Karlsruhe Institute of Technology

Probabilistic forecasting and comparative model assessment, with focus on extreme events

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von
Dipl.-Math. Sebastian Lerch
aus
Gießen

Tag der mündlichen Prüfung: 18. Mai 2016
Referent: Prof. Dr. Tilmann Gneiting
1. Korreferentin: Dr. Thordis L. Thorarinsdottir
2. Korreferentin: Prof. Dr. Vicky Fasen-Hartmann

Abstract

Probabilistic forecasts in the form of probability distributions over future quantities or events allow to quantify the prediction uncertainty and are essential for informed decision making. The main focus of the work presented in this thesis are topics in making and evaluating probabilistic forecasts.

First, we focus on forecast verification and investigate how to evaluate forecasts with an emphasis on extreme events. In public discussions of forecast quality, attention typically focuses on the predictive performance in cases of extreme events. However, the restriction of conventional forecast evaluation methods to subsets of extreme observations has undesired effects and is bound to discredit skillful forecasts. Any procedure of hand-picking extreme events when comparing and ranking competing forecasters is incompatible with the theoretical assumptions of established forecast verification methods, thus confronting forecasters with what we refer to as the *forecaster's dilemma*. Using theoretical arguments, simulation experiments, and a real data study on probabilistic forecasts of U.S. inflation and gross domestic product growth, we illustrate and discuss the forecaster's dilemma along with potential remedies.

In Bayesian implementations of forecasting models, the forecast distribution of interest is often only available indirectly through a simulated sample, typically generated via Markov chain Monte Carlo algorithms. In the second part of this thesis, we conduct a systematic analysis of how to make and evaluate probabilistic forecast distributions based on such simulation output. Building on the mathematical framework of forecast evaluation, we propose a notion of consistency for assessing the adequacy of methods for estimating the unknown forecast distribution. We then review asymptotic results and derive conditions under which choices from the extant literature satisfy this notion of consistency. The theoretical considerations are illustrated in simulation and case studies in order to assess the efficiency of various approximation methods in practical applications.

The third part focuses on applications of probabilistic forecasting in numerical weather prediction where non-homogeneous regression approaches are used to statistically postprocess forecast ensembles obtained as output from multiple runs of numerical weather prediction models. For wind speed, the standard regression model is given by a truncated normal distribution with parameters derived from the ensemble. We propose alternative models based on log-normal and generalized extreme value distributions, as well as combinations and mixtures thereof. In three case studies for different ensemble prediction systems and wind quantities, the novel models show improved predictive performance, particularly for high wind speed observations. Further, we investigate new similarity-based approaches to parameter estimation for postprocessing models where training data for a specific observation station are augmented with corresponding forecast cases from stations with similar characteristics. In a case study over Europe, the proposed similarity-based semi-local models show improved predictive performance compared to standard estimation methods and allow for efficiently estimating complex models without numerical stability issues.

Zusammenfassung

Probabilistische Vorhersagen in der Form von Wahrscheinlichkeitsverteilungen erlauben eine Quantifizierung der Unsicherheit der Vorhersage und sind damit von essentieller Bedeutung für Entscheidungsprozesse. Das Hauptaugenmerk der vorliegenden Arbeit liegt auf der Erstellung und Bewertung solcher probabilistischen Vorhersagen.

Zunächst wenden wir uns der Verifikation von Vorhersagen zu, insbesondere untersuchen wir geeignete Verfahren zur Bewertung probabilistischer Vorhersagen für Extremereignisse. Diskussionen von Vorhersagequalität in den Medien beschränken sich meist auf die Bewertung von Vorhersagen für ausgewählte Extremereignisse. Es kann jedoch gezeigt werden, dass jede solche Einschränkung der Verifikation auf Teilmengen der Beobachtungen unerwartete und unerwünschte Effekte hat und optimale Vorhersagen benachteiligt. Unter Verwendung von theoretischen Argumenten, Simulationsexperimenten und einer ökonomischen Fallstudie illustrieren und untersuchen wir das resultierende Dilemma und mögliche Auswege.

In Bayesschen Vorhersagemodellen ist die Vorhersageverteilung meist nur indirekt durch eine simulierte Stichprobe zugänglich. Im zweiten Teil der vorliegenden Dissertation wenden wir uns der Erstellung und Bewertung probabilistischer Vorhersagen basierend auf solchen simulierten Stichproben zu. Aufbauend auf dem theoretischen Rahmen etablierter Methoden zur Bewertung probabilistischer Vorhersagen führen wir einen Konsistenzbegriff ein, welcher die Analyse verschiedener Approximationsmethoden zur Schätzung der Vorhersageverteilung erlaubt. Mithilfe asymptotischer Resultate leiten wir Bedingungen her, unter welchen in der Literatur verwendete Approximationsmethoden den eingeführten Konsistenzbegriff erfüllen. Diese mathematischen Betrachtungen werden von Anwendungen in Simulationsexperimenten und Fallstudien begleitet, um eine Untersuchung der Effizienz in praktischen Beispielen zu ermöglichen.

Der dritte Teil der Arbeit beschäftigt sich mit Anwendungen in der numerischen Wettervorhersage. Dort werden heteroskedastische Regressionsmodelle zur statistischen Nachbearbeitung von Ensemblevorhersagen verwendet, welche man aus mehreren Durchläufen numerischer Wettermodelle erhält. Das standardmäßig verwendete Regressionsmodell für Windgeschwindigkeit basiert auf trunkierten Normalverteilungen. Wir untersuchen zunächst alternative parametrische Modelle basierend auf Log-Normal und Extremwertverteilungen sowie geeigneten Kombinationen und Mischungen. In Fallstudien für verschiedene Ensembles zeigen diese neuen Modelle Verbesserungen in der Vorhersagequalität, insbesondere für hohe Windgeschwindigkeiten. Des Weiteren untersuchen wir neue ähnlichkeitsbasierte Ansätze zur Parameterschätzung für Vorhersagemodelle, bei welchen die Trainingsdaten für spezifische Beobachtungsstationen durch Daten von Stationen mit ähnlichen Eigenschaften ergänzt werden. In einer Fallstudie zeigen diese neuen Ansätze Verbesserungen im Vergleich zu Standardmethoden und erlauben die effiziente Schätzung komplexer Modelle ohne numerische Probleme.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Tilmann Gneiting and Thordis Thorarinsdottir for their support and guidance over the last years. I have greatly enjoyed pursuing a Ph.D. under their excellent supervision and have learned a lot from them. I further thank Vicky Fasen for agreeing to serve as a reviewer on my Ph.D. committee, and for helpful discussions prior to the final submission of this thesis.

Financial support by the Volkswagen Foundation through the project "Mesoscale Weather Extremes – Theory, Spatial Modeling and Prediction (WEX-MOP)", by the Deutsche Forschungsgemeinschaft through Research Training Group (RTG) 1953 "Statistical Modeling of Complex Systems and Processes", and through the project "C7 – Statistical Postprocessing and Stochastic Physics for Ensemble Predictions" within Transregional Collaborative Research Center 165 "Waves to Weather" is gratefully acknowledged.

I have benefited from the help and advice of many colleagues over the course of this Ph.D. project. In particular, I would like to thank Sándor Baran, Werner Ehm, Kira Feldmann, Nadine Gissibl, Stephan Hemri, Alexander Jordan, Fabian Krüger, Evgeni Ovcharov, Francesco Ravazzolo, Roman Schefzik and Michael Scheuerer for countless helpful discussions, insightful comments, and for sharing code, data, ideas and experiences. The comments of various reviewers, editors and participants of conferences and workshops have helped improve the quality of the presented research.

I further thank the members of the Institute of Applied Mathematics at Heidelberg University, the Computational Statistics group at the Heidelberg Institute for Theoretical Studies and the members of the Institute of Stochastics at the Karlsruhe Institute of Technology for sharing their knowledge and for creating an enjoyable work environment.

Parts of the work presented in this thesis were made during a research stay at the Norwegian Computing Center in Oslo in the summer of 2014. I thank Alex Lenkoski, Thordis Thorrarinsdottir, and the colleagues at the Norwegian Computing Center for their hospitality.

On a personal note, I would like to extend my gratitude to Sarah, my friends, and my family, and thank them for their support.

List of abbreviations

AE	Absolute Error
ALADIN-HUNEPS	Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System
AR	Autoregressive (models)
BMA	Bayesian Model Averaging
BS	Brier Score
CDF	Cumulative Distribution Function
CL	Conditional Likelihood (scoring rule)
CRPS	Continuous Ranked Probability Score
CSL	Censored Likelihood (scoring rule)
DSS	Dawid-Sebastiani Score
ECDF	Empirical Cumulative Distribution Function
ECMWF	European Centre for Medium-Range Weather Forecasts
EMOS	Ensemble Model Output Statistics
EPS	Ensemble Prediction System
ESS	Error-Spread Score
FAR	False Alarm Rate
GDP	Gross Domestic Product
GEV	Generalized Extreme Value (distribution)
GLAMEPS	Grand Limited Area Model Ensemble Prediction System
HR	Hit Rate
HS	Hyvärinen Score
IG	Inverse Gamma (distribution)
KDE	Kernel Density Estimation
KL	Kullback-Leibler (divergence)
LN	Log-Normal (distribution)
LogS	Logarithmic Score
LRT	Likelihood Ratio Test
MAE	Mean Absolute Error
MCMC	Markov chain Monte Carlo
ML	Maximum Likelihood
NWP	Numerical Weather Prediction
PDF	Probability Density Function
PIT	Probability Integral Transform
QS	Quadratic Score
rCRPS	Restricted Continuous Ranked Probability Score
rDSS	Restricted Dawid-Sebastiani Score
rLogS	Restricted Logarithmic Score

rMAE	Restricted Mean Absolute Error
rMSE	Restricted Mean Squared Error
SE	Squared Error
SEDI	Symmetric Extremal Dependence Index
TN	Truncated Normal (distribution)
TVP-SV	Time-Varying Parameters and Stochastic Volatility specification for AR/VAR models
twCRPS	threshold-weighted Continuous Ranked Probability Score
twCRPSS	threshold-weighted Continuous Ranked Probability Skill Score
U.S.	United States
UWME	University of Washington Mesoscale Ensemble
VAR	Vector Autoregressive (models)
VR	Verification Rank

Contents

1 Introduction

> We demand rigidly defined areas of doubt and uncertainty![1]
>
> Vroomfondel, in Douglas Adams'
> *The Hitchhiker's Guide to the Galaxy*, 1979

Making forecasts for an uncertain future is a key desire in many aspects of human activity. Any prediction is typically surrounded by uncertainty, and forecasts should thus be probabilistic in nature, taking the form of full probability distributions over future quantities or events (Dawid, 1984; Gneiting, 2008, 2011). Probabilistic forecasts allow to quantify the inherent uncertainty which is essential for good decision making, and have thus become popular over the past few decades. A shift of paradigms from point forecasts to probabilistic forecasts can be observed in various key applications including meteorology, hydrology, seismology, economics, finance, demography, and political science.

