histograms. Both univariate EMOS versions neglect spatial dependence by design, which can be detected from the $\cup$-shape of the corresponding histograms indicating that the correlation in the forecasts is too low. Similar to the results in Table 4.5, the quantile-based variants of ECC and the Schaake shuffle perform better as their histograms are flatter than those of the random sampling approaches. The raw ensemble, ECC-R, ECC-Q, Schaake-R and Schaake-Q display similar $\cap$-shapes, which are associated with too strong correlation patterns in the predictions. As the resolution of all TIGGE's sub-ensembles is lower than the native resolution of approximately $\sim 0.281° \times 0.281°$ for the verification set ERA5, forecast models might neglect local weather phenomena. Without these local variations in temperature, the forecast ensembles can overestimate the correlation of in reality independent grid points. In contrast the verification set accounts for these small-scale effects and associates the corresponding grid points with a smaller correlation, resulting in the $\cap$-shape histograms for the raw ensemble and ECC. Within the non-parametric methods, especially Schaake-Q shows this particular shape the least, as the dependence pattern is based on past verifications. In this case, spatial postprocessing benefits from

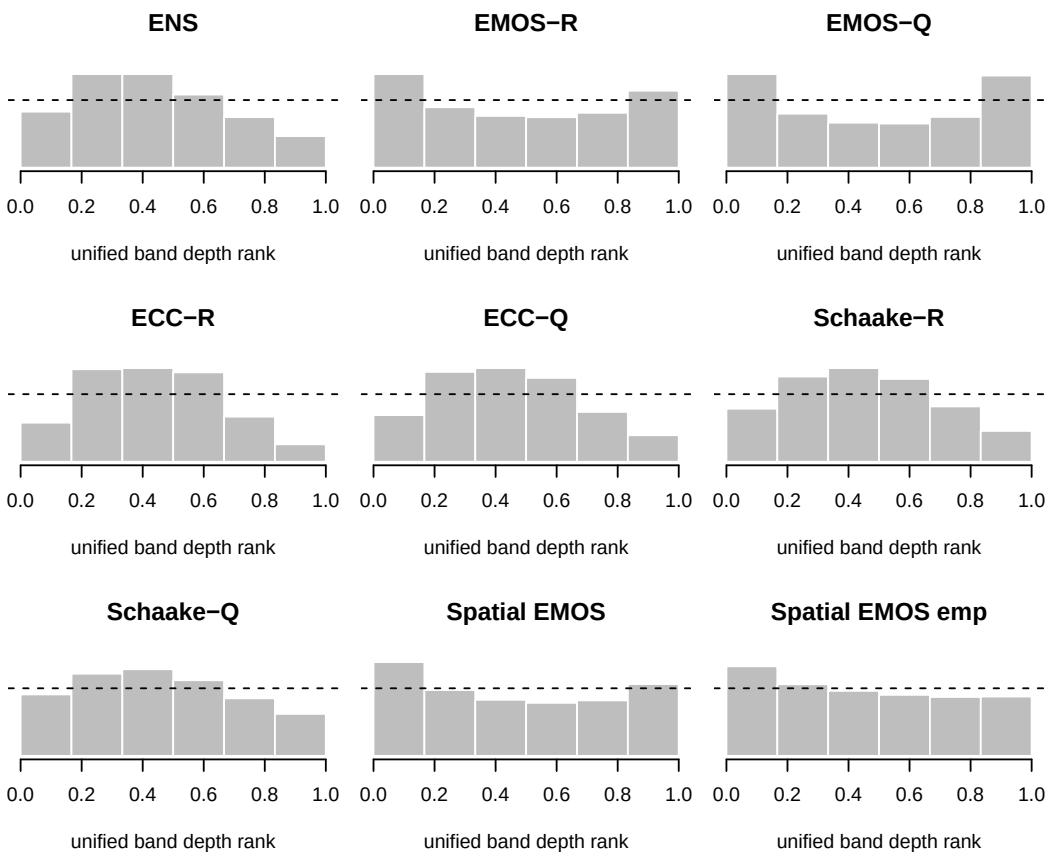

Figure 4.10: Unified band depth rank histograms for raw and postprocessed TIGGE temperature forecasts initialized at 12 UTC and verified against ERA5 across 5×5 grid boxes at a prediction horizon of 2 days accumulated over the test period from starting January, 1, 2018 until December 31, 2018.

fitting a correlation function which does not only mimic the spatial pattern of the verification set or raw EPS as indicated by the flatter histograms for spatial EMOS with non-parametric correlation functions.

The shapes of histograms for unified average ranking, displayed in Figure 4.11 and band depth ranking (see Figure 4.10) almost coincide, but can differ in interpretation. According to Thorarinsdottir et al. (2016), a $\cap$-shaped histogram implies too high correlation in terms of the average ranking and lack of predictive correlation when evaluated with band depth ranks. A $\cup$-shape histogram implies the opposite interpretations for each diagnostic tool. When evaluating against reanalyses instead of observations, often the forecasts and the verification coincide, discussed further in Chapter 6. Due to this dependency and thereby limited room for benefits through postprocessing, inferences drawn from these tools might still be inconclusive.

To nevertheless further understand the shape of the histograms for ECC, we reconsider the argument by Schefzik et al. (2013) stating that the benefit of ECC diminishes

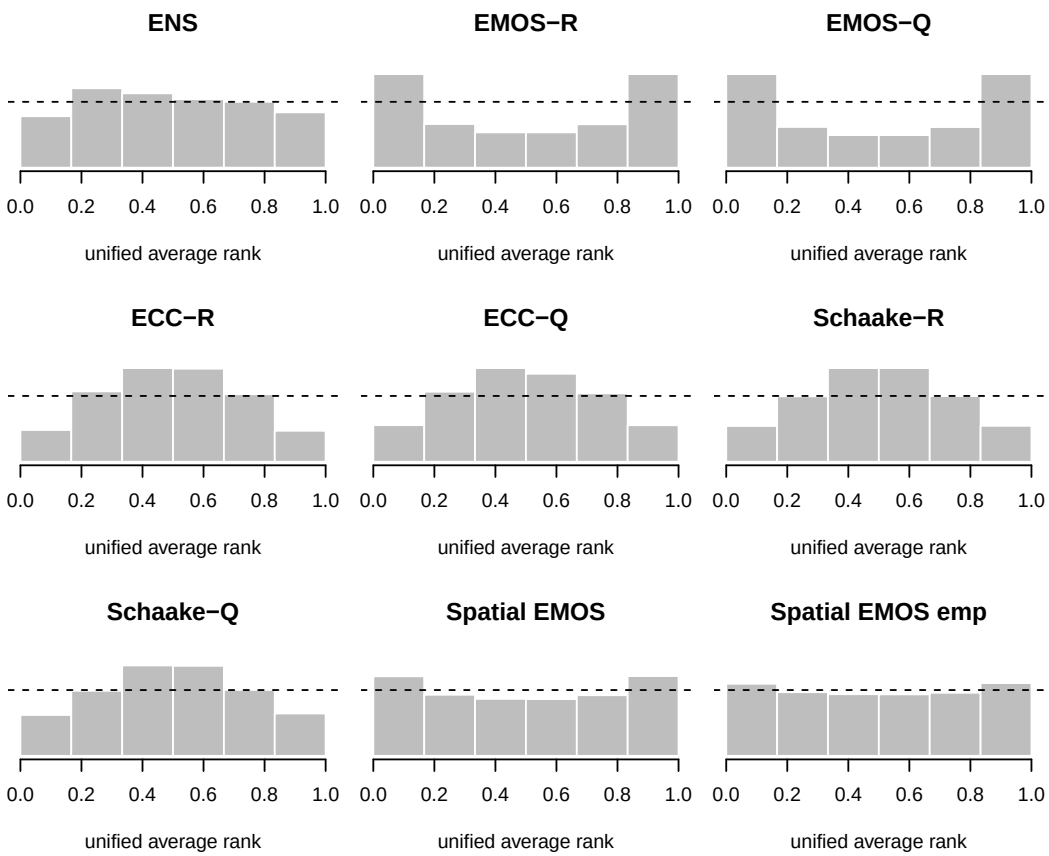

Figure 4.11: Unified average rank histograms for raw and postprocessed TIGGE temperature forecasts initialized at 12 UTC verified against ERA5 across 5×5 grid boxes at a prediction horizon of 2 days accumulated over the test period from January 1, 2018 until December 31, 2018.

beyond 400km distance between sites in case of temperature forecasts. The spatial evaluation within $5 \times 5 = 25$ grid boxes results in distances up to 786km among grid centers. In order to reduce these distances to a more meaningful size, we additionally assess ECC-Q forecasts in grid boxes of size $3 \times 3 = 9$ and $2 \times 2 = 4$. Note that other publications applying ECC like e.g. Schefzik et al. (2013) or Schefzik (2017) solely evaluate forecasts of dimension 3. The unified averaged rank histograms of ECC-Q are shown in Figure 4.12 for the reduced grid boxes. The smaller the dimensions and thus the maximal distance between grid centers, the closer the histogram becomes to being uniform. The remaining hump in the middle might be attributed to the higher resolution of the verification set, as mentioned before.

Because the considered grid boxes contain 25 points, the pre-ranks of the multivariate rank histogram cannot be uniquely determined resulting in many ties, which are resolved at random. Hence, the histograms of most models are perfectly uniform and we omit showing them. As an example for the verification results of aggregated quantities, Figure

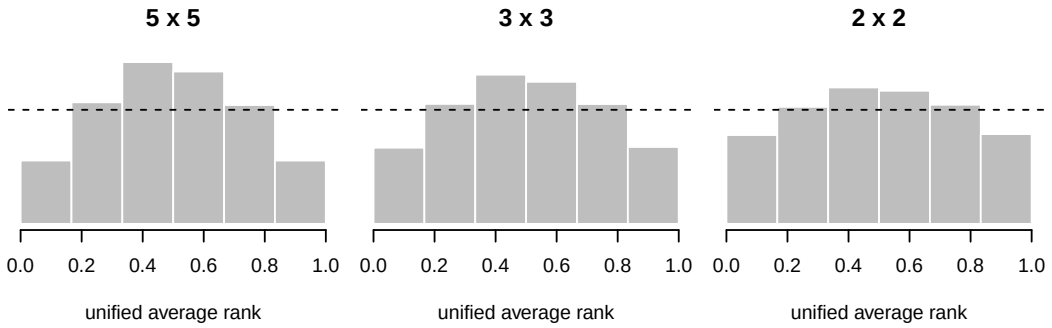

Figure 4.12: Unified average rank histograms for ECC-Q postprocessed TIGGE temperature forecasts initialized at 12 UTC verified against ERA5 across 5×5, 3×3 and 2×2 grid boxes at a prediction horizon of 2 days accumulated over the test period from January 1, 2018 until December 31, 2018.

4.13 pictures the uPIT histograms of maximum temperature. Most predictions suffer from biases and insufficient calibration, while Schaake-R and spatial EMOS with an empirical correlation function yield histograms closest to uniformity.

4.3 Discussion

The case studies in this chapter have confirmed the great benefit of combining predictions to ensemble forecasts as envisioned in the TIGGE project. Already the direct model output delivers skillful forecasts, when verified against reanalysis data; this data is however not independent of the prediction model. Most postprocessing approaches are trained and verified on past observational data sets instead and generate great benefit in forecast skill (see Chapter 6). In the current context, the potential scope of improving predictions through postprocessing is limited because of the high quality of raw TIGGE predictions and the dependence between the verification set and the prediction model. This holds especially true for the more complex spatial postprocessing approaches as the implications and sensitivity of multivariate evaluation tools are not fully understood and still further research is needed. In particular, the verification results for the methods EMOS, the Schaake shuffle and ECC demonstrate a strong dependence on the applied sampling scheme. While the random sample approach results in performance comparable to the other spatial methods, the quantile approach largely outperforms the competing multivariate postprocessing techniques.