With the proliferation of probabilistic forecasting arises the need for decision theoretically principled tools to evaluate the appropriateness of models and predictions. The main focus of the work presented in this dissertation are facets of probabilistic forecasting and comparative model assessment. It will be demonstrated throughout that the tasks of making and evaluating probabilistic forecasts are closely connected. This introduction will serve to outline the central questions that will be addressed in the thesis at hand.

After a brief review of relevant theoretical foundations in Chapter 2, we begin with a focus on forecast verification. Chapter 3 addresses the question how to evaluate probabilistic forecasts with an emphasis on extreme events. In public discussions of the quality of forecasts, attention typically focuses on the predictive performance in cases of extreme events. However, the restriction of conventional forecast evaluation methods to subsets of extreme observations has unexpected and undesired effects, and is bound to discredit skillful forecasts when the signal-to-noise ratio in the data generating process is low. Conditioning on outcomes is incompatible with the theoretical assumptions of established forecast evaluation methods, thereby confronting forecasters with what we refer to as the forecaster's dilemma. For probabilistic forecasts, proper weighted scoring rules have been proposed as decision theoretically justifiable alternatives for forecast evaluation with an emphasis on extreme events. Using theoretical arguments, simulation experiments, and a real data study on probabilistic forecasts of U.S. inflation and gross domestic product growth, we illustrate and discuss the forecaster's dilemma

[1] Adams (1996, p. 115), first edition published 1979.

along with potential remedies.

In Chapter 4, we move the focus from forecast verification to close connections of making and evaluating probabilistic forecasts. A rapidly growing literature uses Bayesian methods to produce probabilistic forecasts of meteorological, economic or financial variables. Thereby, the posterior predictive distribution of interest comes as a simulated sample, typically generated by a Markov chain Monte Carlo (MCMC) algorithm. We conduct a systematic analysis of how to make and evaluate probabilistic forecasts based on such simulation output. Utilizing the mathematical framework provided by the theory of proper scoring rules (Gneiting and Raftery, 2007), we develop a notion of consistency that allows for assessing the adequacy of methods for estimating the stationary distribution underlying the simulation output. We then review asymptotic results that account for the salient features of Bayesian posterior simulators, and derive conditions under which choices from the literature satisfy this notion of consistency. Importantly, these conditions depend on the scoring rule being used, such that the choices of approximation method and scoring rule are intertwined. The theoretical considerations are illustrated in a simulation study and a case study of a popular model for economic time series in order to assess consistency and efficiency of the various approximation methods in practical applications.

Understanding what makes a good probabilistic forecasts is essential for developing forecasting models. In Chapters 5 and 6, we turn to applications in meteorology and investigate probabilistic wind speed forecasting. Nowadays, forecasts of wind speed are usually based on output of numerical weather prediction models which describe the dynamical and physical behavior of the atmosphere through nonlinear partial differential equations. Single deterministic predictions produced by single runs of such models fail to account for uncertainties in the initial conditions and the numerical model. Therefore, models are typically run several times with varying initial conditions and model physics, resulting in an ensemble of forecasts (Palmer, 2002; Gneiting and Raftery, 2005). While the implementation of such ensemble prediction systems is an important step towards probabilistic forecasting, ensemble forecasts tend to be underdispersive and subject to systematic bias, and therefore require statistical postprocessing.

Chapters 5 and 6 address topics in postprocessing methods for ensemble forecasts. In Chapter 5, we investigate the choice a suitable statistical model for wind speed. Building on the non-homogeneous regression approach of Gneiting et al. (2005), we compare different parametric models for wind speed. The standard model based on truncated normal distributions (Thorarinsdottir and Gneiting, 2010) often fails to resolve the heavy right tail of wind speed observations. We therefore propose alternative approaches based on log-normal and generalized extreme value distributions. We further investigate combination models that select one of the candidate distributions based on covariate information, and mixture models where we combine lighter and heavier tailed distributions as weighted mixtures. The various models are compared in three case studies with different ensemble prediction systems and observed wind quantities, and are demonstrated to outperform the basic truncated normal model.

In Chapter 6, we address the question of how to select training sets for estimating the parameters of postprocessing models. In particular, we propose two similarity-based semi-local approaches to parameter estimation where training data for a specific observation station are augmented with corresponding forecast cases from stations with similar characteristics. Similarities between stations are determined using either distance functions or clustering based on various features of the climatology, forecast errors, ensemble predictions and locations of the observation stations. In a case study on wind speed over Europe, the proposed similarity-based semi-local models show improved predictive performance compared to standard estimation methods and allow for efficiently estimating complex models without numerical stability issues.

Chapter 7 concludes this thesis with a summary and discussion of the main results and an outlook to future work.

1.1 Relation to previous and published work

The work presented in this thesis has resulted in the following research articles. All of the articles have been written jointly with one or more coauthors, the specific contributions of individual coauthors are identified in the following.

Chapter 3 and parts of Sections 2.1–2.4 are based on the following research article.

Lerch et al. (2016) Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. and Gneiting, T. (2016). Forecaster's dilemma: Extreme events and forecast evaluation. In revision at *Statistical Science*. Preprint available at `http://arxiv.org/abs/1512.09244`.

The Bayesian models of GDP growth and inflation for the United States used in the case study in Section 3.4 were implemented by Francesco Ravazzolo who provided simulation draws which I used to evaluate the forecasts and produce the figures and tables in Section 3.4.

Initial case studies in a comparable direction have already been investigated in the Diplom thesis of Lerch (2012), however, the presentation in the dissertation at hand will provide insights from a theoretical perspective, as well as extensive simulation evidence and a novel case study.

Chapter 4 is based on the following draft paper which is being prepared for submission at the time of writing.

Krüger et al. (2016) Krüger, F., Lerch, S., Thorarinsdottir, T. L. and Gneiting, T. (2016). Probabilistic forecasting and comparative model assessment based on MCMC output.

In its current form, it has been written jointly by Fabian Krüger and myself, with comments and suggestions by Tilmann Gneiting and Thordis Thorarinsdottir. Specifically, the theoretical considerations and results presented in Sections 4.2 and 4.3 are my own work, whereas Fabian Krüger designed and implemented the

simulation and case studies in Sections 4.4 and 4.5. Fabian Krüger also provided `R` code and data from which I produced the figures in the respective sections.

Chapter 5 is based on the following three published research articles and provides a detailed comparison of the postprocessing models introduced therein.

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

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

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

Sándor Baran provided `R` code for the implementation of EMOS models based on log-normal distributions and weighted mixtures of log-normal and truncated normal distributions, and Thordis Thorarinsdottir provided `R` code for the truncated normal distribution based models.

A simplified variant of the generalized extreme value distribution based model studied by Lerch and Thorarinsdottir (2013) was already investigated in Lerch (2012), however, the results presented here and in Lerch and Thorarinsdottir (2013) are based on substantial extensions. The model formulation and the parameter estimation have been significantly revised, and in new case studies based on different data sets, the model is now compared to novel alternatives that have been proposed in subsequent research.

Chapter 6 is based on the following research article written jointly with Sándor Baran.

Lerch and Baran (2016) Lerch, S. and Baran, S. (2016). Similarity-based semi-local estimation of EMOS models. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, accepted for publication. Preprint available at `http://arxiv.org/abs/1509.03521`.

Further, the following software package for the statistical programming language `R` (R Core Team, 2015) has been developed over the course of this Ph.D. project in joint work with Alexander Jordan and Fabian Krüger.

Jordan et al. (2016) Jordan, A., Krüger, F. and Lerch, S. (2016). `scoringRules` package for `R`. Source code and a manual are available at `https://github.com/FK83/scoringRules`.

The `scoringRules` package is described in detail in Section 2.3.4, parts of which are based on the forthcoming manual that has been written jointly with Alexander Jordan and Fabian Krüger.

2 Preliminaries on probabilistic forecasting and forecast verification

> It seems to me that the condition of confidence or otherwise forms a very important part of the prediction, and ought to find expression.[1]
>
> W. Ernest Cooke, 1906

Probabilistic forecasts in the form of full probability distributions over future quantities or events have become popular over the past few decades, and in various key applications there has been a shift of paradigms from point forecasts to probabilistic forecasts, as reviewed by Tay and Wallis (2000), Timmermann (2000), Gneiting (2008), and Gneiting and Katzfuss (2014), among others. With the proliferation of probabilistic forecasts arises the need for theoretically principled tools to evaluate the quality and appropriateness of models and forecasts in a generalized way.

In this chapter, we introduce important facets of the quality of probabilistic forecasts as well as tools for their verification. Various aspects of the introduced concepts will be revisited and investigated in more detail later.

2.1 Mathematical framework and notation

In a seminal paper on the evaluation of point forecasts, Murphy and Winkler (1987) introduce a mathematical framework for forecast verification based on the joint distribution of forecasts and observations. Gneiting and Ranjan (2013), Ehm et al. (2016), and Strähl and Ziegel (2015) extend and adapt this framework to include the case of potentially multiple probabilistic forecasts. This general setting considers the joint distribution of forecasts and observations on a probability space $(\Omega, \mathcal{A}, \mathbb{Q})$, where the elements of the sample space Ω can be identified with tuples

$$(F_1, \ldots, F_k, Y),$$

the distribution of which is specified by the probability measure $\mathbb{Q}$. The probabilistic forecasts $F_1, \ldots, F_k$, are probability measures on the outcome space $(\Omega_Y, \mathcal{A}_Y)$ for the observation Y. Unless stated otherwise, we restrict our attention to real-valued observations where $\Omega_Y = \mathbb{R}$, and identify probabilistic

[1] Cooke (1906, p. 23)

forecasts F with the associated cumulative distribution function (CDF) F or probability density function (PDF) f.

This measure-theoretic framework allows us to now review the theory on basic aspects of the quality of probabilistic forecasts. We will revisit this framework for forecast evaluation in Section 3.2.1.

2.2 Calibration and sharpness

As argued concisely by Gneiting et al. (2007), the general aim of probabilistic forecasting is to maximize the sharpness of the predictive distribution subject to calibration. *Calibration* is a joint property of the predictive distribution F and the associated observation Y. It essentially requires that the observation is indistinguishable from a random draw from the predictive distribution. *Sharpness* refers to the concentration of the predictive distribution and is a property of the forecasts only.

Various notions of calibration have been proposed. For now, we restrict our attention to *probabilistic calibration*. A probabilistic forecast F is probabilistically calibrated if the *probability integral transform* (PIT) $F(Y)$ is uniformly distributed, with suitable technical adaptations in cases in which F may have a discrete component (Gneiting et al., 2007; Gneiting and Ranjan, 2013). Alternative notions of calibration will be reviewed in Section 3.2.1. Given that a probabilistic forecast is calibrated, it should be as sharp as possible, as clearly, more concentrated forecast distributions indicate a higher information content in the predictions, subject to calibration.

There exist various empirical tools for assessing calibration and sharpness in practical applications. For CDF-valued probabilistic forecasts, checks of the uniformity of PIT values provide an essential device. Given a sample of pairs of probabilistic forecasts and corresponding observations $(F_t, y_t),\ t = 1, \ldots, T$, calibration can be assessed by visual inspection of the histogram of PIT values $F_t(y_t),\ t = 1, \ldots, T$ (Dawid, 1984; Diebold et al., 1998; Gneiting et al., 2007). Deviations from the desired uniform distribution indicate miscalibration. The shape of the histogram can further point towards the reasons of miscalibration, e.g., U-shaped histograms indicate underdispersed forecast distributions with too narrow prediction intervals, whereas inverse-U-shaped histograms correspond to overdispersed forecast distributions with too wide prediction intervals. However, despite their value and popularity, checks of calibration via the uniformity of PIT histograms should be accompanied by an assessment of sharpness, as otherwise misspecifications in the forecast distributions can remain undetected, see Hamill (2001) and Gneiting et al. (2007) for details. Examples of PIT histograms in a practical application will be provided in Chapter 5.

Apart from the visual inspection of PIT histograms, formal statistical test of uniformity can be used to assess calibration. Suitably adapted tests that account for the complex dependence structures in PIT values of sequential k-step-ahead forecasts in time series settings have been proposed in the econometric literature,

see, e.g., Diebold et al. (1998), Corradi and Swanson (2006), and Knüppel (2015), among others. For details, see Section 5.3 where we employ a moment-based test of uniformity proposed by Knüppel (2015) in a comparative assessment of calibration of competing models in probabilistic weather forecasting.

The coverage and width of prediction intervals provide alternative tools to assess calibration and sharpness of predictive distributions. The coverage of an $(1-\alpha)100\,\%,\ \alpha \in (0,1)$, central prediction interval is the proportion of validating observations located between the lower and upper $\alpha/2$ quantiles of the predictive distribution and should be around $(1-\alpha)100\,\%$ for a calibrated probabilistic forecast. Sharper distributions correspond to narrower central prediction intervals, their width thus constitutes a natural measure of sharpness. See Chapter 5 for applications in the verification of probabilistic weather forecasts.

2.3 Proper scoring rules

In the preceding section we have introduced calibration and sharpness as key aspects of the quality of probabilistic forecasts. Proper scoring rules assess calibration and sharpness simultaneously and play key roles in the comparative evaluation and ranking of competing forecasts (Gneiting and Raftery, 2007). Specifically, let $\mathcal{F}$ denote a class of probability distributions on Ω_Y, the set of possible values of the observation Y. A *scoring rule* is a mapping

$$S : \mathcal{F} \times \Omega_Y \longrightarrow \mathbb{R} \cup \{\infty\}$$

that assigns a numerical penalty based on the predictive distribution $F \in \mathcal{F}$ and observation $y \in \Omega_Y$. A scoring rule is *proper* relative to the class $\mathcal{F}$ if

$$\mathbb{E}_{Y\sim G}\, S(G,Y) \leq \mathbb{E}_{Y\sim G}\, S(F,Y) \tag{2.1}$$

for all probability distributions $F, G \in \mathcal{F}$. It is *strictly proper* relative to the class $\mathcal{F}$ if the above holds with equality only if $F = G$. We generally take scoring rules to be negatively oriented penalties that forecasters wish to minimize, i.e., smaller scores indicate better predictions.

Scoring rules provide summary measures of predictive performance, and in practical applications, competing forecasting methods are compared and ranked in terms of the mean score over the cases in a test set. Propriety is a critically important property that encourages honest and careful forecasting, as the expected score is minimized if the quoted predictive distribution agrees with the actually assumed distribution G under which the expectation in (2.1) is computed (Gneiting and Raftery, 2007; Bröcker and Smith, 2007). The use of improper scoring rules can lead to misguided decision-making and inferential procedures, see, e.g., Hilden and Gerds (2014).

For a detailed mathematical analysis of properties and characterizations of proper scoring rules, we refer to Gneiting and Raftery (2007). Measure-theoretic representations reveal connections to convex analysis and provide insight into properties of scoring rules and associated divergences, see Section 2.3.2.

For related work on local proper scoring rules which only depend on the forecast density through both its value and the values of its derivatives at the observation y, see also Ehm and Gneiting (2012), Parry et al. (2012), and Ovcharov (2015a).

2.3.1 Proper scoring rules for real-valued quantities

The most popular proper scoring rules for real-valued quantities are the *logarithmic score* (LogS), defined as

$$\mathrm{LogS}(F, y) = -\log f(y), \tag{2.2}$$

where f denotes the density of F (Good, 1952), which applies to absolutely continuous distributions only, and the *continuous ranked probability score* (CRPS), which is defined as

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(z) - \mathbb{1}\{y \leq z\}\right)^2 \mathrm{d}z \tag{2.3}$$

directly in terms of the predictive CDF (Matheson and Winkler, 1976). Here $\mathbb{1}\{y \leq z\}$ denotes the indicator function which is 1 if $y \leq z$, and 0 otherwise. For any distribution F with finite first moment,

$$\mathrm{CRPS}(F, y) = \mathbb{E}_F|X - y| - \frac{1}{2}\,\mathbb{E}_F|X - X'| \tag{2.4}$$

where X and X' are two independent random variables with distribution F. By representation (2.4), the CRPS is given in the same unit as the observations and generalizes the absolute error.

The CRPS can be interpreted as the integral of the proper Brier score (Brier, 1950; Gneiting and Raftery, 2007),

$$\mathrm{BS}_z(F, y) = \left(F(z) - \mathbb{1}\{y \leq z\}\right)^2, \tag{2.5}$$

for the induced probability forecast for the binary event of the observation not exceeding the threshold value z. Alternative representations of the CRPS are discussed in Gneiting and Raftery (2007) and Gneiting and Ranjan (2011).

The logarithmic score is strictly proper relative to the class $\mathcal{L}_1$ of probability measures with a Lebesgue density, and the CRPS is strictly proper relative to the class of probability measures with finite first moment.

Various other proper scoring rules have been proposed and employed in the theoretical and applied literature. The *quadratic score* (QS) given by

$$\mathrm{QS}(F, y) = \|f\|_2^2 - 2f(y)$$

is strictly proper relative to $\mathcal{F} = \mathcal{L}_2$, the class of probability measures with Lebesgue density f satisfying $\|f\|_2 = \left(\int |f(z)|^2\,\mathrm{d}z\right)^{1/2} < \infty$. The *Hyvärinen score* (HS)

$$\mathrm{HS}(f, y) = 2\,\frac{f''(y)}{f(y)} - \left(\frac{f'(y)}{f(y)}\right)^2 \tag{2.6}$$

proposed by Hyvärinen (2005) is a local proper scoring rule that also involves derivatives of the predictive density. It is proper relative to the class of probability distributions on $\mathbb{R}$ with probability density functions f that are twice continuously differentiable and $(\log f(x))' \to 0$ as $|x| \to \infty$ (Parry et al., 2012). The HS can be computed without knowledge of the normalizing constant of f and thus allows for estimating statistical models where such normalizing constants are unavailable (Gneiting and Katzfuss, 2014).

LogS, QS and HS are restricted to density forecasts which can be impractical, e.g., if the forecasts are only available through a simulated sample from the predictive distribution as we will discuss in Chapter 4. The CRPS is defined in terms of the predictive CDF, however, can be hard to compute analytically for complex classes forecast distributions. Alternatives are given by proper scoring rules that are solely based on the moments of the forecast distributions. The *Dawid-Sebastiani score* (DSS; Dawid and Sebastiani, 1999; Gneiting and Raftery, 2007), is given by

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

where μ_F and σ_F^2 denote the mean and variance of the predictive distribution F, and is proper relative to the class of probability measures with finite second moment. The propriety is strict if the members of the class of probability measures $\mathcal{F}$ are fully characterized by the first two moments. The *error-spread score* (ESS; Christensen et al., 2015) further includes normalized central third moments $\gamma_F = \mathbb{E}_{X \sim F}((X - \mu_F)/\sigma_F)^3$ of the predictive distributions and is given by

$$\mathrm{ESS}(F, y) = \left(\sigma_F^2 - (\mu_F - y)^2 - (\mu_F - y)\sigma_F \gamma_F\right)^2. \tag{2.7}$$

Similar to the DSS, the ESS is only strictly proper if the members of $\mathcal{F}$ are fully characterized by the first three moments.

Some of these alternatives will be discussed in Chapters 3 and 4. In Chapter 3 we will further investigate weighted version of proper scoring rules that allow for emphasizing specific regions of interest.

2.3.2 Score divergences

Denote the expected score of a probabilistic forecast F under the true distribution G by

$$\mathcal{S}(F, G) = \mathbb{E}_{Y \sim G} S(F, Y).$$

Then

$$d_S(F, G) = \mathcal{S}(F, G) - \mathcal{S}(G, G) \tag{2.8}$$

is the *score divergence* associated with the scoring rule S (Gneiting and Raftery, 2007; Thorarinsdottir et al., 2013). Clearly, $d_S(F, G) \geq 0$ for all $F, G \in \mathcal{F}$ if S is proper relative to $\mathcal{F}$. The score divergence associated with the LogS is the Kullback-Leibler divergence (Kullback, 1959)

$$d_{\mathrm{LogS}}(F, G) = \mathrm{KL}(f, g) = \int_{-\infty}^{\infty} g(z) \log\left(\frac{g(z)}{f(z)}\right) \mathrm{d}z,$$

and the score divergence associated with the CRPS is given by

$$d_{\mathrm{CRPS}}(F, G) = \int_{-\infty}^{\infty} (F(z) - G(z))^2 \, \mathrm{d}z,$$

see Chapter 4 for details.

Score divergences are closely related to the concept of Bregman divergences (Bregman, 1967). These connections can be established through representations of proper scoring rules as supergradients of concave functions, see Gneiting and Raftery (2007), Hendrickson and Buehler (1971). In particular, if $\mathcal{F}$ is a convex class of probabilistic forecasts and S is proper relative to $\mathcal{F}$, then the expected score function (or *entropy*)

$$e(F) = \mathcal{S}(F, F)$$

is concave and d_S is a Bregman divergence. In case of infinite sample spaces, e.g., if $\Omega_Y = \mathbb{R}$, technical modifications such as extensions to functional Bregman divergences (Frigyik et al., 2008) are required, see Ovcharov (2015b) for a detailed mathematical analysis. Connections between Bregman divergences and scoring rules have also been studied by Grünwald and Dawid (2004), Buja et al. (2005), and Abernethy and Frongillo (2012), among others.

The representations of proper scoring rules and the connection to Bregman divergences illustrate the close relation of proper scoring rules and convex analysis. Score divergences will be revisited in Chapter 4.