In Chapter 3, we proposed a globally applicable spatial postprocessing methods that delivers a multivariate predictive distribution. While the size of the prediction ensemble generated via ECC or the Schaake shuffle is restricted by the size of the original ensemble or number of available training cases, we can create an ensemble of any desired size with this proposed technique which can be beneficial for the end user. Though this spatial

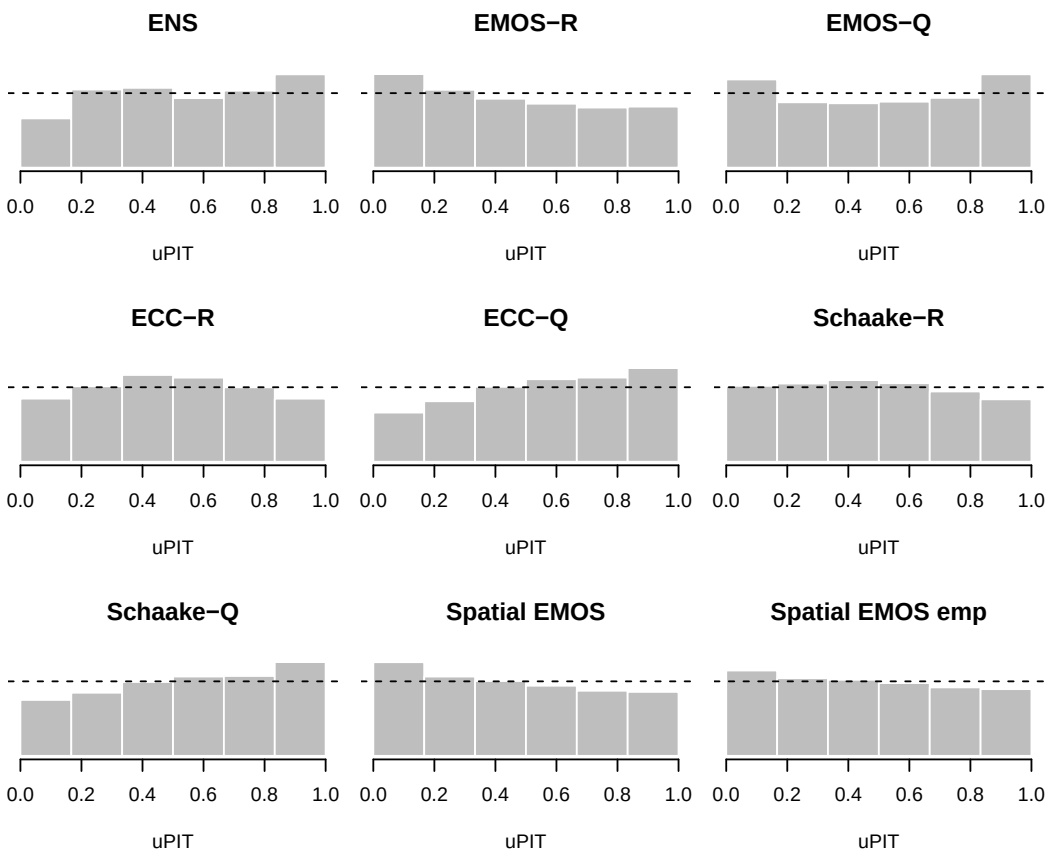

Figure 4.13: uPIT histograms for raw and postprocessed forecasts of maximal temperature across 5×5 grid boxes provided by TIGGE initialized at 12 UTC verified against ERA5 at a prediction horizon of 2 days, accumulated over the test period from January 1, 2018 until December 31, 2018.

EMOS version does not outperform the references standards, ECC-Q and Schaake-Q, in terms of ES and VS (see Tables 4.5 and 4.7), results for spatial calibration are superior to the benchmarks as suggested by the histograms in Figures 4.10 and 4.11.

In the case of spatial EMOS with an empirical correlation function, the sliding window approach yields more skillful forecasts than a fixed window for the training period. Also for ECC and the Schaake shuffle, the spatial dependence template differs daily capturing seasonal changes in the structure. Spatial EMOS with a parametric correlation function might also benefit from a sliding window training period, which we omit due to computational capacity.

The partially superior performance of spatial EMOS with an empirical correlation function might indicate that the pre-made assumptions on the structure on the spatial dependence are too restrictive. Based on covariance tapering and principal component analysis, Heinrich et al. (2021) propose a more flexible multivariate approach for seasonal temperature forecasts which also allows for negative correlations of the forecast errors.

When applied solely to sea surface temperature, this method in certain cases outperforms the Schaake shuffle and might also improve the quality of TIGGE temperature forecasts.

As the resolutions of the contributing ensembles are lower than for the verifying reanalyses (see Table 4.1), small scale variations present in the verification data might not be represented by the ensembles. Thus, the ensemble can overestimate the dependence between grid points, especially seen in the multivariate histograms for ECC. When verifying an EPS with its associated reanalyses, this issue does not occur, as both are produced on the same spatio-temporal scale. However, due to the multi-model nature of TIGGE, ensembles of different resolutions are merged, which can lead to misspecifications of spatial dependence between grid points. In particular for spatial modeling of probabilistic forecasts, we recommend careful selection of the forecasting model and verification data to ensure congruence.

Chapter 5

Postprocessing for ECMWF forecasts

As supported in case studies by for instance Buizza et al. (2005) or Hagedorn et al. (2012), we have demonstrated in the previous chapter that the ECMWF EPS issues the most skillful forecasts among the contributing TIGGE sub-ensembles. Due to its outstanding predictive performance, the ECMWF ensemble has been subject to many research studies – see e.g. Hemri et al. (2014), Schefzik (2017) or Vogel et al. (2020). In this chapter, we focus on exploring the benefits of statistical postprocessing for this EPS individually with again a special emphasis on modeling the spatial dependence of the probabilistic forecasts by applying the methods introduced in Chapter 3.

This chapter begins with further information on the ECMWF data set in Section 5.1. We apply EMOS (see Section 2.2) to global temperature forecasts by this ensemble system. In order to account for spatial dependence of the calibrated temperature forecast fields, we employ ECC, the Schaake shuffle and a number of variants of spatial EMOS (see Sections 3.2.1, 3.2.2 and 3.4, respectively) to 3-days ahead predictions. The verification results are reported in Sections 5.2 and 5.3. These suggest that the assumption of temperature forecasts being normally distributed might not be true for this data and in the spirit of Gebetsberger et al. (2017), we conduct more specialized experiments with different distribution families over the restricted area of Europe. The chapter closes with a discussion of the results.

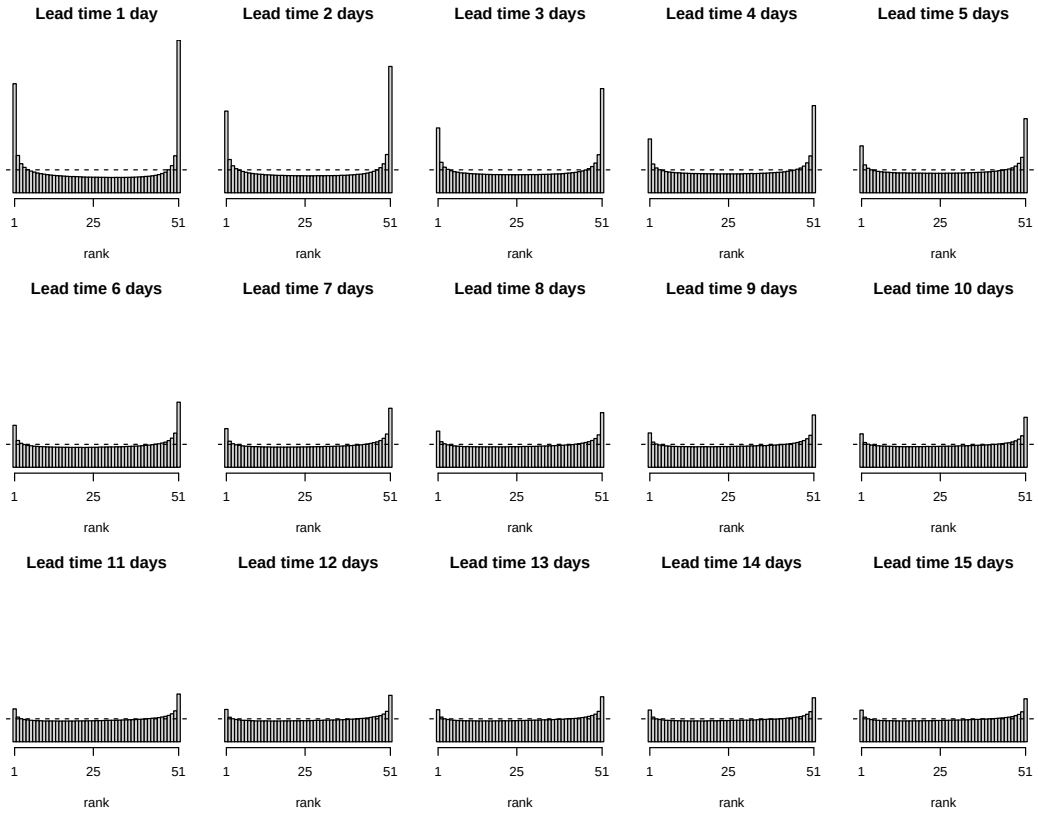

Figure 5.1: Rank histograms for temperature forecasts by the ECMWF ensemble, verified against ERA5, from November 1, 2016 until December 7, 2017 for lead times from 1 to 15 days ahead.

5.1 Data set

We study surface (2m) temperature forecasts issued between March 1, 2016 and December 7, 2017. Specifically, we consider the 50 perturbed members of the ECMWF ensemble. The forecasts are initialized at 12 UTC, valid at lead times from 1 to 15 days, and available globally on a $1.5^\circ \times 1.5^\circ$ grid, for a total of $240 \times 121 = 29,040$ grid points on planet Earth. To train and verify we use its associated reanalyses ERA5 (Hersbach et al., 2020). While generally we consider the entire globe, some more specialized analyses use a restricted data set, where the evaluation region is over Europe (10.5°W-30°E; $36^\circ - 69^\circ$N), resulting in $28 \times 23 = 644$ grid points only. The unit used is degrees Celsius.

Figure 5.1 displays rank histograms of the ECMWF ensemble for consecutive lead times of 1, 2, ..., 15 days. All histograms show a minor bias as they display a skewed shape with many values falling in the upper bins. More importantly, at shorter lead times the histograms are strongly underdispersive, with the ECMWF ensemble underestimating the uncertainty in the forecast due to insufficient spread, as illustrated further in Figure 5.2. At higher lead times, the underdispersion becomes less and less

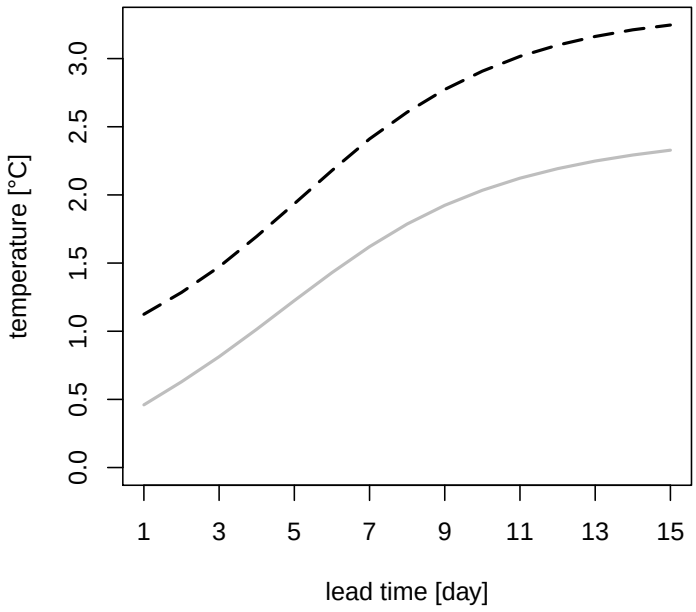

Figure 5.2: Mean ensemble standard deviation (grey) and root mean squared error of the ECMWF ensemble mean (black-dashed) for temperature forecasts from November 1, 2016 until December 7, 2017.

pronounced. Further information about the ECMWF's Integrated Forecasting System and its performance is available in Molteni et al. (1996), Richardson (2000), Buizza et al. (2005) and ECMWF Directorate (2012), among many other documents, and via the ECMWF website at `https://www.ecmwf.int/`.

5.2 Univariate postprocessing

Our evaluation period includes forecasts issued November 1, 2016 through December 7, 2017, amounting to 402 forecasting days and allowing for training periods of a maximal length of 200 days. Excluding the North Pole and South Pole, we consider $240 \times 119 = 28,560$ grid points, resulting in a total of 11,481,120 univariate temperature forecast cases. To this data, we apply the EMOS neighborhood variants presented in Section 2.2.2.

In Figure 5.3, we compare the predictive performance of the postprocessed temperature forecasts and the raw ensemble in terms of the CRPS. At lead times up to about 10 days, postprocessing consistently ameliorates upon the raw ensemble forecasts. For longer lead times, postprocessing fails to yield improvement at training periods of