2.3.3 Optimum score estimation

Proper scoring rules provide valuable tools for parameter estimation. Following the general optimum score estimation approach of Gneiting and Raftery (2007), the parameters of a distribution are determined by optimizing the average value of a proper scoring rule as a function of the parameters over a training set. Optimum score estimation based on minimizing the logarithmic score in (2.2) corresponds to maximum likelihood (ML) estimation.

Minimum CRPS estimation, that is, optimum score estimation based on the CRPS in (2.3), provides a robust alternative to ML estimation if closed form expressions for the CRPS of the distribution family of interest are available. As argued by Gneiting et al. (2005) and Gneiting and Raftery (2007), optimum score estimation can be viewed within the framework of M-estimation (Huber, 1964). Asymptotic results for optimum score estimators such as consistency theorems can thus be derived directly from the corresponding results for M-estimators (Huber, 1967). For applications of optimum score estimation in a meteorological problem and further considerations from an applied perspective, see Chapter 5.

In the light of Section 2.3.2, optimum score estimation corresponds to finding parameter values that minimize the score divergence between the empirical distribution of the observations and the class of parametric distributions at hand. Parameter estimation strategies that rely on minimizing Bregman divergences have been employed in various applications in information theory, computer science and statistics, see, e.g., Banerjee et al. (2005), Gutmann and Hirayama

(2011), Stummer and Vajda (2012), and Holland and Ikeda (2016) for overviews. Dawid et al. (2016) study parametric inference based on proper scoring rules from a theoretical perspective.

2.3.4 `R` package `scoringRules`

Over the course of this Ph.D. project, a software package for the statistical programming language `R` (R Core Team, 2015) has been developed in joint work with Alexander Jordan and Fabian Krüger. Source code and a manual are available at `https://github.com/FK83/scoringRules`.

The `scoringRules` package (Jordan et al., 2016) aims to be a convenient dictionary-like reference for computing scoring rules. It offers implementations of the CRPS and the LogS for a variety of distributions F that come up in applied work. Two main classes are parametric distributions like normal, t, or gamma distributions, and distributions that are not known analytically, but are indirectly described through a sample of simulation draws.

Forecasts given as parametric distributions appear for example in probabilistic weather forecasts that are obtained via statistical postprocessing of the output of numerical weather prediction models. Such examples will be studied in detail in Chapter 5. The integral in the definition of the CRPS in equation (2.3) can be expressed in a closed form for many parametric families which allows for an efficient computation, see Jordan (2015) for an extensive list. The `scoringRules` package offers implementations of many of these previously unavailable analytical expressions of the CRPS. For a full list of the implemented parametric families, see Jordan et al. (2016).

In Bayesian forecasting, the posterior predictive distribution of interest is often available only indirectly through a simulated sample typically generated by a Markov Chain Monte Carlo algorithm. In order to compute the value of a proper scoring rule, one must convert the simulated sample into a closed-form distribution via some approximation method. Chapter 4 provides a systematic analysis of this issue from a theoretical and applied perspective. The implementation choices and default settings in the `scoringRules` package follow the findings presented there.

There exist other `R` packages which allow for computing the values of proper scoring rules. The `ensembleBMA` (Fraley et al., 2015) and `ensembleMOS` (Yuen et al., 2013) packages include implementations of the CRPS for normal and gamma distributions, as well as normal and gamma mixtures, see also Fraley et al. (2011) for a detailed description. However, these implementations are tailored to the specific data structures in the application to statistical postprocessing of ensemble weather forecasts, and are not straightforward to apply to more general settings. Further, the `scoringRules` package includes many other parametric distributions that come up in applied work, see, e.g., Chapter 5. The `verification` (National Center for Atmospheric Research, 2015) and `SpecsVerification` (Siegert, 2015) packages offer implementations of the CRPS for distributions given as simulated samples. The `scoringRules` pack-

age additionally provides corresponding implementations for the LogS based on the analysis presented in Chapter 4. For an overview of verification software for a variety of programming languages from a broader perspective, see Pocernich (2012).

2.4 Consistent scoring functions

Traditionally, forecasts used to be deterministic, i.e., given in the form of point predictions for future events. Although it has been widely recognized that forecasts should be probabilistic, some practical applications still require point forecasts for reasons of decision making, reporting requirements or communications (Gneiting and Katzfuss, 2014). The quality of point forecasts is typically assessed by means of a *scoring function* $s(x, y)$ that assigns a numerical score based on the point forecast, x, and the respective observation, y. As in the case of proper scoring rules, competing forecasting methods are compared and ranked in terms of the mean score over the cases in a test set. Popular scoring functions include the *squared error* (SE),

$$\mathrm{SE}(x, y) = (x - y)^2,$$

and the *absolute error* (AE),

$$\mathrm{AE}(x, y) = |x - y|.$$

To avoid misguided inferences, the scoring function and the forecasting task have to be matched carefully, either by specifying the scoring function ex ante, or by employing scoring functions that are *consistent* for a target functional T, relative to the class $\mathcal{F}$ of predictive distributions at hand, in the technical sense that

$$\mathbb{E}_{Y \sim F}\, s(T(F), Y) \leq \mathbb{E}_{Y \sim F}\, s(x, Y)$$

for all $x \in \mathbb{R}$ and $F \in \mathcal{F}$ (Gneiting, 2011). For instance, the squared error scoring function is consistent for the mean or expectation functional relative to the class of the probability measures with finite first moment, and the absolute error scoring function is consistent for the median functional.

Consistent scoring functions become proper scoring rules if the point forecast is chosen to be the Bayes rule or optimal point forecast $\hat{x}$ under the respective predictive distribution, i.e.,

$$\hat{x} = \arg\min_x \mathbb{E}_{Y \sim F} s(x, Y).$$

In other words, if the scoring function s is consistent for the functional T, then

$$S(F, y) = s(T(F), y)$$

defines a proper scoring rule relative to the class $\mathcal{F}$ (Gneiting, 2011). For instance, the squared error can be interpreted as a proper scoring rule provided the point forecast is the mean of the respective predictive distribution, and the absolute error yields a proper scoring rule if the point forecast is the median of the predictive distribution.

3 Forecaster's dilemma: Extreme events and forecast evaluation

Quod male consultum cecidit feliciter, Ancus,
Arguitur sapiens, quo modo stultus erat.
Quod prudenter erat provisum, si male vortat,
Ipse Cato (populo iudice) stultus erat.[1]

John Owen, 1607

In this chapter, we focus on forecast verification and investigate how to evaluate probabilistic forecasts with an emphasis on extreme events. In particular, we discuss the dilemma that occurs if forecast evaluation is restricted to subsets of extreme observations, and study suitably weighted proper scoring rules that can be flexibly tailored to the situation at hand and allow for a decision theoretically principled forecast evaluation.

3.1 Introduction

Extreme events are inherent in natural or man-made systems and may pose significant societal challenges. The development of the theoretical foundations for the study of extreme events started in the middle of the last century and has received considerable interest in various applied domains, including but not limited to meteorology, climatology, hydrology, finance, and economics. Topical reviews can be found in the work of Gumbel (1958), Embrechts et al. (1997), Easterling et al. (2000), Coles (2001), Katz et al. (2002), Beirlant et al. (2004), and Albeverio et al. (2006), among others. Not surprisingly, accurate predictions of extreme events are of great importance and demand. In many situations distinct models and forecasts are available, thereby calling for a comparative assessment of their predictive performance with particular emphasis placed on extreme events.

In the public, forecast evaluation often only takes place once an extreme event has been observed, in particular, if forecasters have failed to predict an event with high economic or societal impact. Table 3.1 gives examples from newspapers, magazines, and broadcasting corporations that demonstrate the focus on extreme

[1] Owen (1607), 216. *Sapientia duce, comite fortuna. In Ancum.* English translation by Edith Sylla (Bernoulli, 2006):
Because what was badly advised fell out happily,
Ancus is declared wise, who just now was foolish;
Because of what was prudently prepared for, if it turns out badly,
Cato himself, in popular opinion, will be foolish.

Table 3.1: Media coverage illustrating the focus on extreme events in public discussions of the quality of forecasts. The sources were accessed January 8, 2016.

Year	Headline	Source
2008	Dr. Doom	The New York Times[1]
2009	How did economists get it so wrong?	The New York Times[2]
2009	He told us so	The Guardian[3]
2010	An exclusive interview with Med Yones - The expert who predicted the financial crisis	CEO Q Magazine[4]
2011	A seer on banks raises a furor on bonds	The New York Times[5]
2013	Meredith Whitney redraws 'map of prosperity'	USA Today[6]
2007	Lessons learned from Great Storm	BBC[7]
2011	Bad data failed to predict Nashville Flood	NBC[8]
2012	Bureau of Meteorology chief says super storm 'just blew up on the city'	The Courier-Mail[9]
2013	Weather Service faulted for Sandy storm surge warnings	NBC[10]
2013	Weather Service updates criteria for hurricane warnings, after Sandy criticism	Washington Post[11]
2015	National Weather Service head takes blame for forecast failures	NBC[12]
2011	Italian scientists on trial over L'Aquila earthquake	CNN[13]
2011	Scientists worry over 'bizarre' trial on earthquake prediction	Scientific American[14]
2012	L'Aquila ruling: Should scientists stop giving advice?	BBC[15]

[1] http://www.nytimes.com/2008/08/17/magazine/17pessimist-t.html
[2] http://www.nytimes.com/2009/09/06/magazine/06Economic-t.html
[3] http://www.guardian.co.uk/business/2009/jan/24/nouriel-roubini-credit-crunch
[4] http://www.ceoqmagazine.com/whopredictedfinancialcrisis/index.htm
[5] http://www.nytimes.com/2011/02/08/business/economy/08whitney.html
[6] http://www.usatoday.com/story/money/business/2013/06/05/meredith-whitney-book-interview/2384905/
[7] http://news.bbc.co.uk/2/hi/science/nature/7044050.stm
[8] http://www.nbc15.com/weather/headlines/January_13_Report_Bad_Data_Failed_To_Predict_Nashville_Flood_113450314.html
[9] http://www.couriermail.com.au/news/queensland/bureau-of-meteorology-under-fire-after-a-weekend-of-wild-weather-and-storms-in-queensland-left-many-unprepared/story-e6freoof-1226519213928
[10] http://www.nbcnewyork.com/news/local/Sandy-Report-Weather-Storm-Surge-Warnings-207545031.html
[11] http://www.washingtonpost.com/blogs/capital-weather-gang/wp/2013/04/04/weather-service-changes-criteria-for-hurricane-warnings-after-sandy-criticism/
[12] http://www.nbcnews.com/storyline/blizzard-15/national-weather-service-head-takes-blame-forecast-failures-n294701
[13] http://edition.cnn.com/2011/09/20/world/europe/italy-quake-trial/
[14] http://www.scientificamerican.com/article/trial-such-as-that-star/
[15] http://www.bbc.co.uk/news/magazine-20097554

events in finance, economics, meteorology, and seismology. Striking examples include the international financial crisis of 2007/08 and the L'Aquila earthquake of 2009. After the financial crisis, much attention was paid to economists who had correctly predicted the crisis, and a superior predictive ability was attributed to them. In 2011, against the protest of many scientists around the world, a group of Italian seismologists was put on trial for not warning the public of the devastating L'Aquila earthquake of 2009 that caused 309 deaths (Hall, 2011). Six scientists and a government official were found guilty of involuntary manslaughter in October 2012 and sentenced to six years of prison each. In November 2015, the scientists were acquitted by the Supreme Court in Rome, whereas the sentence of the deputy head of Italy's civil protection department, which had been reduced to two years in 2014, was upheld.