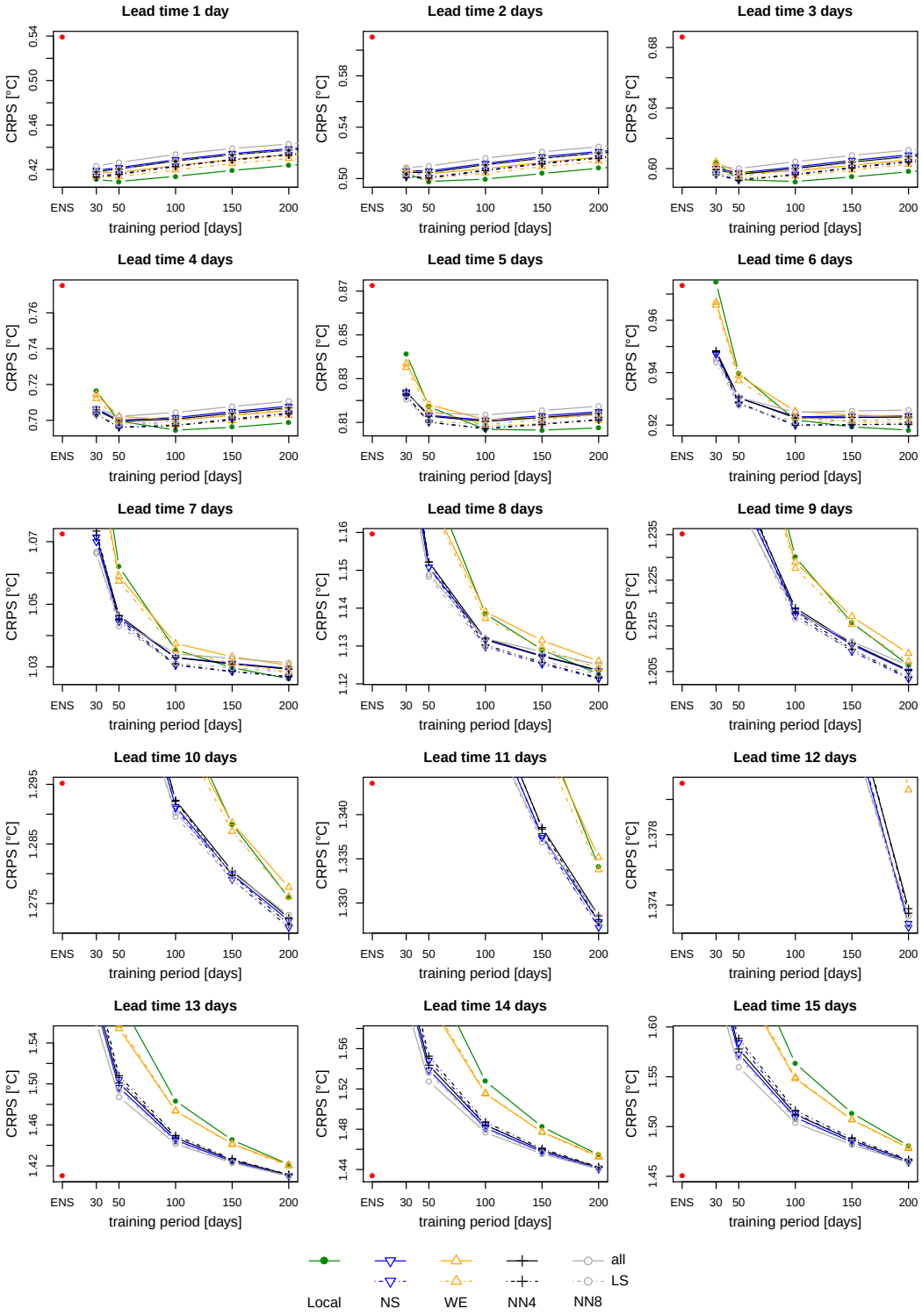

Figure 5.3: Globally averaged CRPS for EMOS postprocessed and raw ECMWF ensemble temperature forecasts over November 1, 2016 through December 7, 2017, verified against ERA5. The scores for the raw ensemble are marked by red dots.

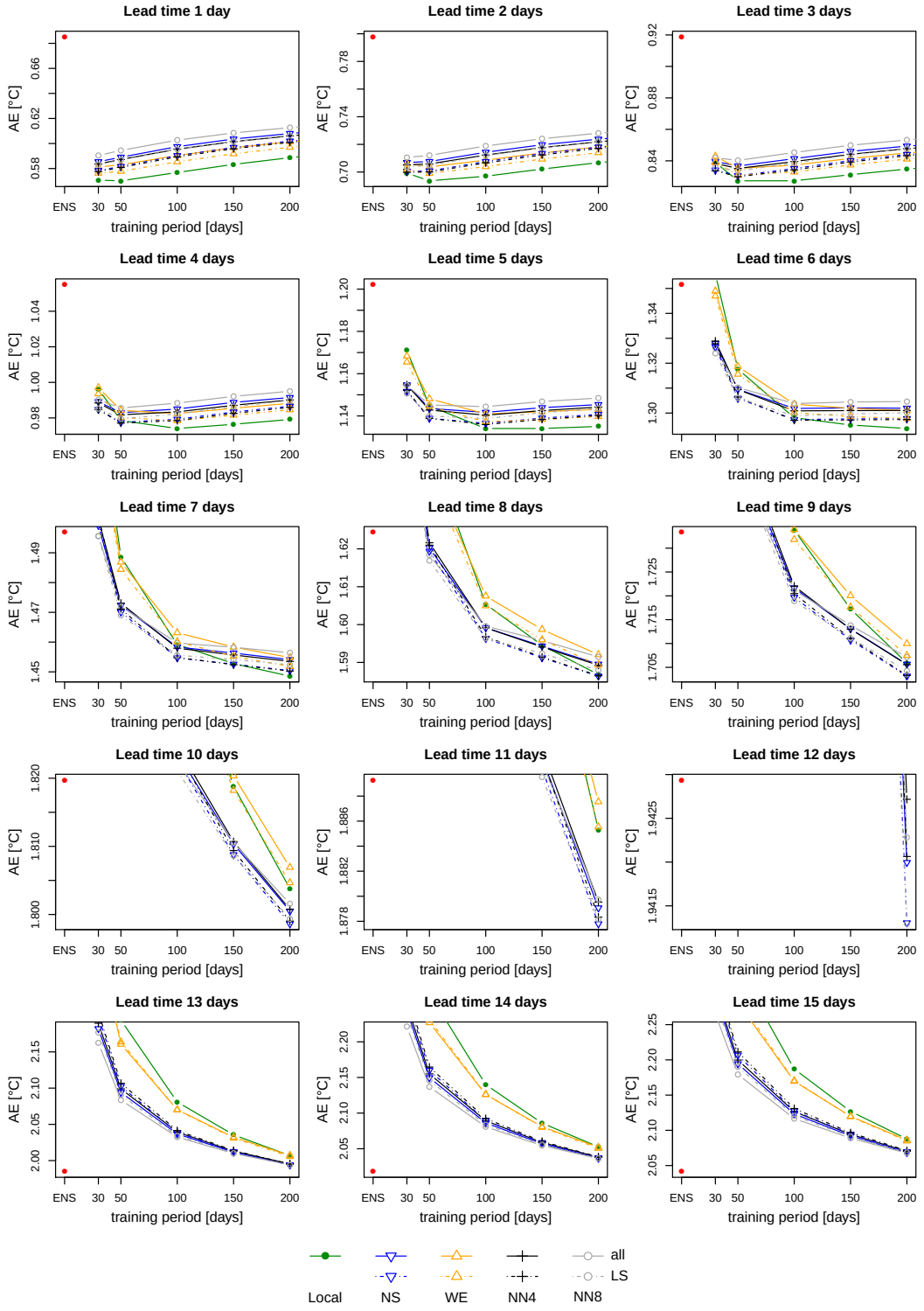

Figure 5.4: Globally averaged AE for EMOS postprocessed and raw ECMWF ensemble temperature forecasts over November 1, 2016 through December 7, 2017, verified against ERA5. The scores for the raw ensemble are marked by red dots.

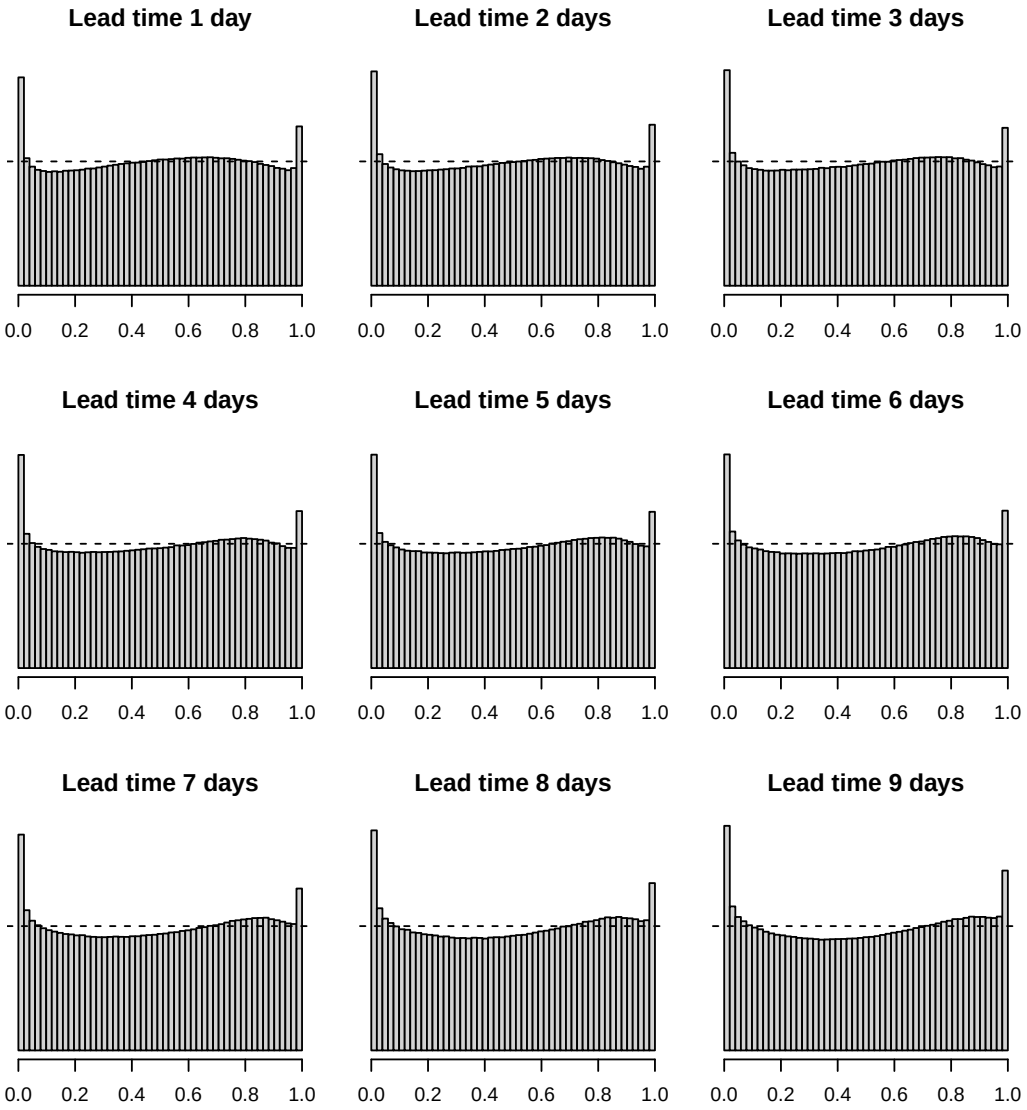

Figure 5.5: PIT histograms for local EMOS postprocessed temperature forecasts by ECMWF with a training period of 100 days, verified against ERA5 from November 1, 2016 until December 7, 2017.

lengths up to 200 days; it may or may not do so if longer training periods or reforecast data sets are used for calibration; see Hagedorn et al. (2008) and Hagedorn et al. (2012). At lead times up to 5 days, the length of the training period tends to be optimal at 50 or 100 days for all estimation methods considered (local, NS, WE, NN4, and NN8, with and without the LS restriction). Local EMOS yields the lowest scores at lead times up to about 5 days. At higher lead times, spatially extended training sets are preferred. The NN8 method eventually performs best among the EMOS techniques, but is outperformed by the raw ensemble forecast at lead times beyond 14 days.

Table 5.1: Globally averaged energy and variogram score for direct model output by ECMWF and postprocessed 3-days ahead forecasts of the temperature fields in 3×3 grid boxes. Models are verified against ERA5 in the grid boxes (containing 9 grid points) from November 1, 2016 through December 7, 2017. The predictive sample is either generated by a random draw (R) or the calculation of equidistant quantiles (Q).

	ES	VS		
Forecast		$p = 0.5$	$p = 1$	$p = 2$
Raw ensemble	2.37	9.65	95.8	21619
EMOS-R	2.12	11.26	98.4	18555
ECC-Q	**2.07**	**8.93**	84.3	18142
Schaake-Q	2.08	9.22	**82.8**	17304
Spatial EMOS emp cor	2.09	9.04	83.8	**17161**

Similar results can be found in Figure 5.4, where the methods are compared in terms of the AE. Local EMOS yields best (lowest) scores at shorter lead times up to several days ahead. At a lead time of 1 day, the mean AE for the raw ensemble is 0.68 degrees Celsius, whereas local EMOS reaches a minimum of less than 0.58 degrees Celsius. At a lead time of 2 days, the AE values for the raw ensemble and local EMOS rise to 0.80 and 0.70 degrees Celsius, respectively. The benefits of postprocessing begin to vanish at lead times of about 10 days, and at lead times of 14 days and more, the AE values for the raw and postprocessed forecasts are in excess of two degrees Celsius.

Figure 5.5 shows globally aggregated PIT histograms for local EMOS postprocessed forecasts with a training period of 100 days at lead times from 1 to 9 days. While the PIT histograms for the EMOS postprocessed forecasts are considerably more uniform than the rank histograms for the raw ensemble forecast (Figure 5.1), they still show underdispersion. This suggests that instead of assuming temperature to be normally distributed, EMOS specifications for probability distributions with heavier tails might be more appropriate for this data. We return to this issue and investigate it in some detail in Section 5.4.

5.3 Spatial postprocessing

After univariate postprocessing, we now turn to EMOS specifications that allow to model and accommodate dependencies in the multivariate postprocessed distributions. In particular, we fit spatial EMOS models, as introduced in Section 3.4, over boxes comprising $3 \times 3 = 9$ grid points, with the center of the grid sliding over the surface of planet Earth, except for the North Pole and South Pole. All scores reported are globally averaged across the respective boxes over space and time. Specifically, we apply the spatial techniques in concert with local EMOS with a rolling training period of length 100 days and at a prediction horizon of 3 days. The test period is the same as before.

Table 5.2: Globally averaged CRPS for direct model output by ECMWF and postprocessed temperature forecasts of minimum (min), maximum (max) and average (ave) temperature across 3×3 grid boxes at a prediction horizon of 3 days. Models are verified against ERA5 in these grid boxes from November 1, 2016 until December 7, 2017.