At first sight, the practice of selecting extreme observations, while discarding non-extreme ones, and to proceed using standard evaluation tools appears to be a natural approach. Intuitively, accurate predictions on the subset of extreme observations may suggest superior predictive ability. However, the restriction of the evaluation to subsets of the available observations has unwanted effects that may discredit even the most skillful forecast available (Denrell and Fang, 2010; Diks et al., 2011; Gneiting and Ranjan, 2011). In a nutshell, if forecast evaluation proceeds conditionally on a catastrophic event having been observed, always predicting calamity becomes a worthwhile strategy. Given that media attention tends to focus on extreme events, skillful forecasts are bound to fail in the public eye, and it becomes tempting to base decision-making on misguided inferential procedures. We refer to this critical issue as the *forecaster's dilemma*.[2]

To demonstrate the phenomenon, we let $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution with mean μ and standard deviation σ and consider the following simple experiment. Let the observation Y satisfy

$$Y \,|\, \mu \sim \mathcal{N}(\mu, \sigma^2) \quad \text{where} \quad \mu \sim \mathcal{N}(0, 1 - \sigma^2). \tag{3.1}$$

Table 3.2 introduces forecasts for Y, showing both the predictive distribution, F, and the associated point forecast, X, which we take to be the respective median or mean. The predictive distributions are symmetric, so their mean and median coincide. We use X in upper case, as the point forecast may depend on μ and τ and, therefore, is a random variable. The perfect forecast has knowledge of μ, while the unconditional forecast is the unconditional standard normal distribution of Y. The deliberately misguided extremist forecast shows a constant bias of $\frac{5}{2}$. As expected, the perfect forecast is preferred under both the mean absolute error (MAE) and the mean squared error (MSE). However, these results change completely if we restrict attention to the largest 5% of the observations, as shown

[2]Our notion of the *forecaster's dilemma* differs from a previous usage of the term in the marketing literature by Ehrman and Shugan (1995), who investigated the problem of influential forecasting in business environments. The forecaster's dilemma in influential forecasting refers to potential complications when the forecast itself might affect the future outcome, for example, by influencing which products are developed or advertised.

Table 3.2: Forecasts in the simulation study, where the observation Y satisfies (3.1) with $\sigma^2 = \frac{2}{3}$ being fixed. The mean absolute error (MAE) and mean squared error (MSE) for the point forecast X are based on a sample of size 10 000; the restricted versions rMAE and rMSE are based on the subset of observations exceeding 1.64 only. The lowest value in each column is in bold.

Forecast	Predictive distribution	X	MAE	MSE	rMAE	rMSE
Perfect	$\mathcal{N}(\mu, \sigma^2)$	μ	**0.64**	**0.67**	1.35	2.12
Unconditional	$\mathcal{N}(0, 1)$	0	0.80	0.99	2.04	4.30
Extremist	$\mathcal{N}(\mu + \frac{5}{2}, \sigma^2)$	$\mu + \frac{5}{2}$	2.51	6.96	**1.16**	**1.61**

in the last two columns of the table, where the misguided extremist forecast receives the lowest mean score.

In this simple example, we have considered point forecasts only, for which there is no obvious way to abate the forecaster's dilemma by adapting existing forecast evaluation methods appropriately, such that particular emphasis can be put on extreme outcomes. Probabilistic forecasts in the form of predictive distributions provide a suitable alternative.

Probabilistic forecasts have become popular over the past few decades, and in various key applications there has been a shift of paradigms from point forecasts to probabilistic forecasts, as reviewed by Tay and Wallis (2000), Timmermann (2000), Gneiting (2008), and Gneiting and Katzfuss (2014), among others, see Chapter 2 for further details. As we will see below, the forecaster's dilemma is not limited to point forecasts and occurs in the case of probabilistic forecasts as well. However, in the case of probabilistic forecasts extant methods of forecast evaluation can be adapted to place emphasis on extremes in decision theoretically coherent ways. In particular, it has been suggested that suitably weighted scoring rules allow for the comparative evaluation of probabilistic forecasts with emphasis on extreme events while retaining propriety (Diks et al., 2011; Gneiting and Ranjan, 2011).

The remainder of this chapter is organized as follows. In Section 3.2 theoretical foundations on forecast evaluation and proper scoring rules are reviewed, serving to analyze and explain the forecaster's dilemma along with potential remedies. In Section 3.3 this is followed up and illustrated in simulation experiments. Furthermore, we elucidate the role of the fundamental lemma of Neyman and Pearson, which suggests the superiority of tests of equal predictive performance that are based on the classical, unweighted logarithmic score. A case study on probabilistic forecasts of gross domestic product (GDP) growth and inflation for the United States comparing the predictive performance of autoregressive and vector-autoregressive models with different specifications of volatility is presented in Section 3.4. We close with a discussion in Section 3.5. The chapter is based on Lerch et al. (2016).

3.2 Forecast evaluation and extreme events

Building on the basic concepts introduced in Chapter 2, we now review relevant theory that is then used to study and explain the forecaster's dilemma.

3.2.1 The joint distribution framework for forecast evaluation

We start by extending our discussion of the mathematical framework for forecast evaluation introduced in Section 2.1. In a seminal paper on the evaluation of point forecasts, Murphy and Winkler (1987) argued that the assessment ought to be based on the joint distribution of the forecast, X, and the observation, Y, building on both the *calibration-refinement factorization*,

$$[X, Y] \;=\; [X]\,[Y|X],$$

and the *likelihood-baserate factorization*,

$$[X, Y] \;=\; [Y]\,[X|Y].$$

Extensions and adaptations of this framework by Gneiting and Ranjan (2013), Ehm et al. (2016), and Strähl and Ziegel (2015) that include the case of potentially multiple probabilistic forecasts have been introduced in Section 2.1. Recall that the joint distribution of the probabilistic forecasts and the observation is then defined on a probability space $(\Omega, \mathcal{A}, \mathbb{Q})$, where the elements of the sample space Ω can be identified with tuples

$$(F_1, \ldots, F_k, Y),$$

the distribution of which is specified by the probability measure $\mathbb{Q}$. The σ-algebra $\mathcal{A}$ can be understood as encoding the information available to forecasters. The predictive distributions $F_1, \ldots, F_k$ are CDF-valued random quantities on the outcome space of the observation, Y. They are assumed to be measurable with respect to their corresponding information sets, which can be formalized as sub-σ-algebras $\mathcal{A}_1, \ldots, \mathcal{A}_k \subseteq \mathcal{A}$. The predictive distribution F_i is *ideal* relative to the information set $\mathcal{A}_i$ if $F_i = [Y|\mathcal{A}_i]$ almost surely. Thus, an ideal predictive distribution makes the best possible use of the information at hand. In the setting of equation (3.1) and Table 3.2, the perfect forecast is ideal relative to knowledge of μ, the unconditional forecast is ideal relative to the empty information set, and the extremist forecast fails to be ideal.

Considering the case of a single probabilistic forecast, F, the above factorizations have immediate analogues in this setting, namely, the calibration-refinement factorization

$$[F, Y] \;=\; [F]\,[Y|F] \tag{3.2}$$

and the likelihood-baserate factorization

$$[F, Y] \;=\; [Y]\,[F|Y]. \tag{3.3}$$

The components of the calibration-refinement factorization (3.2) can be linked to the sharpness and the calibration of a probabilistic forecast (Gneiting et al., 2007). Sharpness refers to the concentration of the predictive distributions and is a property of the marginal distribution of the forecasts only. Calibration can be interpreted in terms of the conditional distribution of the observation, Y, given the probabilistic forecast F.

Various notions of calibration have been proposed. In Section 2.2 we have introduced the notion of probabilistic calibration. Recall that a forecast F is probabilistically calibrated if the probability integral transform $F(Y)$ is uniformly distributed, with suitable technical adaptations in cases in which F may have a discrete component (Gneiting et al., 2007; Gneiting and Ranjan, 2013). Among the alternative notions of calibration, the concept of auto-calibration is particularly strong. Specifically, a probabilistic forecast F is *auto-calibrated* if

$$[Y|F] = F \tag{3.4}$$

almost surely (Tsyplakov, 2013). This property carries over to point forecasts, in that, given any functional T, such as the mean or expectation functional, or a quantile, auto-calibration implies $T([Y|F]) = T(F)$. Furthermore, if the point forecast $X = T(F)$ characterizes the probabilistic forecast, as is the case in Table 3.2, where T can be taken to be the mean or median functional, then auto-calibration implies

$$T([Y|X]) = T([Y|F]) = T(F) = X. \tag{3.5}$$

This property can be interpreted as unbiasedness of the point forecast $X = T(F)$ that is induced by the predictive distribution F. To relate to probabilistic calibration, note that an ideal probabilistic forecast is necessarily auto-calibrated, and an auto-calibrated predictive distribution is necessarily probabilistically calibrated (Gneiting and Ranjan, 2013; Strähl and Ziegel, 2015).

In contrast, the interpretation of the second component $[F|Y]$ in the likelihood-baserate factorization (3.3) is much less clear. While the conditional distribution of the forecast given the observation can be viewed as a measure of discrimination ability, it was noted by Murphy and Winkler (1987) that forecasts can be perfectly discriminatory although they are uncalibrated. Therefore, discrimination ability by itself is not informative, and forecast assessment might be misguided if one stratifies by the realized value of the observation. To demonstrate this, we return to the simpler setting of point forecasts and revisit the simulation example of equation (3.1) and Table 3.2, with $\sigma^2 = \frac{2}{3}$ being fixed. Figure 3.1 shows the perfect forecast, the deliberately misspecified extremist forecast, and the observation in this setting. The bias of the extremist forecast is readily seen when all forecast cases are taken into account. However, if we restrict attention to cases where the observation exceeds a high threshold of 2, it is not obvious whether the perfect or the extremist forecast is preferable. To provide analytical results, $X_{\text{perfect}}|Y = y \sim \mathcal{N}((1-\sigma^2)y, \sigma^2(1-\sigma^2))$ and $X_{\text{extremist}}|Y = y \sim \mathcal{N}\big((1-\sigma^2)y + \frac{5}{2}, \sigma^2(1-\sigma^2)\big)$.