	CRPS			AE		
Forecast	min	max	ave	min	max	ave
Raw ensemble	0.712	0.594	0.599	0.943	0.794	0.798
EMOS-R	0.677	0.611	0.544	0.906	0.816	0.717
ECC-Q	**0.592**	0.522	**0.514**	**0.829**	0.726	**0.715**
Schaake-Q	0.593	**0.515**	**0.514**	**0.829**	**0.719**	0.719
Spatial EMOS emp cor	0.599	0.524	0.519	0.834	0.728	0.724

Table 5.3: Mean ES and VS for direct model output by ECMWF and postprocessed 3-days ahead forecasts of temperature fields in 3×3 grid boxes over Europe. Models are verified against ERA5 from November 1, 2016 through December 7, 2017.

	ES	VS		
Forecast		$p = 0.5$	$p = 1$	$p = 2$
Raw ensemble	2.531	11.96	109.9	12957
EMOS-R	2.432	12.53	110.6	11757
ECC-Q	**2.212**	**10.40**	**89.4**	9632
Schaake-Q	2.242	10.90	91.2	**9595**
Regional Spatial EMOS	2.247	10.82	92.6	9880
Spatial EMOS	2.240	10.68	91.5	9769
Spatial EMOS emp cor	2.391	11.87	96.4	9826

Table 5.1 compares the raw ensemble forecast and the standard (independent) implementation of the (local) EMOS postprocessed forecast (EMOS) to various explicitly spatial variants thereof, namely, ensemble copula coupling (ECC), the Schaake shuffle (Schaake), and a non-parametric version that relies on empirical correlations (Spatial EMOS emp cor), in terms of globally averaged energy and variogram scores. The methods that account for spatial structure yield lower scores than univariate EMOS and the raw ensemble, but do not differ much between themselves.

Spatial dependence structure is of critical importance when predicting aggregated univariate quantities, such as minima, maxima, totals, or averages over a region. For our boxes of 3×3 grid points we consider the minimum, maximum and average temperature. Then we apply univariate evaluation methods such as the CRPS presented in Table 5.2. Spatial postprocessing improves the raw ensemble forecast and the spatially independent EMOS implementation for the aggregated quantities by an impressive margin, with ECC and the Schaake shuffle yielding the lowest scores overall.

Table 5.4: Mean CRPS for direct model output by ECMWF and postprocessed forecasts of minimum (min), maximum (max) and average (ave) temperature across 3×3 grid boxes in Europe at a prediction horizon of 3 days. Models are verified against ERA5 in the grid boxes points from November 1, 2016 through December 7, 2017.

	CRPS			AE		
Forecast	min	max	av	min	max	av
Raw ensemble	0.774	0.600	0.519	1.028	0.813	0.694
EMOS-R	0.735	0.662	0.538	0.991	0.895	0.681
ECC-Q	**0.610**	0.561	0.451	**0.848**	0.782	**0.624**
Schaake-Q	0.617	**0.549**	**0.450**	0.862	**0.768**	0.627
Regional Spatial EMOS	0.626	0.575	0.458	0.870	0.798	0.632
Spatial EMOS	0.623	0.566	0.455	0.866	0.787	0.634
Spatial EMOS emp cor	0.677	0.601	0.486	0.896	0.806	0.636

5.4 Further analyses over Europe

We close this case study with a number of more specialized experiments that we conduct with a restricted data set, where the evaluation region is over Europe (10.5°W–30°E; 36°–69°N), resulting in $28 \times 23 = 644$ grid points only. The test period is the same as before.

The PIT histograms for the EMOS postprocessed forecasts in Figure 5.5 suggest that the normal assumption may not be ideal. A potential explanation is that the tails of the Gaussian forecast distributions in Eq. (2.2) are too light, and in the following we explore the use of alternatives with heavier tails, such as logistic or Student's t-distributions with degrees of freedom fixed at 3 and 5, as proposed by Gebetsberger et al. (2017). The respective mean CRPS values for forecasts at a prediction horizon of 3 days can be found in Figure 5.6. For training periods up to 150 days, the standard EMOS model with normal distributions, indicated by red, yields the lowest score. For a longer training period of 200 days, the EMOS model with a logistic distribution performs slightly better. As the improvement is negligible, we continue to employ the Gaussian EMOS model.

We turn to spatial postprocessing, where we present a full-fledged comparison, including a further competitor, namely the regional Spatial EMOS technique, where we fit a single Matérn model of the form (3.10) over all of Europe applying the model across the 644 corresponding grid points simultaneously. Otherwise, the setting is the same as in Section 5.3. Table 5.3 shows evaluation results in terms of energy and variogram scores. Again, EMOS paired with ECC or the Schaake shuffle yield the lowest scores. In Table 5.4 we assess forecasts of aggregated quantities. As before, spatial postprocessing improves the raw ensemble or spatially independent EMOS forecasts, and EMOS paired with ECC or the Schaake shuffle yields the best results.

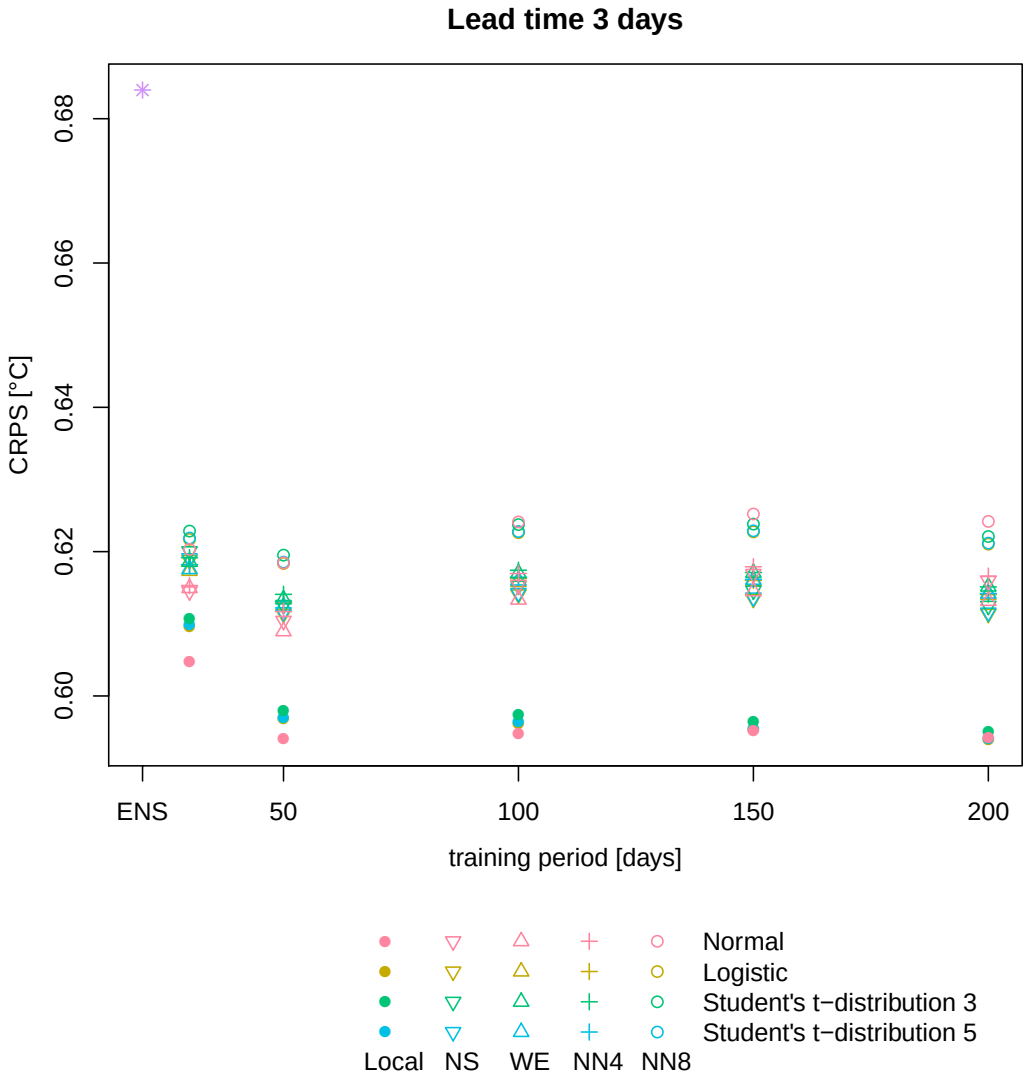

Figure 5.6: CRPS for EMOS postprocessed ECMWF temperature forecasts under normal, logistic, and Student's t-distribution (with degrees of freedom fixed at 3 and 5) assumptions, verified against ERA5, averaged over Europe from November 1, 2016 until December 7, 2017, at a prediction horizon of 3 days. The violet asterisk indicates the result for the raw ensemble.

5.5 Discussion

In this chapter, we have seen how postprocessing can improve temperature forecasts by the ECMWF ensemble system. For longer lead times of about 10 days, the benefit of postprocessing begins to vanish. When applying the spatial methods, we thus choose a shorter prediction horizon and have demonstrated that the skill of the 3-days ahead

forecasts improves through the reintroduction of the spatial dependence (see Tables 3.4 and 5.2).

In most assessments, ECC combined with EMOS yields the lowest scores, followed by the Schaake shuffle. The ECC approach reintroduces the spatial template of the raw ensemble, while the Schaake shuffle retains the spatial structure of past verification data. In this chapter, we validate the ECMWF forecasts against its native reanalyses which are both produced on the same spatio-temporal scale. Thus, distortion which might occur due to the misrepresentation of e.g. small scale effects between the forecasting model and verification data are prevented (in contrast to Chapter 4). Especially when both data sets are based on the same resolution, the ECC approach might add more valuable information to the marginally calibrated forecasts and thus improve the quality slightly more.

For the global case study, we fit spatial EMOS with a Mátern correlation function within the 3×3 grid boxes independently to save computational power. Over Europe, we compared the simultaneous fit over all grid points to the boxes approach. As the parameters of the correlation function differ regionally, the boxes approach outperforms the simultaneous estimation. For univariate EMOS, we experiment with the replacement of the normal distribution by logistic and Student's t-distributions, but find solely slight improvements in the forecast quality and choose to continue employing the normal distribution family.

Chapter 6

Grid- vs. station-based postprocessing of ensemble forecasts

Any application and evaluation of postprocessing methods (described in Chapters 2 and 3) relies on the availability of training and verification data. The choice thereof is of crucial importance, and a fundamental decision is to be made between using gridded data, or station-based data at spatially scattered meteorological stations across planet Earth. So far into this manuscript we have used (re)analyses for this purpose, since they cover the entire globe on the same scale as the forecasting model. In this chapter, we aim to close a gap in the extant literature, by providing a systematic comparison of the effects of this choice for both raw and statistically postprocessed temperature forecasts from the ECMWF ensemble system (see also Section 5.1).

In current practice, the most common approach is to apply ensemble postprocessing techniques to training and verification data at spatially scattered meteorological stations. To this end, observations at surface weather stations are required, and gridded output from NWP models is bi-linearly interpolated to the station locations (e.g. Hemri et al., 2014). Arguably, this commonly used approach is natural, as it aims to perfect forecasts of weather quantities that are directly observed. An alternative lies in the use of a gridded reanalyses with a retrospective best estimate of the state of the atmosphere. This approach has the advantage of the temporal and spatial scale of the training and verification data being consistent with the NWP model output, and satisfies user needs of gridded postprocessed guidance. For a thoughtful, up-to-date discussion of the trade-offs in the choice of training and verification data for postprocessing see Hamill (2018, Section 7.3.2).

In a detailed study of ECMWF ensemble forecasts of surface wind speed, Pinson and Hagedorn (2012) compared the effects of the grid-based and station-based approaches in terms of the forecast quality. Not surprisingly, standard performance measures for probabilistic forecasts, such as the CRPS, indicate a lesser performance of the raw