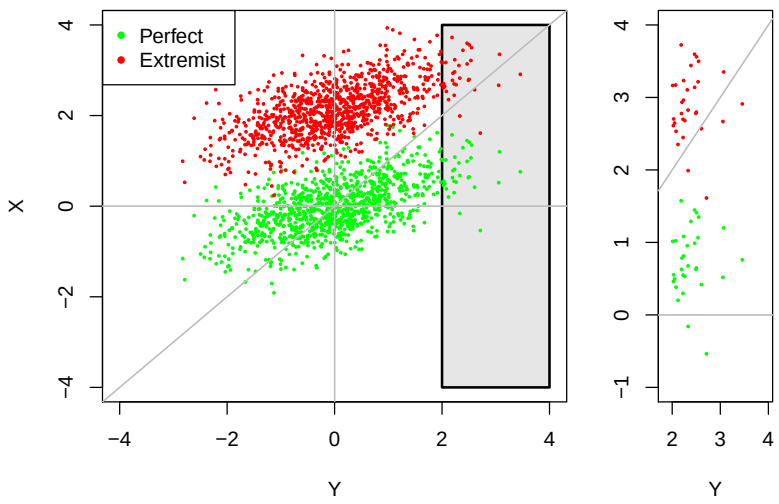


Figure 3.1: The sample illustrates the conditional distribution of the perfect forecast (green) and the extremist forecast (red) given the observation in the setting of equation (3.1) and Table 3.2, where $\sigma^2 = \frac{2}{3}$. The vertical stripe, which is enlarged at right, corresponds to cases where the respective point forecast exceeds a threshold value of 2.

In this simple example, we have seen that if we stratify by the value of the realized observation, a deliberately misspecified forecast may appear appealing, while an ideal forecast may appear flawed, even though the forecasts are based on the same information set. Fortunately, unwanted effects of this type are avoided if we stratify by the value of the forecast. To see this, note that ideal predictive distributions and their induced point forecasts satisfy the auto-calibration property (3.4) and, subject to conditions, the unbiasedness property (3.5), respectively.

In Section 2.3 we introduced proper scoring rules as key tools in the comparative evaluation and ranking of competing forecasts. Recall that a scoring rule $S : \mathcal{F} \times \Omega_Y \longrightarrow \mathbb{R} \cup \{\infty\}$ is proper if

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

for all probability distributions $F, G \in \mathcal{F}$. Proper scoring rules provide summary measures of predictive performance and assess calibration and sharpness simultaneously. Important examples introduced in Section 2.3.1 are the logarithmic score

$$\mathrm{LogS}(F, y) = -\log f(y), \tag{3.6}$$

and the continuous ranked probability score

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 \, \mathrm{d}z. \tag{3.7}$$

3.2.2 Understanding the forecaster's dilemma

We are now in the position to analyze and understand the forecaster's dilemma both within the joint distribution framework and from the perspective of proper scoring rules. While there is no unique definition of extreme events in the literature, we follow common practice and take extreme events to be observations that fall into the tails of the underlying distribution. In public discussions of the quality of forecasts, attention often falls exclusively on cases with extreme observations. As we have seen, under this practice even the most skillful forecasts available are bound to fail in the public eye, particularly when the signal-to-noise ratio in the data generating process is low. In a nutshell, if forecast evaluation is restricted to cases where the observation falls into a particular region of the outcome space, forecasters are encouraged to unduly emphasize this region.

Within the joint distribution framework of Section 3.2.1, any stratification by, and conditioning on, the realized values of the outcome is problematic and ought to be avoided, as general theoretical guidance for the interpretation and assessment of the resulting conditional distribution $[F|Y]$ does not appear to be available. In view of the likelihood-baserate factorization (3.3) of the joint distribution of the forecast and the observation, the forecaster's dilemma arises as a consequence. Fortunately, stratification by, and conditioning on, the values of a point forecast or probabilistic forecast is unproblematic from a decision theoretic perspective, as the auto-calibration property (3.4) lends itself to practical tools and tests for calibration checks, as discussed by Gneiting et al. (2007), Held et al. (2010), and Strähl and Ziegel (2015), among others.

From the perspective of proper scoring rules, Gneiting and Ranjan (2011) showed that a proper scoring rule S_0 is rendered improper if the product with a non-constant weight function $w(y)$ is formed. Specifically, consider the weighted scoring rule

$$S(F, y) = w(y)\, S_0(F, y). \tag{3.8}$$

Then if Y has distribution G with density g, the expected score $\mathbb{E}_{Y\sim G} S(F, Y)$ is minimized by the predictive distribution F with density

$$f(y) = \frac{w(y)g(y)}{\int w(z)g(z)\,\mathrm{d}z}, \tag{3.9}$$

which is proportional to the product of the weight function, w, and the true density, g. In other words, forecasters are encouraged to deviate from their true beliefs and misspecify their predictive densities, with multiplication by the weight function (and subsequent normalization) being an optimal strategy. Therefore, the scoring rule S in (3.8) is improper.

To connect to the forecaster's dilemma, consider the indicator weight function $w_r(y) = \mathbb{1}\{y \geq r\}$. The use of the weight function w_r does not directly correspond to restricting the evaluation set to cases where the observation exceeds or equals the threshold value r, as instead of excluding these cases, a score of zero is assigned to them. However, when forecast methods are compared, the use of the indicator weighted scoring rule corresponds to a multiplicative scaling of the restricted

score, and so the ranking of competing forecasts is the same as that obtained by restricting the evaluation set.

3.2.3 Tailoring proper scoring rules

The forecaster's dilemma gives rise to the question how one might apply scoring rules to probabilistic forecasts when particular emphasis is placed on extreme events, while retaining propriety. To this end, Diks et al. (2011) and Gneiting and Ranjan (2011) consider the use of proper weighted scoring rules that emphasize specific regions of interest.

Diks et al. (2011) propose the *conditional likelihood* (CL) score,

$$\mathrm{CL}(F, y) = -w(y) \log\left(\frac{f(y)}{\int_{-\infty}^{\infty} w(z) f(z)\, \mathrm{d}z} \right), \tag{3.10}$$

and the *censored likelihood* (CSL) score,

$$\mathrm{CSL}(F, y) = -w(y) \log f(y) - (1 - w(y)) \log\left(1 - \int_{-\infty}^{\infty} w(z) f(z)\, \mathrm{d}z \right). \tag{3.11}$$

Here, w is a weight function such that $0 \leq w(z) \leq 1$ and $\int w(z) f(z)\, \mathrm{d}z > 0$ for all potential predictive distributions, where f denotes the density of F. When $w(z) \equiv 1$, both the CL and the CSL score reduce to the unweighted logarithmic score (3.6). Gneiting and Ranjan (2011) propose the *threshold-weighted continuous ranked probability score* (twCRPS), defined as

$$\mathrm{twCRPS}(F, y) = \int_{-\infty}^{\infty} w(z) \left(F(z) - \mathbb{1}\{y \leq z\} \right)^2 \mathrm{d}z, \tag{3.12}$$

where, again, w is a non-negative weight function. When $w(z) \equiv 1$, the twCRPS reduces to the unweighted CRPS (3.7). For recent applications of the twCRPS and a quantile-weighted version of the CRPS see, for example, Cooley et al. (2012), Lerch and Thorarinsdottir (2013), and Manzan and Zerom (2013). Further examples are provided in Chapter 5.

As noted, these scoring rules are proper and can be tailored to the region of interest. When interest centers on the right tail of the distribution, we may choose $w(z) = \mathbb{1}\{z \geq r\}$ for some high threshold r. However, the indicator weight function might result in violations of the regularity conditions for the CL and CSL scoring rule, unless all predictive densities considered are strictly positive. Furthermore, predictive distributions that are identical on $[r, \infty)$, but differ on $(-\infty, r)$, cannot be distinguished. Weight functions based on CDFs as proposed by Amisano and Giacomini (2007) and Gneiting and Ranjan (2011) provide suitable alternatives. For instance, we can set $w(z) = \Phi(z \,|\, r, \sigma^2)$ for some $\sigma > 0$, where $\Phi(\cdot \,|\, \mu, \sigma^2)$ denotes the CDF of a normal distribution with mean μ and variance σ^2. Weight functions emphasizing the left tail of the distribution can be constructed similarly, by using $w(z) = \mathbb{1}\{z \leq r\}$ or $w(z) = 1 - \Phi(z \,|\, r, \sigma^2)$

for some low threshold r. In practice, the weighted integrals in (3.10), (3.11), and (3.12) may need to be approximated by discrete sums, which corresponds to the use of a discrete weight measure, rather than a weight function, as discussed by Gneiting and Ranjan (2011).

In what follows we focus on the above proper variants of the LogS and the CRPS. However, further types of proper weighted scoring rules can be developed. Pelenis (2014) introduces the penalized weighted likelihood score

$$\mathrm{PWL}(F, y) = -w(y) \log f(y) + \int_{-\infty}^{\infty} w(z) f(z) \, \mathrm{d}z - w(y),$$

and the incremental CPRS

$$\mathrm{IncCRPS}(F, y) = \int_{-\infty}^{\infty} w(z)(F(z) - F(z_w(z)) - \mathbb{1}\{z_w(z) \leq y \leq z\})^2 \, \mathrm{d}z,$$

where $z_w(z) = \sup\{A_w^c \cap (-\infty, z]\}$ and $A_w^c = \{y \in \Omega_Y | w(y) = 0\}$. Tödter and Ahrens (2012) and Juutilainen et al. (2012) propose the *continuous ranked logarithmic score* (CRLS),

$$\mathrm{CRLS}(F, y) = -\int_{-\infty}^{\infty} \log |F(z) - \mathbb{1}\{y > z\}| \, \mathrm{d}z,$$

a logarithmic scoring rule that depends on the predictive CDF rather than the predictive density. As hinted at by Juutilainen et al. (2012, p. 466), this score can be generalized to a weighted version, which we call the *threshold-weighted continuous ranked logarithmic score* (twCRLS),

$$\mathrm{twCRLS}(F, y) = -\int_{-\infty}^{\infty} w(z) \log |F(z) - \mathbb{1}\{y > z\}| \, \mathrm{d}z. \tag{3.13}$$

In analogy to the twCRPS (3.12) being a weighted integral of the Brier score in equation (2.5), the twCRLS (3.13) can be interpreted as a weighted integral of the discrete *logarithmic score* (LS) (Good, 1952; Gneiting and Raftery, 2007),

$$\begin{aligned} \mathrm{LS}_z(F, y) &= -\log |F(z) - \mathbb{1}\{y > z\}| \\ &= -\mathbb{1}\{y \leq z\} \log F(z) - \mathbb{1}\{y > z\} \log(1 - F(z)), \end{aligned} \tag{3.14}$$

for the induced probability forecast for the binary event of the observation not exceeding the threshold value z. The aforementioned weight functions and discrete approximations can be employed.