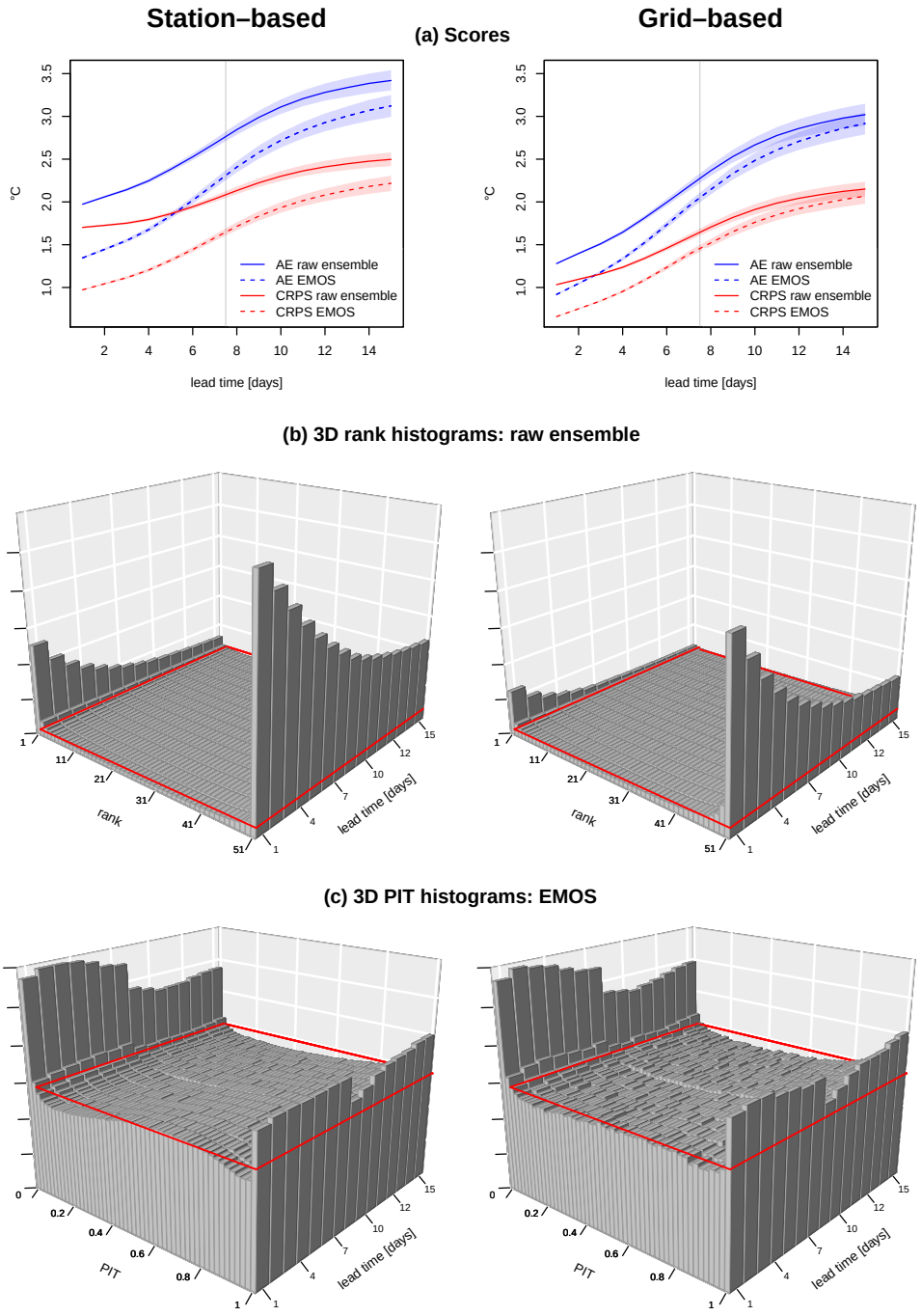

Figure 6.1: Comparison of station-based forecasts of surface temperature, trained and verified against observations at WMO stations, to matched grid-based forecasts, trained and verified against ERA5, in terms of (a) mean CRPS and mean AE, (b) rank histograms for the raw ensemble, and (c) PIT histograms for the EMOS postprocessed forecasts, at lead times from one to 15 days. The thick red lines in the three-dimensional histograms correspond to uniformity.

ensemble forecast in the station-based approach. Furthermore, when evaluated against a gridded analysis, the forecasts tend to be more reliable and better calibrated. The superior performance in the grid-based approach can be attributed to the absence of representativeness error and subgrid variability, in stark contrast to the station-based approach. However, Pinson and Hagedorn (2012) restrict attention to the raw ensemble forecast and do not consider statistical postprocessing.

In this chapter, we analyze the effects of using grid-based vs. station-based training and verification data on both raw and statistically postprocessed ensemble forecasts. For postprocessing, we again adopt the widely used EMOS approach (see Section 2.2). The case study is based on the same gridded data as of Section 5.1 and further expanded by observational data, described in Section 6.1. We apply EMOS to gridded temperature forecasts from the ECMWF 50-member ensemble, with a reanalysis being employed for training and verification, and to bi-linearly interpolated forecasts with station-based observational data used to train and verify. The two approaches are compared in Section 6.2, and discussed in Section 6.3. The chapter closes with an appendix where we desribe of the moving blocks bootstrap procedure which is implemented to generate the error bars of the verification scores in the case study.

6.1 Data set

For our case study, we consider the same surface (2m) temperature forecasts as described in Section 5.1. In the grid-based approach, we work directly on the grid and again use ECMWF's ERA5 reanalysis product for training and verification. In the station-based approach, the ensemble member forecasts are bi-linearly interpolate to the locations of 9,103 World Meteorological Organization (WMO) stations worldwide, and we use station observations for training and verification. Forecasts initialized December 31, 2016 are missing, and we exclude this day from our comparison. Throughout the chapter, the unit used is degrees Celsius.

A major challenge in any comparison of grid- vs. station-based approaches is the distinct coverage of verification data; see, e.g., Buizza (2018). In the grid-based approach, the spatial coverage of the verification data is uniform. In the station-based approach, verification sites correspond to spatially scattered meteorological stations, which cluster in more densely populated and more developed parts of the world. Coverage is sparse over the oceans, in polar regions, and in large parts of Africa.

For a meaningful comparison of grid- vs. station-based postprocessing, we match every single available forecast case at a WMO station to a forecast case at a grid point. Specifically, for any WMO station we consider the four surrounding grid points. The nearest of these four grid points that is of the same surface type (land or sea) as the station is then matched with the station. If none of the surrounding grid points are of the same surface type, we simply match with the nearest grid point, regardless of the land/sea distinction. When multiple stations are matched to the same grid point, the forecast case at this grid point contributes repeatedly. All verification results are

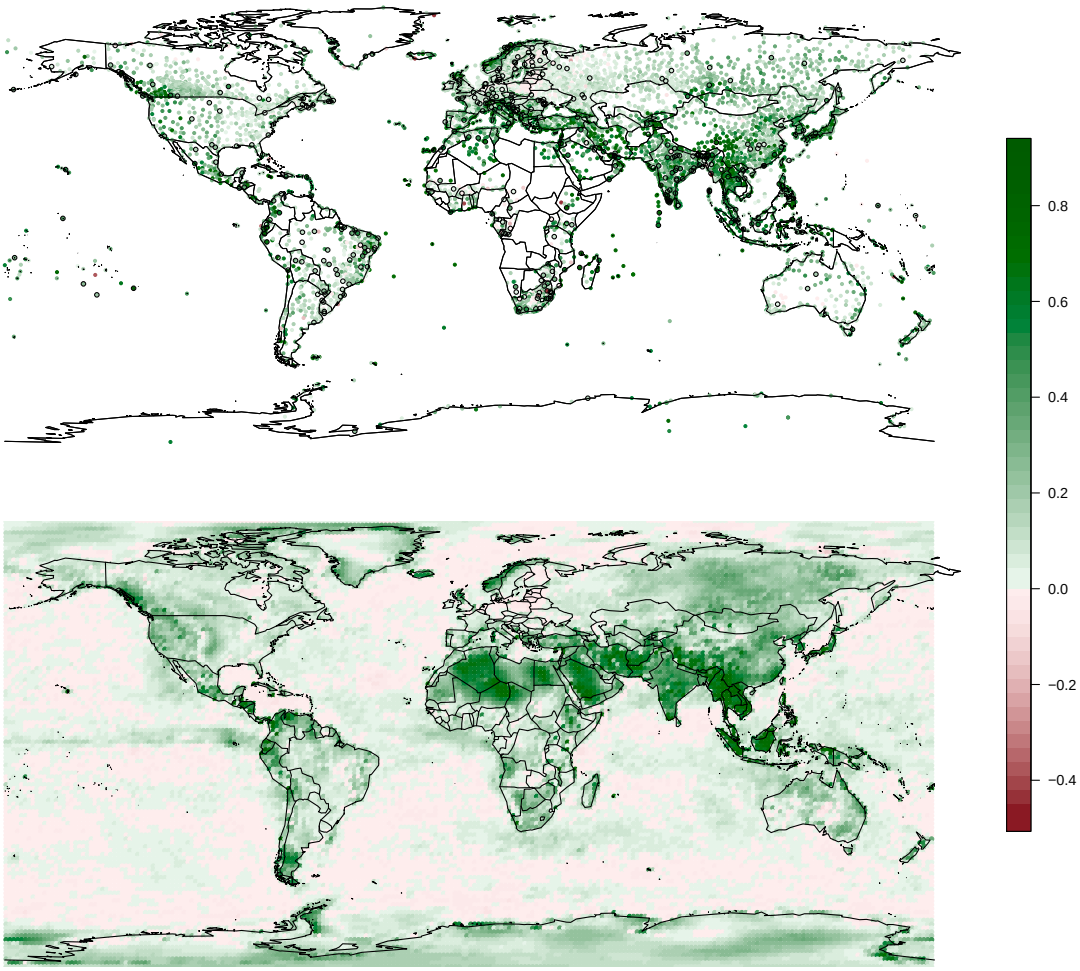

Figure 6.2: CRPS skill for station-based (top panel) and grid-based (bottom panel) EMOS postprocessed forecasts of surface temperature, relative to the raw ECMWF ensemble and at a lead time of three days.

aggregated over the forecast cases from the matched pairs in our evaluation period, which ranges from November 1, 2016 to December 7, 2017.

6.2 Results

Figure 6.1 gives an overall summary of the results from this comparison. The three-dimensional versions of the histograms in (b) and (c) as proposed by Hemri et al. (2015) allow to accommodate all lead times simultaneously. A first, crucial insight from panel (a) is that in both the grid- and station-based approaches statistical postprocessing improves the raw ensemble forecast at all lead times. A day ahead, the raw ensemble

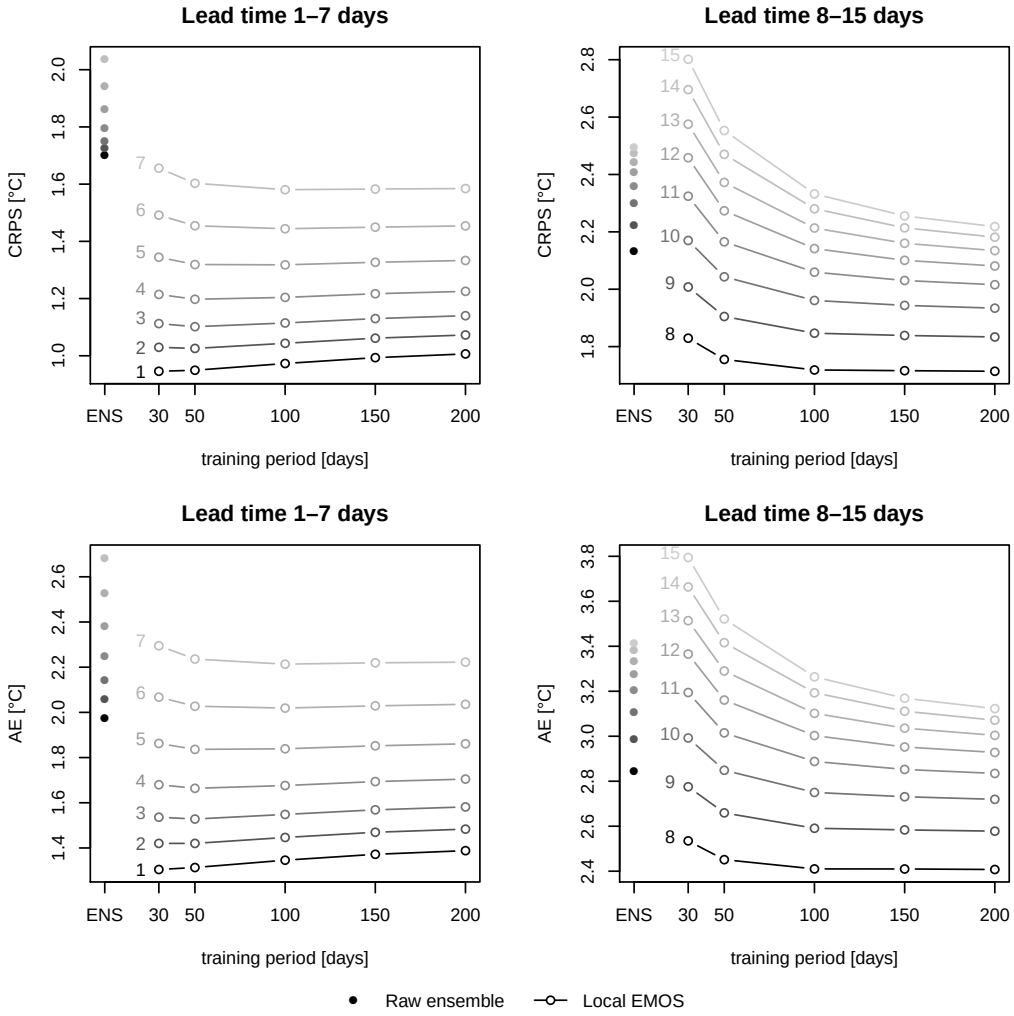

Figure 6.3: Globally averaged CRPS and AE for raw and EMOS postprocessed ECMWF ensemble forecasts of surface temperature in the station-based approach. The scores for the raw ensemble are marked by filled dots, and the numbers from 1 to 15 stand for lead time in days with the corresponding color.

shows mean CRPS and mean AE of 1.03 and 1.28 on the grid, and 1.72 and 1.97 degrees Celsius at stations. Postprocessing with EMOS reduces these numbers to .66 and .92 on the grid, and to .97 and 1.35 at stations, respectively. It is remarkable and indicative of the vast improvement in raw and postprocessed NWP guidance (Alley et al., 2019) that both on the grid and at stations mean CRPS values below a single degree Celsius have been achieved.