3.2.4 Diebold-Mariano tests

Formal statistical tests of equal predictive performance have been widely used, particularly in the economic literature. Turning now to a time series setting, we consider probabilistic forecasts F_t and G_t for an observation y_{t+k} that lies k time

steps ahead. Given a proper scoring rule S, we denote the respective mean scores on a test set ranging from time $t = 1, \ldots, n$ by

$$\bar{S}_n^F = \frac{1}{n}\sum_{t=1}^{n} S(F_t, y_{t+k}) \qquad \text{and} \qquad \bar{S}_n^G = \frac{1}{n}\sum_{t=1}^{n} S(G_t, y_{t+k}),$$

respectively. Diebold and Mariano (1995) proposed the use of the test statistic

$$t_n = \sqrt{n}\,\frac{\bar{S}_n^F - \bar{S}_n^G}{\hat{\sigma}_n}, \tag{3.15}$$

where $\hat{\sigma}_n^2$ is a suitable estimator of the asymptotic variance of the score difference. Under the null hypothesis of a vanishing expected score difference and standard regularity conditions, the test statistic t_n in (3.15) is asymptotically standard normal (Diebold and Mariano, 1995; Giacomini and White, 2006; Diebold, 2015). When the null hypothesis is rejected in a two-sided test, F is preferred if the test statistic t_n is negative, and G is preferred if t_n is positive.

For $j = 0, 1, \ldots$ let $\hat{\gamma}_j$ denote the lag j sample autocovariance of the sequence $S(F_1, y_{1+k}) - S(G_1, y_{1+k}), \ldots, S(F_n, y_{n+k}) - S(G_n, y_{n+k})$ of score differences. Diebold and Mariano (1995) noted that for ideal forecasts at the k step ahead prediction horizon the respective errors are at most $(k-1)$-dependent. Motivated by this fact, Gneiting and Ranjan (2011) use the estimator

$$\hat{\sigma}_n^2 = \begin{cases} \hat{\gamma}_0 & \text{if } k = 1, \\ \hat{\gamma}_0 + 2\sum_{j=1}^{k-1} \hat{\gamma}_j & \text{if } k \geq 2. \end{cases} \tag{3.16}$$

for the asymptotic variance in the test statistic (3.15). While the at most $(k-1)$-dependence assumption might be violated in practice for various reasons, this appears to be a reasonable and practically useful choice nonetheless. Diks et al. (2011) propose the use of the heteroskedasticity and autocorrelation consistent estimator

$$\hat{\sigma}_n^2 = \hat{\gamma}_0 + 2\sum_{j=1}^{J}\left(1 - \frac{j}{J}\right)\hat{\gamma}_j, \tag{3.17}$$

where J is the largest integer less than or equal to $n^{1/4}$. When this latter estimator is used, larger estimates of the asymptotic variance and smaller absolute values of the test statistic (3.15) tend to be obtained, as compared to using the estimator (3.16), particularly when the sample size n is large.

3.3 Simulation studies

We now present simulation studies. In Section 3.3.1 we mimic the experiment reported on in Table 3.2 for point forecasts, now illustrating the forecaster's dilemma on probabilistic forecasts. Furthermore, we consider the influence of the signal-to-noise ratio in the data generating process. Thereafter in the following

Table 3.3: Mean scores for the probabilistic forecasts in Table 3.2, where the observation Y satisfies (3.1) with $\sigma^2 = \frac{2}{3}$ being fixed. The CRPS and LogS are computed based on all observations, whereas the restricted versions (rCRPS and rLogS) are based on observations exceeding 1.64, the 95th percentile of the population, only. The lowest value in each column is shown in bold.

Forecast	CRPS	LogS	rCRPS	rLogS
Perfect	**0.46**	**1.22**	0.96	2.30
Unconditional	0.57	1.42	1.48	3.03
Extremist	2.05	5.90	**0.79**	**1.88**

Table 3.4: Mean scores for the probabilistic forecasts in Table 3.2, where the observation Y satisfies (3.1) with $\sigma^2 = \frac{2}{3}$ being fixed, under the proper weighted scoring rules twCRPS, CL, and CSL. For each weight function and column, the lowest value is shown in bold.

Threshold r	Forecast	twCRPS	CL	CSL
Indicator weight function, $w(z) = \mathbb{1}\{z \geq 1.64\}$				
1.64	Perfect	**0.018**	$<$ **0.001**	**0.164**
	Unconditional	0.019	0.002	0.204
	Extremist	0.575	0.093	2.205
Gaussian weight function, $w_r(z) = \Phi(z \,\vert\, 1.64, 1)$				
1.64	Perfect	**0.053**	**−0.043**	**0.298**
	Unconditional	0.062	−0.028	0.345
	Extremist	0.673	0.379	1.625

sections, we investigate whether or not there is a case for the use of proper weighted scoring rules, as opposed to their unweighted counterparts, when interest focuses on extremes. As it turns out, the fundamental lemma of Neyman and Pearson (1933) provides theoretical guidance in this regard. All results in this section are based on 10 000 replications.

3.3.1 The influence of the signal-to-noise ratio

Let us recall that in the simulation setting of equation (3.1) the observation satisfies $Y \,|\, \mu \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu \sim \mathcal{N}(0, 1 - \sigma^2)$. In Table 3.2 we have considered three competing point forecasts — termed the perfect, unconditional, and extremist forecasts — and have noted the appearance of the forecaster's dilemma when the quality of the forecasts is assessed on cases of extreme outcomes only.

We now turn to probabilistic forecasts and study the effect of the parameter $\sigma \in (0, 1)$ that governs predictability. Small values of σ correspond to high signal-

to-noise ratios, and large values of σ to small signal-to-noise ratios, respectively. Marginally, Y is standard normal for all values of σ. In the limit as $\sigma \to 0$ the perfect predictive distribution approaches the point measure in the random mean μ; as $\sigma \to 1$ it approaches the unconditional standard normal distribution. The perfect probabilistic forecast is ideal in the technical sense of Section 3.2.1 and thus will be preferred over any other predictive distribution (with identical information basis) by any rational user (Diebold et al., 1998; Tsyplakov, 2013).

In Table 3.3 we report mean scores for the three probabilistic forecasts when $\sigma^2 = \frac{2}{3}$ is fixed. Under the CRPS and LogS the perfect forecast outperforms the others, as expected, and the extremist forecast performs by far the worst. However, these results change drastically if cases with extreme observations are considered only. In analogy to the results in Table 3.2, the perfect forecast is discredited under the restricted scores rCRPS and rLogS, whereas the misguided extremist forecast appears to excel, thereby demonstrating the forecaster's dilemma in the setting of probabilistic forecasts. As shown in Table 3.4, under the proper weighted scoring rules introduced in Section 3.2.3 with weight functions that emphasize the right tail, the rankings under the unweighted CRPS and LogS are restored.

Next we investigate the influence of the signal-to-noise ratio in the data generating process on the appearance and extent of the forecaster's dilemma. As noted, predictability increases with the parameter $\sigma \in (0, 1)$. Figure 3.2 shows the mean CRPS and LogS for the three probabilistic forecasts as a function of σ. The scores for the unconditional forecast do not depend on σ. The predictive performance of the perfect forecast decreases in σ, which is natural, as it is less beneficial to know the value of μ when σ is large. The extremist forecast yields better scores as σ increases, which can be explained by the increase in the predictive variance that allows for a better match between the probabilistic forecast and the true distribution. For the improper restricted scoring rules rCRPS and rLogS, the same general patterns can be observed in Figure 3.3, the mean score increases in σ for the perfect forecast and decreases for the extremist forecast. In accordance with the forecaster's dilemma, the extremist forecast is now perceived to outperform its competitors for all sufficiently large values of σ. However, for small values of σ, when the signal in μ is strong, the rankings are the same as under the CRPS and LogS in Figure 3.2. This illustrates the intuitively obvious observation that the forecaster's dilemma is tied to stochastic systems with moderate to low signal-to-noise ratios, so that predictability is weak.

3.3.2 Power of Diebold-Mariano tests: Diks et al. (2011) revisited

While thus far we have illustrated the forecaster's dilemma, the unweighted CRPS and LogS are well able to distinguish between the perfect forecast and its competitors. In the subsequent sections we investigate whether there are benefits to using proper weighted scoring rules, as opposed to their unweighted versions.

To begin with, we adopt the simulation setting in Section 4 of Diks et al.

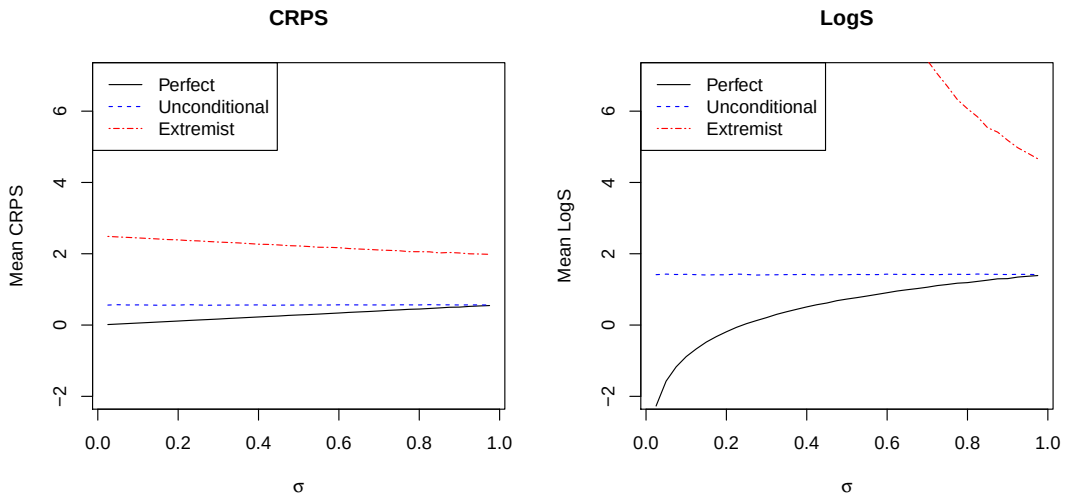


Figure 3.2: Mean CRPS and LogS for the probabilistic forecasts in the setting of equation (3.1) and Table 3.2 as functions of the parameter $\sigma \in (0, 1)$.