At all lead times, raw and postprocessed ensemble forecasts show higher mean CRPS and higher mean AE in the station-based approach, well in line with reports by Pinson and Hagedorn (2012) for the raw ensemble. Not surprisingly, the effects of calibration are more pronounced at stations than on the grid. While postprocessing

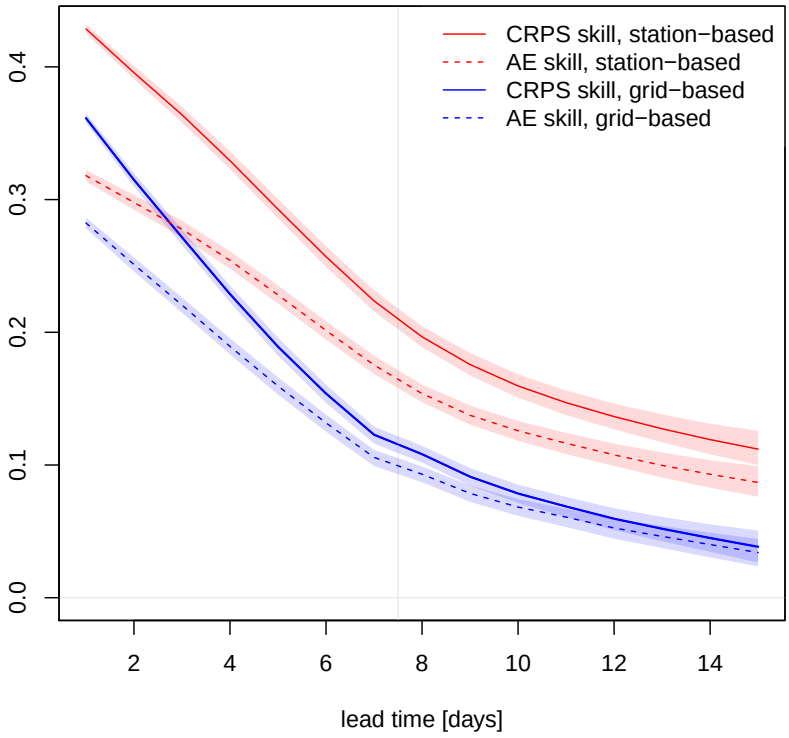

Figure 6.4: CRPS and AE skill score for matched station- and grid-based EMOS postprocessed forecasts of surface temperature relative to the raw ECMWF ensemble.

remains effective beyond week two, its benefits become smaller at larger lead times, and appear to phase out in the grid-based approach, as opposed to the station-based approach, where representativeness error needs to be addressed. The confidence intervals are at the 90% level and have been generated by a moving blocks bootstrap procedure in the spirit of Künsch (1989), Wilks (1997) and Hamill (1999), for which we refer to Section 6.4. Briefly, we pool globally to account for spatial autocorrelation, and for forecasts h days ahead, we use moving blocks of length h days, owing to the stylized fact that verification measures for ideal h-step ahead forecasts are correlated at lags up to $h-1$ time units, but not at larger lags Diebold et al. (1998).

At a lead time of one day postprocessing reduces the mean CRPS from 1.70 to .97 degrees Celsius in the station-based approach, and from 1.03 to .66 in the grid-based approach. This suggests that an improvement of about $1.03 - .66 = .37$ degrees Celsius can be attributed to the reduction of model error, and of roughly $(1.70 - .97) - .37 = .36$ degrees Celsius to the reduction of representativeness error. At a lead time of 15 days, these attributions are .08 and 0.21 degrees Celsius, respectively, for a considerably stronger relative contribution of the latter.

The rank histograms in panel (b) demonstrate that the raw ensemble is underdispersive, particularly at smaller lead times, with the underdispersion being more pronounced in the station-based approach. While the PIT histograms for the EMOS postprocessed

forecasts in panel (c) are much closer to uniformity, some small but notable degree of underdispersion remains. Due to this shape, we explored EMOS with heavier tails – namely logistic or Student-t distributions in Section 5.4. But this results in minimal, if any, improvement and so we retain the Gaussian assumption. The slight structural shift in the three-dimensional PIT histograms between lead times of seven and eight days reflects the increase from 100 to 200 days in the length of the rolling training period that we use for parameter estimation.

For an analysis of any spatial patterns in the predictive performance, Figure 6.2 shows the global distribution of the CRPS skill score for EMOS postprocessed forecasts relative to the raw ensemble, exemplarily at a lead time of three days. The top panel concerns the station-based approach, where the black circles mark the WMO stations with sufficient data to allow for training every single day in the evaluation period. To avoid distortions of the color scale, we exclude the ten stations with the most negative values of the skill score from the visual display. Overall, postprocessing has a thoroughly positive effect, with the skill scores at a vast majority of the stations being positive. The benefits are strongest along the west coast of the Americas and Scandinavia and in tropical and subtropical areas, such as Northern Africa, the Arabian Peninsula, India, South East Asia, and Japan.

The bottom panel in Figure 6.2 turns to the grid-based approach. Generally, the EMOS postprocessed forecasts have positive skill over land, whereas over the oceans skill tends to be slightly negative. The latter may relate to the facts that very few surface observations are available over water, and that spatial variability is much less pronounced over the oceans than over land. On the continents, the patterns in the relative benefits of postprocessing resemble those seen under the station-based approach.

After the comparison of the two verification options, we will close this section with further analysis containing the results of experiments with raw and EMOS postprocessed ECMWF temperature forecasts at observation sites. Similar to Figures 5.3 and 5.4, Figure 6.3 assess the effects of the length of the rolling training period for the estimation of EMOS parameters in the station-based approach. For all lead times the ensemble benefits from postprocessing. For shorter lead times, there is a clear minimum for the length of the training period, whereas forecasts for lead times beyond 7 days might from training periods longer than 200 days.

6.3 Discussion

In a first-ever comprehensive comparison, we have evaluated raw and postprocessed ECMWF ensemble forecasts of surface temperature under both grid-based and station-based approaches. Figure 6.4 summarizes our findings on the effects of EMOS postprocessing in terms of global CRPS and AE skill. The confidence intervals are at the 90% level and have been generated by the moving blocks bootstrap procedure of Appendix 6.4. While under both the station- and grid-based approaches, calibration has

positive effects beyond week two, postprocessing yields less pronounced improvement against gridded (re)analyses, where subgrid variability is not a major concern. Over the continents the benefits of statistical postprocessing are near ubiquitous, with calibration showing positive effects through week two and beyond. These findings extend from the station-based setting studied earlier by Hemri et al. (2014) to the more challenging grid-based approach.

A caveat to these findings is that they are based on pairs of forecast cases that are matched to WMO stations, and so they give emphasis to temperature forecasts in more densely populated and more developed parts of the world. Our choice of either station observations or gridded ERA5 products corresponds to extreme ends within the spectrum options for training and verification data. For variables such as quantitative precipitation, an intermediary alternative lies in the use of gridded, satellite-based observations for training and verification, as exemplified by Vogel et al. (2018).

6.4 Appendix: Moving blocks bootstrap procedure

We now describe the moving blocks bootstrap procedure used to generate the error bars in panel (a) of Figure 6.1 and Figure 6.4. As noted, we pool globally to account for spatial autocorrelation, and for forecasts h days ahead, we use moving blocks of length h days, owing to the stylized fact that verification measures for ideal h-step ahead forecasts are correlated at lags up to $h-1$ time units, but not at larger lags (Diebold et al., 1998).

We present the technique for general types of spatio-temporal verification data observed on successive time units $t = 1, \ldots, T$, at a prediction horizon of h time units ahead, and using a generic bootstrap sample size M. In our work, the evaluation period comprises $T = 401$ days, the prediction horizons considered range from $h = 1$ to $h = 15$ days, and we use $M = 9999$ bootstrap replicates, so that the 500th and 9500th ordered value define a 90% confidence interval. We ignore the complications that arise due to the lack of forecasts initialized on December 31, 2016 and proceed as if the data were observed on $T = 401$ successive days.

Moving blocks bootstrap procedure for a mean score:

(a) At time $t = 1, \ldots, T$ let s_t be the (spatially aggregated) mean score considered, and let n_t be the respective number of spatially scattered observations. Let T_h be the integer part of T/h.

(b) For $m = 1, \ldots, M$ repeat:

- Draw $k_{m,1}, \ldots, k_{m,T_h}$ at random with replacement from $\{1, \ldots, T-h+1\}$.
- Let $n_m = \sum_{t=1}^{T_h} \sum_{j=1}^{h} n_{k_{m,t}+j-1}$.
- Find $S_m = \sum_{t=1}^{T_h} \sum_{j=1}^{h} n_{k_{m,t}+j-1} s_{k_{m,t}+j-1} / n_m$.

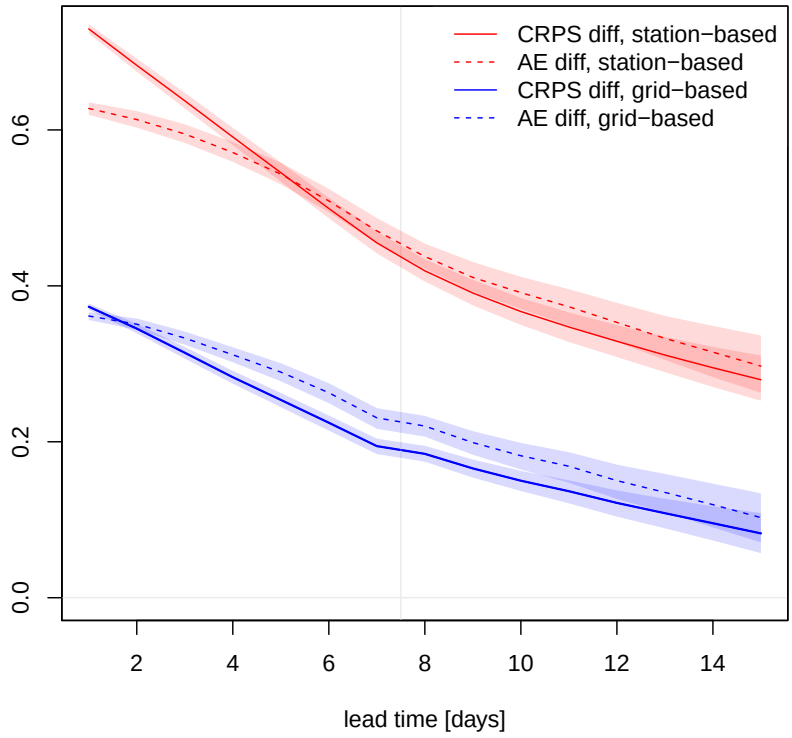

Figure 6.5: Mean score differences in the setting of panel (a) in Figure 6.1. Positive score differences correspond to superior performance of the EMOS postprocessed forecasts relative to the raw ensemble.

(c) Find a confidence interval based on the order statistics of the bootstrap replicates $S_1, ..., S_M$.

With obvious modifications, this scheme yields bootstrap confidence intervals for score differences as well, as exemplified in Figure 6.5. The intervals in this figure do not include a difference of zero, despite the overlap of the respective intervals for the raw ensemble and the EMOS postprocessed forecasts in the grid-based approach in panel (a) of Figure 6.1 at larger lead times. This phenomenon can readily be explained by the joint temporal variation of the mean level of the score for either type of forecast across the evaluation period, which necessitates paired comparisons if tests of the hypothesis of equal predictive performance are desired (Hamill, 1999, Section 3a2).

To obtain confidence intervals for skill scores, slight adaptations are needed, as described now. In our paper the reference method is the raw ensemble forecast, and the method at hand is the EMOS postprocessed forecast.

Moving blocks bootstrap procedure for a skill score:

(a) At time $t = 1, \ldots, T$ let r_t be the (spatially aggregated) mean score under the reference method, let s_t be the (spatially aggregated) mean score under the method

at hand, and let n_t be the respective number of spatially scattered observations. Let T_h be the integer part of T/h.

(b) For $m = 1, \ldots, M$ repeat:

- Draw $k_{m,1}, \ldots, k_{m,T_h}$ at random with replacement from $\{1, \ldots, T - h + 1\}$.
- Let $n_m = \sum_{t=1}^{T_h} \sum_{j=1}^{h} n_{k_{m,t}+j-1}$.
- Find $R_m = \sum_{t=1}^{T_h} \sum_{j=1}^{h} n_{k_{m,t}+j-1} r_{k_{m,t}+j-1}/n_m$ and $S_m = \sum_{t=1}^{T_h} \sum_{j=1}^{h} n_{k_{m,t}+j-1} s_{k_{m,t}+j-1}/n_m$.
- Find the bootstrap replicated skill score α_m based on R_m and S_m.

(c) Find a confidence interval based on the order statistics of the bootstrap replicates $\alpha_1, ..., \alpha_M$.

For a review of resampling based methods in the evaluation of predictive performance under spatial and/or temporal autocorrelation we refer to Sections 5.3.5 and 8.10.5 in Wilks (2011).

Chapter 7

Conclusion

In this thesis, we have seen how statistical postprocessing can systematically improve output from NWP models. Especially for operational use of postprocessing techniques, modeling of spatial, temporal and inter-variable dependence is crucial in order to produce calibrated forecast fields. The main focus of this work was placed on incorporating the spatial dependence in the statistically corrected predictions, while simultaneously exploring the impact on postprocessing when trained and verified against observational or reanalyses data.

True multivariate postprocessing for weather forecasts bears multiple challenges. In Chapter 3, we presented a fully probabilistic postprocessing approach applicable to temperature forecast located on the sphere. Combining EMOS with a GRF model delivers a predictive multivariate distribution which allows for nonstationary and anisotropic correlation of the forecast errors. While the reference standards ECC and the Schaake shuffle base the dependence template on the structure of the forecast ensemble and past verifications, respectively, spatial EMOS benefits from modeling the spatial dependence of the EMOS residuals, which can be viewed as combining the before mentioned approaches.

In the case study in Chapter 4, this global spatial EMOS version yields superior results than the references standards, ECC-Q and Schaake-Q, in terms of spatial calibration, but does not outperform these benchmarks when evaluated by the proper scoring rules ES and VS. This can indicate that the assumptions on the structure of the correlation function might be too restrictive. Other covariance models on spheres as e.g., reviewed in Jeong et al. (2017) could lead to more skillful probabilistic forecast fields.

Up to this date, the sensitivity and implications of multivariate verification measures are not fully understood. While earlier works by Pinson and Girard (2012), Pinson and Tastu (2013) or Scheuerer and Hamill (2015b) suggested that the ES can fail to identify

misspecification in the correlation structure, recent simulation studies by Ziel and Berk (2019) and Lerch et al. (2020) do not support this view and conclude that the VS and ES exhibit similar discrimination ability. We find that these scores can easily detect deficiencies in the raw ensemble, but seem to be sensitive to the sampling scheme of EMOS, ECC and the Schaake shuffle as also suggested by Schefzik (2011, 2017). In general, the ranking of the different multivariate postprocessing models highly depends on the applied proper scoring rule. Thus, these verification measures might benefit from further research to improve mathematical understanding and hence implications for model comparisons.

In the case studies in Chapters 4 and 5, we trained and verified the postprocessing models against reanalysis data. Due to the construction, the analyses are not truly independent of the output of the forecasting model, delivering similar values especially over regions where observational stations are scattered sparsely. Thus the potential benefits through postprocessing diminish, which is especially apparent for TIGGE in Chapter 4. Although Hemri et al. (2014) conclude for station-based data that calibration will improve the direct model output of weather prediction models for the foreseeable future, these highly skillful TIGGE predictions leave limited room for benefits through postprocessing – especially when verified against reanalyses.

This hypothesis is confirmed by the results of the case study in Chapter 6, in which we compare the benefits of postprocessing against observational data versus analysis data. The positive impact on the forecast skill is greater at stations than on a grid. Although postprocessing remains effective for all forecasting horizons up to 15 days, the effect decreases with increasing lead times and appears to diminish for gridded data, in contrast to the observational data.

Although the impressive univariate performance results by TIGGE in Chapter 4 might be partially attributed to the verification against reanalyses, as a multi-model ensemble, TIGGE combines truly independent forecast ensembles from different weather centers. For deterministic forecasts, Ebert (2001) states that merging them from multiple NWP centers produces an EPS with greater skill than the contributing NWP models provide individually. Within the TIGGE project, not only the deterministic forecasts, but the whole EPSs are united to generate the multi-model ensemble. While each of the contributing EPSs might be subject to biases and dispersion errors, the combination of them transcends much of the individual shortcomings.

For the future of postprocessing in weather forecasting, the ultimate aim should be to account for all three types of dependences – spatial, temporal, and inter-variable. Such would enable the production of coherent, calibrated and physically realistic forecast fields that benefit the end user.

Bibliography

Alley, R. B., K. Emanuel and F. Zhang (2019) Advances in weather prediction. *Science*, **363**, 342–344.

Aminyavari, S. and B. Saghafian (2019) Probabilistic streamflow forecast based on spatial post-processing of TIGGE precipitation forecasts. *Stochastic Environmental Research and Risk Assessment*, **33**, 1939–1950.

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

Banerjee, S. (2005) On geodetic distance computations in spatial modeling. *Biometrics*, **61**, 617–625.

Bao, L., T. Gneiting, E. P. Grimit, P. Guttorp and A. E. Raftery (2010) Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Monthly Weather Review*, **138**, 1811–1821.

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. and S. Lerch (2015) Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, **141**, 2289–2299.

Baran, S. and S. Lerch (2016) Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, **27**, 116–130.

Baran, S. and A. Möller (2015) Joint probabilistic forecasting of wind speed and temperature using Bayesian model averaging. *Environmetrics*, **26**, 120–132.

Baran, S. and A. Möller (2017) Bivariate ensemble model output statistics approach for joint forecasting of wind speed and temperature. *Meteorology and Atmospheric Physics*, **129**, 99–112.

Baran, S. and D. Nemoda (2016) Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, **27**, 280–292.

Barnes, C., C. M. Brierley and R. E. Chandler (2019) New approaches to postprocessing of multi-model ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **145**, 3479–3498.

Bellier, J., G. Bontron and I. Zin (2017) Using meteorological analogues for reordering postprocessed precipitation ensembles in hydrological forecasting. *Water Resources Research*, **53**, 10085–10107.

Bellier, J., I. Zin and G. Bontron (2018) Generating coherent ensemble forecasts after hydrological postprocessing: Adaptations of ECC-based methods. *Water Resources Research*, **54**, 5741–5762.

Ben Bouallègue, Z., T. Heppelmann, S. E. Theis and P. Pinson (2016) Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach. *Monthly Weather Review*, **144**, 4737–4750.

Bentzien, S. and P. Friederichs (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., A. E. Raftery and T. Gneiting (2007) Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, **135**, 1386–1402.

Berrocal, V. J., A. E. Raftery and T. Gneiting (2008) Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Annals of Applied Statistics*, **2**, 1170–1193.

Billingsley, P. (2012) *Probability and Measure*. John Wiley & Sons, Anniversary edition.

Bjerknes, V. (1904) Das Problem der Wettervorhersage betrachtet von dem Standpunkt der Mechanik und der Physik. *Meteorologische Zeitschrift*, **21**, 1–7.

Bolin, D., F. Lindgren et al. (2011) Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Annals of Applied Statistics*, **5**, 523–550.

Bougeault, P., Z. Toth, C. Bishop, B. Brown, D. Burridge, D. H. Chen, B. Ebert, M. Fuentes, T. M. Hamill, K. Mylne, J. Nicolau, T. Paccagnella, Y.-Y. Park, D. Parsons, B. Raoult, D. Schuster, P. S. Dias, R. Swinbank, Y. Takeuchi, W. Tennant, L. Wilson and S. Worley (2010) The THORPEX Interactive Grand Global Ensemble. *Bulletin of the American Meteorological Society*, **91**, 1059–1072.

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.

Buizza, R. (2018) Ensemble forecasting and the need for calibration. *In: Statistical Postprocessing of Ensemble Forecasts* (eds. S. Vannitsem, D. S. Wilks and J. Messner), 15–48, Elsevier.

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

Castruccio, S. and J. Guinness (2017) An evolutionary spectrum approach to incorporate large-scale geographical descriptors on global processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **66**, 329–344.

Castruccio, S. and M. L. Stein (2013) Global space-time models for climate ensembles. *The Annals of Applied Statistics*, **7**, 1593–1611.

Chaloulos, G. and J. Lygeros (2007) Effect of wind correlation on aircraft conflict probability. *Journal of Guidance, Control, and Dynamics*, **30**, 1742–1752.

Chmielecki, R. M. and A. E. Raftery (2011) Probabilistic visibility forecasting using Bayesian model averaging. *Monthly Weather Review*, **139**, 1626–1636.

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagalopalan and R. Wilby (2004) The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, **5**, 243–262.

Cressie, N. A. C. (1985) Fitting variogram models by weighted least squares. *Mathematical Geology*, **17**, 563–586.

Dabernig, M., G. J. Mayr, J. W. Messner and A. Zeileis (2017) Spatial ensemble post-processing with standardized anomalies. *Quarterly Journal of the Royal Meteorological Society*, **143**, 909–916.

Das, B. (2000) *Global Covariance Modeling: A Deformation Approach to Anisotropy.* Ph.D. thesis, University of Washington.

Dawid, A. P. (1984) Statistical theory: The prequential approach. *Journal of the Royal Statistical Society Series A*, **147**, 278–292.

Dawid, A. P. and P. Sebastiani (1999) Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, **27**, 65–81.

Dee, D. P., S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, d. P. Bauer, others, D. Dee, S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. Geer, L. Haimberger, S. Healy, H. Hersbach, E. Holm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. McNally, B. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thepaut and F. Vitart (2011) The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137**, 553–597.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**, 1–22.

Diebold, F. X., T. A. Gunther and A. S. Tay (1998) Evaluating density forecasts with applications to financial risk management. *International Economic Review*, **39**, 863–883.

Diebold, F. X. and R. S. Mariano (1995) Comparing predictive accuracy. *Journal of Business and Economic Statistics*, **20**, 134–144.

Doblas-Reyes, F. J., R. Hagedorn and T. Palmer (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting – II. Calibration and combination. *Tellus A: Dynamic Meteorology and Oceanography*, **57**, 234–252.

Ebert, E. E. (2001) Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Monthly Weather Review*, **129**, 2461–2480.

ECMWF Directorate (2012) Describing ECMWF's forecasts and forecasting system. *ECMWF Newsletter*, **133**, 11–13.

Epstein, E. S. (1969) Stochastic dynamic prediction. *Tellus*, **21**, 739–759.

Feldmann, K., D. S. Richardson and T. Gneiting (2019) Grid- versus station-based postprocessing of ensemble temperature forecasts. *Geophysical Research Letters*, **46**, 7744–7751.

Feldmann, K., M. Scheuerer and T. L. Thorarinsdottir (2015) Spatial postprocessing of ensemble forecasts for temperature using non-homogeneous Gaussian regression. *Monthly Weather Review*, **143**, 955–971.

Fraley, C., A. Raftery, T. Gneiting, M. Sloughter and V. Berrocal (2011) Probabilistic weather forecasting in R. *R Journal*, **3**, 55–63.

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

Friederichs, P. and A. Hense (2007) Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review*, **135**, 2365–2378.

Gebetsberger, M., J. W. Messner, G. J. Mayr and A. Zeileis (2017) Fine-tuning non-homogeneous regression for probabilistic precipitation forecasts: Unanimous predictions, heavy tails, and link functions. *Monthly Weather Review*, **145**, 4693–4708.

Gebetsberger, M., J. W. Messner, G. J. Mayr and A. Zeileis (2018) Estimation methods for non-homogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood. *Monthly Weather Review*, **146**, 4323–4338.

Gel, Y., A. E. Raftery and T. Gneiting (2004) Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation method. *Journal of the American Statistical Association*, **99**, 575–583.

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

Gneiting, T. (2013) Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, **19**, 1327–1349.

Gneiting, T., F. Balabdaoui and A. E. Raftery (2007) Probabilistic forecasts, calibration and sharpness. *Royal Statistical Society*, **69**, 243–268.

Gneiting, T. and M. Katzfuss (2014) Probabilistic forecasting. *Annual Review of Statistics and its Application*, **1**, 125–151.

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

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

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

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

Gneiting, T. and R. Ranjan (2013) Combining predictive distributions. *Electronic Journal of Statistics*, **7**, 1747–1782.

Gneiting, T., Z. Sasvári and M. Schlather (2001) Analogies and correspondences between variograms and covariance functions. *Advances in Applied Probability*, **33**, 617–630.

Gneiting, T., L. I. Stanberry, E. P. Grimit, L. Held and N. A. Johnson (2008) Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, **17**, 211–235.

Good, I. J. (1952) Rational decisions. *Journal of the Royal Statistical Society B*, **135**, 4226–4230.

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

Grimmett, G. and D. Stirzaker (2020) *Probability and Random Processes*. Oxford University Press, fourth edition.

Guinness, J. and M. Fuentes (2016) Isotropic covariance functions on spheres: Some properties and modeling considerations. *Journal of Multivariate Analysis*, **143**, 143–152.

Guttorp, P. and Gneiting, T. (2006) Studies in the history of probability and statistics XLIX: On the Matérn correlation family. *Biometrika*, **93**, 989–995.

Hagedorn, R., R. Buizza, T. M. Hamill, M. Leutbecher and T. Palmer (2012) Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **138**, 1814–1827.

Hagedorn, R., F. J. Doblas-Reyes and T. N. Palmer (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus A: Dynamic Meteorology and Oceanography*, **57**, 219–233.

Hagedorn, R., T. M. Hamill and J. S. Whitaker (2008) Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Monthly Weather Review*, **136**, 2608–2619.

Hamill, T., R. Hagedorn and J. Whitaker (2008) Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Monthly Weather Review*, **136**, 2379–2390.

Hamill, T. M. (1999) Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*, **14**, 155–167.

Hamill, T. M. (2001) Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, **129**, 550–560.

Hamill, T. M. (2018) Practical aspects of statistical postprocessing. *In: Statistical Postprocessing of Ensemble Forecasts* (eds. S. Vannitsem, D. S. Wilks and J. Messner), 187–217, Elsevier.

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

Harper, K., L. W. Uccellini, E. Kalnay, K. Carey and L. Morone (2007) 50th anniversary of operational numerical weather prediction. *Bulletin of the American Meteorological Society*, **88**, 639–650.

Heaton, M., M. Katzfuss, C. Berrett and D. Nychka (2014) Constructing valid spatial processes on the sphere using kernel convolutions. *Environmetrics*, **25**, 2–15.

Heinrich, C., K. H. Hellton, A. Lenkoski and T. L. Thorarinsdottir (2021) Multivariate postprocessing methods for high-dimensional seasonal weather forecasts. *Journal of the American Statistical Association*, **116**, 1048–1059.

Hemri, S., D. Lisniak and B. Klein (2015) Multivariate postprocessing techniques for probabilistic hydrological forecasting. *Water Resources Research*, **51**, 7436–7451.

Hemri, S., M. Scheuerer, F. Pappenberger, K. Bogner and T. Haiden (2014) Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, **41**, 9197–9205.

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

Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume and J.-N. Thépaut (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146**, 1999–2049.

Higdon, D. (1998) A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, **5**, 173–190.

Hu, Y., M. J. Schmeits, S. J. Van Andel, J. S. Verkade, M. Xu, D. P. Solomatine and Z. Liang (2016) A stratified sampling approach for improved sampling from a calibrated ensemble forecast distribution. *Journal of Hydometeorology*, **17**, 2405–2417.

Hyndman, R. J. and Y. Khandakar (2008) Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, **26**, 1–22.

Jacobson, J., W. Kleiber, M. Scheuerer and J. Bellier (2020) Beyond univariate calibration: Verifying spatial structure in ensembles of forecast fields. *Nonlinear Processes in Geophysics*, **27**, 411–427.

Jeong, J. and M. Jun (2015a) A class of Matérn-like covariance functions for smooth processes on a sphere. *Spatial Statistics*, **11**, 1–18.

Jeong, J. and M. Jun (2015b) Covariance models on the surface of a sphere: When does it matter? *Stat*, **4**, 167–182.

Jeong, J., M. Jun and M. G. Genton (2017) Spherical process models for global spatial statistics. *Statistical Science*, **32**, 501–513.

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

Jones, R. H. (1963) Stochastic processes on a sphere. *Annals of Mathematical Statistics*, **34**, 213–218.

Jordan, A., F. Krüger and S. Lerch (2019) Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, **90**, 1–37.

Joslyn, S. L. and J. E. LeClerc (2012) Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, **18**, 126–140.

Jun, M. (2011) Non-stationary cross-covariance models for multivariate processes on a globe. *Scandinavian Journal of Statistics*, **38**, 726–747.

Jun, M. (2014) Matérn-based nonstationary cross-covariance models for global processes. *Journal of Multivariate Analysis*, **128**, 134–146.

Jun, M. and M. L. Stein (2008) Nonstationary covariance models for global data. *Annals of Applied Statistics*, **2**, 1271–1289.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S.-K. Yang, J. Hnilo, M. Fiorino and G. Potter (2002) NCEP-DOE AMIP-II reanalysis (R-2). *Bulletin of the American Meteorological Society*, **83**, 1631–1644.

Kleiber, W. and D. Nychka (2012) Nonstationary modeling for multivariate spatial processes. *Journal of Multivariate Analysis*, **112**, 76–91.

Kruskal, J. B. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, **7**, 48–50.

Künsch, H. R. (1989) The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, **17**, 1217–1241.

Laio, F. and S. Tamea (2007) Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, **11**, 1267–1277.

Lang, M. N., G. J. Mayr, R. Stauffer and A. Zeileis (2019) Bivariate gaussian models for wind vectors in a distributional regression framework. *Advances in Statistical Climatology, Meteorology and Oceanography*, **5**, 115–132.

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

Lerch, S. and S. Baran (2017) Similarity-based semilocal estimation of post-processing models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **66**, 29–51.

Lerch, S., S. Baran, A. Möller, J. Groß, R. Schefzik, S. Hemri and M. Graeter (2020) Simulation-based comparison of multivariate ensemble post-processing methods. *Nonlinear Processes in Geophysics*, **27**, 349–371.

Lerch, S. and T. L. Thorarinsdottir (2013) Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A: Dynamic Meteorology and Oceanography*, **65**, 21206.

Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo and T. Gneiting (2017) Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, **32**, 106–127.

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

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

Lorenz, E. N. (1963) Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, **20**, 130–141.

Lynch, P. (2006) *The emergence of numerical weather prediction: Richardson's dream*. Cambridge University Press.

Lynch, P. (2008) The ENIAC forecasts: a re-creation. *Bulletin of the American Meteorological Society*, **89**, 45–56.

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

Matheson, J. E. and R. L. Winkler (1976) Scoring rules for continuous probability distributions. *Management Science*, **22**, 1087–1096.

Messner, J. W., G. J. Mayr, D. S. Wilks and A. Zeileis (2014) Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Monthly Weather Review*, **142**, 3003–3014.

Möller, A., A. Lenkoski and T. L. Thorarinsdottir (2013) Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, **139**, 982–991.

Molteni, F., R. Buizza, T. N. Palmer and T. Petroliagis (1996) The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, **122**, 73–119.

Murphy, A. H. and R. L. Winkler (1987) A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330–1338.

Nelsen, R. B. (2007) *An Introduction to Copulas*. Springer.

Paciorek, C. J. (2003) *Nonstationary Gaussian processes for regression and spatial modelling*. Ph.D. thesis, Carnegie Mellon University.

Park, Y.-Y., R. Buizza and M. Leutbecher (2008) TIGGE: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society*, **134**, 2029–2050.

Pinson, P. (2013) Wind energy: Forecasting challenges for its operational management. *Statistical Science*, **28**, 564–585.

Pinson, P. and R. Girard (2012) Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, **96**, 12–20.

Pinson, P. and R. Hagedorn (2012) Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. *Meteorological Applications*, **19**, 484–500.

Pinson, P., H. Madsen, H. A. Nielsen, G. Papaefthymiou and B. Klöckl (2009) From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy*, **12**, 51–62.

Pinson, P. and J. W. Messner (2018) Application of postprocessing for renewable energy. *Statistical Postprocessing of Ensemble Forecasts* (eds. S. Vannitsem, D. S. Wilks and J. Messner), 241–266, Elsevier.

Pinson, P. and J. Tastu (2013) Discrimination ability of the energy score. Technical report, URL `http://orbit.dtu.dk/fedora/objects/orbit:122326/datastreams/file_b919613a-9043-4240-bb6c-160c88270881/content`.

Porcu, E., J. Mateu and G. Christakos (2009) Quasi-arithmetic means of covariance functions with potential applications to space-time data. *Journal of Multivariate Analysis*, **100**, 1830–1844.

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

Richardson, D. S. (2000) Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **126**, 649–667.

Richardson, L. (1922) *Weather Prediction by Numerical Processes*. Cambridge University Press.

Roquelaure, S. and T. Bergot (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.

Roulston, M. S. and L. A. Smith (2002) Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, **130**, 1653–1660.

Sampson, P. D. and P. Guttorp (1992) Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, **87**, 108–119.

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

Schefzik, R. (2015) Multivariate discrete copulas, with applications in probabilistic weather forecasting. URL `https://arxiv.org/abs/1512.05629`.

Schefzik, R. (2016) A similarity-based implementation of the Schaake shuffle. *Monthly Weather Review*, **144**, 1909–1921.

Schefzik, R. (2017) Ensemble calibration with preserved correlations: Unifying and comparing ensemble copula coupling and member-by-member postprocessing. *Quarterly Journal of the Royal Meteorological Society*, **143**, 999–1008.

Schefzik, R. and A. Möller (2018) Ensemble postprocessing methods incorporating dependence structures. *In: Statistical Postprocessing of Ensemble Forecasts* (eds. S. Vannitsem, D. S. Wilks and J. Messner), 91–125, Elsevier.

Schefzik, R., T. L. Thorarinsdottir and T. Gneiting (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 L. Büermann (2014) Spatially adaptive post-processing of ensemble forecasts for temperature. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63**, 405–422.

Scheuerer, M. and T. M. Hamill (2015a) Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, **143**, 4578–4596.

Scheuerer, M. and T. M. Hamill (2015b) Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, **143**, 1321–1334.

Scheuerer, M., T. M. Hamill, B. Whitin, M. He and A. Henkel (2017) A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resources Research*, **53**, 3029–3046.

Scheuerer, M. and G. König (2014) Gridded, locally calibrated, probabilistic temperature forecasts based on ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, **140**, 2582–2590.

Scheuerer, M. and D. Möller (2015) Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Annals of Applied Statistics*, **9**, 1328–1349.

Schmeits, M. J. and K. J. Kok (2010) A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Monthly Weather Review*, **138**, 4199–4211.

Schmidt, A. M. and P. Guttorp (2020) Flexible spatial covariance functions. *Spatial Statistics*, 100416.

Schuhen, N., T. L. Thorarinsdottir and T. Gneiting (2012) Ensemble model output statistics for wind vectors. *Monthly Weather Review*, **140**, 3204–3219.

Simmons, A., S. Uppala, D. Dee and S. Kobayashi (2007) ERA-Interim: New ECMWF reanalysis products from 1989 onwards. *ECMWF Newsletter*, **110**, 25–35.

Sklar, M. (1959) Fonctions de répartition à dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, **8**, 229–231.

Sloughter, J. M., T. Gneiting and A. E. Raftery (2010) Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, **105**, 25–35.

Sloughter, J. M. L., T. Gneiting and A. E. Raftery (2013) Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Monthly Weather Review*, **141**, 2107–2119.

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

Smith, L. A. (2001) Disentangling uncertainty and error: On the predictability of nonlinear systems. *In: Nonlinear Dynamics and Statistics* (ed. A. I. Mees), 31–64, Boston, MA: Birkhäuser Boston.

Smith, L. A. and J. A. Hansen (2004) Extending the limits of ensemble forecast verification with the minimum spanning tree histogram. *Monthly Weather Review*, **132**, 1522–1528.

Statistisches Bundesamt (2021) Bruttostromerzeugung in Deutschland. URL `http://www.destatis.de/DE/Themen/Branchen-Unternehmen/Energie/Erzeugung/Tabellen/bruttostromerzeugung.html`.

Stein, M. L. (2005) Nonstationary spatial covariance functions. *Unpublished technical report*, URL `http://www-personal.umich.edu/~jizhu/jizhu/covar/Stein-Summary.pdf`.

Stein, M. L. (2007) Spatial variation of total column ozone on a global scale. *Annals of Applied Statistics*, **1**, 191–210.

Strähl, C. and J. Ziegel (2017) Cross-calibration of probabilistic forecasts. *Electronic Journal of Statistics*, **11**, 608–639.

Swinbank, R., M. Kyouda, P. Buchanan, L. Froude, T. M. Hamill, T. D. Hewson, J. H. Keller, M. Matsueda, J. Methven, F. Pappenberger, M. Scheuerer, H. A. Titley, L. Wilson and M. Yamaguchi (2016) The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, **97**, 49–67.

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

Tao, Y., Q. Duan, A. Ye, W. Gong, Z. Di, M. Xiao and K. Hsu (2014) An evaluation of post-processed TIGGE multimodel ensemble precipitation forecast in the Huai river basin. *Journal of Hydrology*, **519**, 2890–2905.

Thorarinsdottir, T. L. and T. Gneiting (2010) Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **173**, 371–388.

Thorarinsdottir, T. L., M. Scheuerer and C. Heinz (2016) Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics*, **25**, 105–122.

Thorarinsdottir, T. L. and N. Schuhen (2018) Verification: Assessment of calibration and accuracy. *In: Statistical Postprocessing of Ensemble Forecasts* (eds. S. Vannitsem, D. S. Wilks and J. Messner), 155–186, Elsevier.

Tracton, M. S. and E. Kalnay (1993) Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Weather and Forecasting*, **8**, 379–398.

Tsyplakov, A. (2013) Evaluation of probabilistic forecasts: Proper scoring rules and moments. *Available at SSRN 2236605*.

Vannitsem, S., J. B. Bremnes, J. Demaeyer, G. R. Evans, J. Flowerdew, S. Hemri, S. Lerch, N. Roberts, S. Theis, A. Atencia, Z. B. Bouallègue, J. Bhend, M. Dabernig, L. De Cruz, L. Hieta, O. Mestre, L. Moret, I. Odak Plenković, M. Schmeits, M. Taillardat, J. Van den Bergh, B. Van Schaeybroeck, K. Whan and J. Ylhaisi (2021) Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, **102**, E681 – E699.

Vardi, Y. and C.-H. Zhang (2000) The multivariate L1-median and associated data depth. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 1423–1426.

Vogel, P., P. Knippertz, A. H. Fink, A. Schlueter and T. Gneiting (2018) Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Weather and Forecasting*, **33**, 369–388.

Vogel, P., P. Knippertz, A. H. Fink, A. Schlueter and T. Gneiting (2020) Skill of global raw and postprocessed ensemble predictions of rainfall in the tropics. *Weather and Forecasting*, **35**, 2367–2385.

Wilks, D. S. (1997) Resampling hypothesis tests for autocorrelated fields. *Journal of Climate*, **10**, 63–82.

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. (2006) Comparison of ensemble-MOS methods in the Lorenz'96 setting. *Meteorological Applications*, **13**, 243–256.

Wilks, D. S. (2011) *Statistical Methods in the Atmospheric Sciences*. Academic Press, third edition.

Wilks, D. S. (2015) Multivariate ensemble Model Output Statistics using empirical copulas. *Quarterly Jounal of the Royal Meteorological Society*, **141**, 945–952.

Wilks, D. S. (2017) On assessing calibration of multivariate ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **143**, 164–172.

Wilks, D. S. (2018) Univariate ensemble postprocessing. *In: Statistical Postprocessing of Ensemble Forecasts* (eds. S. Vannitsem, D. S. Wilks and J. Messner), 49–89, Elsevier.

Yadrenko, M. I. (1983) *Spectral Theory of Random Fields*. Optimization Software.

Yaglom, A. M. (1987) *Correlation Theory of Stationary and Related Random Functions.* Volume I, Springer-Verlag.

Yuen, R., S. Baran, C. Fraley, T. Gneiting, S. Lerch, M. Scheuerer and T. Thorarinsdottir (2018) *ensembleMOS: Ensemble Model Output Statistics*. R package version 0.8.2.

Ziegel, J. and T. Gneiting (2014) Copula calibration. *Electronic Journal of Statistics*, **8**, 2619–2638.

Ziel, F. and K. Berk (2019) Multivariate forecasting evaluation: On sensitive and strictly proper scoring rules. URL `https://arxiv.org/abs/1910.07325`.