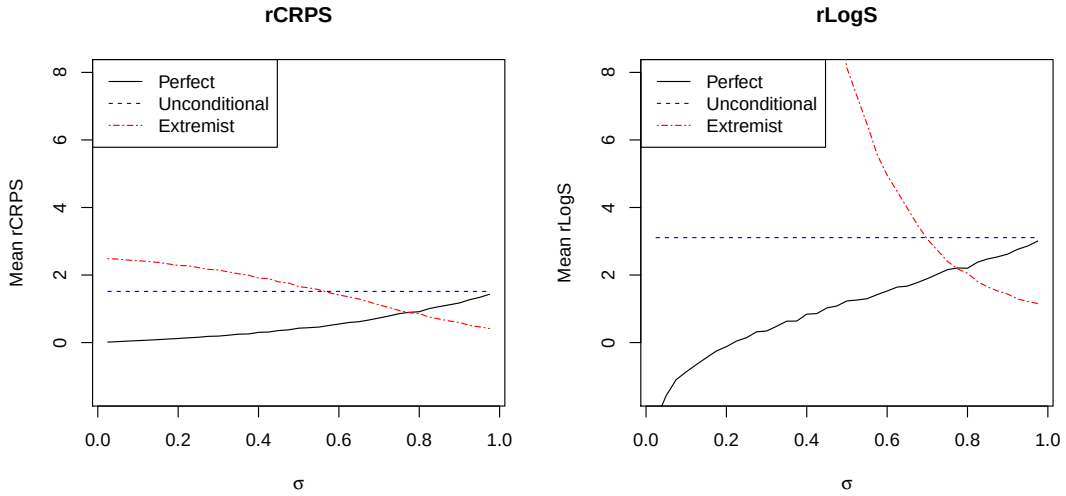


Figure 3.3: Mean of the improper restricted scoring rules rCRPS and rLogS for the probabilistic forecasts in the setting of equation (3.1) and Table 3.2 as functions of the parameter $\sigma \in (0, 1)$. The restricted mean scores are based on the subset of observations exceeding 1.64 only.

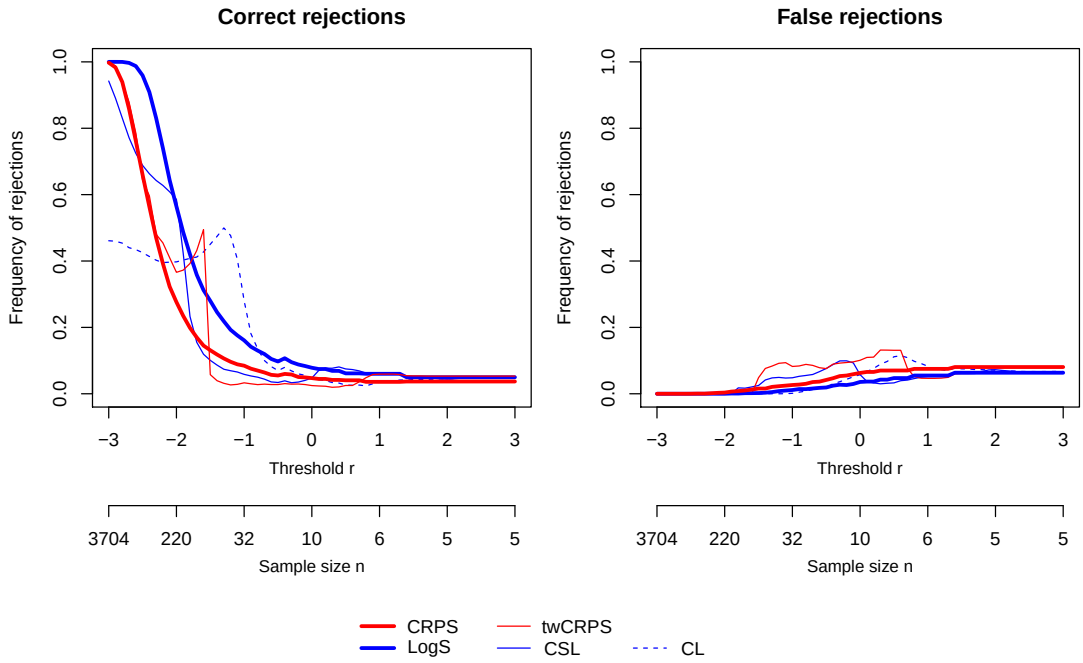


Figure 3.4: Frequency of correct rejections (in favor of the standard normal distribution, left panel) and false rejections (in favor of the Student t distribution, right panel) in two-sided Diebold-Mariano tests in the simulation setting described in Section 3.3.2. The panels correspond to those in the left-hand column of Figure 5 in Diks et al. (2011). The sample size n for the tests depends on the threshold r in the indicator weight function $w(z) = \mathbb{1}\{z \leq r\}$ for the twCRPS, CL, and CSL scoring rules such that under the standard normal distribution there are five expected observations in the relevant interval $(-\infty, r]$.

(2011). Suppose that at time $t = 1, \ldots, n$, the observations y_t are independent standard normal. We apply the two-sided Diebold-Mariano test of equal predictive performance to compare the ideal probabilistic forecast, the standard normal distribution, to a misspecified competitor, a Student t distribution with five degrees of freedom, mean 0, and variance 1. Following Diks et al. (2011), we use the nominal level 0.05, the variance estimate (3.17), and the indicator weight function $w(z) = \mathbb{1}\{z \leq r\}$, and we vary the sample size, n, with the threshold value r in such a way that under the standard normal distribution the expected number, $c = 5$, of observations in the relevant region $(-\infty, r]$ remains constant.

Figure 3.4 shows the proportion of rejections of the null hypothesis of equal predictive performance in favor of either the standard normal or the Student t distribution, respectively, as a function of the threshold value r in the weight function. Rejections in favor of the standard normal distribution represent true power, whereas rejections in favor of the misspecified Student t distribution are

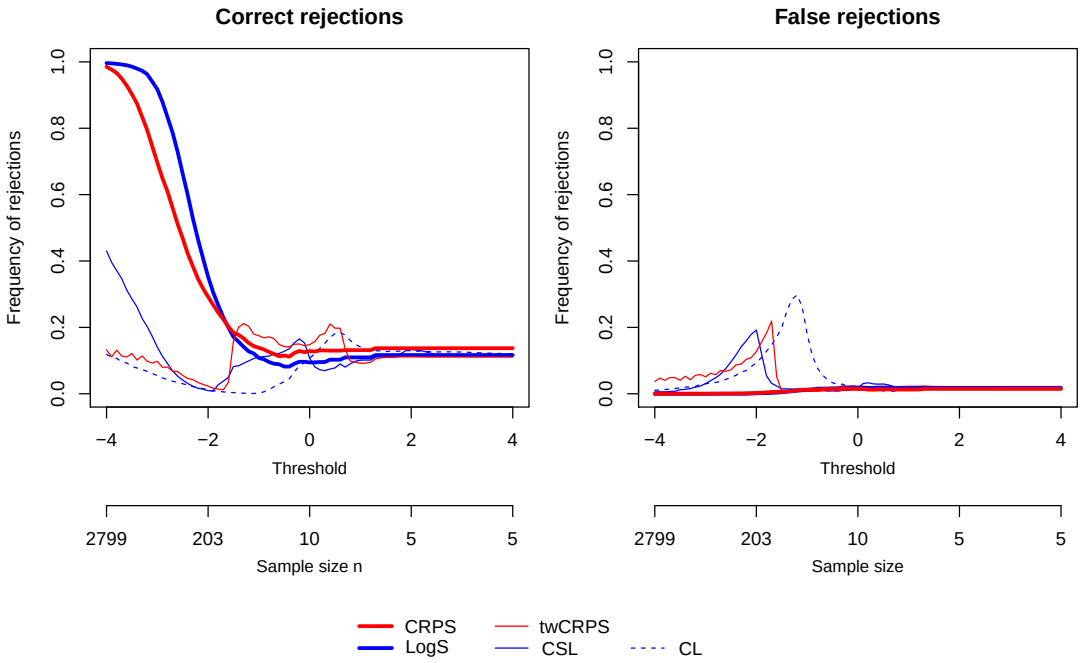


Figure 3.5: Frequency of correct rejections (in favor of the standardized Student t distribution, left panel) and false rejections (in favor of the standard normal distribution, right panel) in two-sided Diebold-Mariano tests in the second variant of the simulation setting described in Section 3.3.2. Compared to Figure 3.4, the roles of true distribution and misspecified forecast distribution are interchanged. The panels correspond to those in the right-hand column of Figure 5 in Diks et al. (2011). The sample size n for the tests depends on the threshold r in the indicator weight function $w(z) = \mathbb{1}\{z \leq r\}$ for the twCRPS, CL, and CSL scoring rules such that under the Student t distribution there are five expected observations in the relevant interval $(-\infty, r]$.

misguided. The curves for the tests based on the twCRPS, CL, and CSL scoring rules agree with those in the left column of Figure 5 of Diks et al. (2011). At first sight, they might suggest that the use of the indicator weight function $w(z) = \mathbb{1}\{z \leq r\}$ with emphasis on the extreme left tail, as reflected by increasingly smaller values of r, yields increased power. At second sight, we need to compare to the power curves for tests using the unweighted CRPS and LogS, based on the same sample size, n, as corresponds to the threshold r at hand. These curves suggest, perhaps surprisingly, that there may not be not be an advantage to using weighted scoring rules. To the contrary, the left-hand panel in Figure 3.4 suggests that tests based on the unweighted LogS are competitive in terms of statistical power.

We further investigate a second variant of the above setting where the roles of

the standard normal distribution and the Student t distribution are interchanged, i.e., the observations $y_t, t = 1, \ldots, n$, are independent realizations of a Student t random variable with mean 0 and variance 1, and the misspecified competitor now is the standard normal distribution. The corresponding rejection rates of two-sided Diebold-Mariano tests are shown in Figure 3.5. In this variant of the simulation study, rejections of the null hypothesis in favor the Student t distribution are correct rejections, and rejections in favor of the standard normal distribution are misguided. As observed before, increasingly smaller values of the threshold r yield higher power if the extreme left tail is emphasized by the proper weighted scoring rules. However, this effect is less pronounced compared to Figure 3.4, and again, the rates of correct rejections in the left panel of Figure 3.5 suggest that tests based on the unweighted variants, particularly the unweighted logarithmic score, are preferable in terms of statistical power.

3.3.3 The role of the Neyman-Pearson lemma

In order to understand this phenomenon, we follow the lead of Feuerverger and Rahman (1992) and draw a connection to a cornerstone of test theory, namely, the fundamental lemma of Neyman and Pearson (1933). In doing so we consider, for the moment, one-sided rather than two-sided tests.

In the simulation setting described by Diks et al. (2011) and in the previous section, any test of equal predictive performance can be re-interpreted as a test of the simple null hypothesis H_0 of a standard normal population against the simple alternative H_1 of a Student t population (and vice versa in the second variant with interchanged roles). We write f_0 and f_1 for the associated density functions and $\mathbb{P}_0$ and $\mathbb{P}_1$ for probabilities under the respective hypotheses. By the Neyman-Pearson lemma (Lehmann and Romano, 2005, Theorem 3.2.1), under H_0 and at any level $\alpha \in (0, 1)$ the unique most powerful test of H_0 against H_1 is the likelihood ratio test. The likelihood ratio test rejects H_0 if $\prod_{t=1}^n f_1(y_t) / \prod_{t=1}^n f_0(y_t) > k$ or, equivalently, if

$$\sum_{t=1}^{n} \log f_1(y_t) - \sum_{t=1}^{n} \log f_0(y_t) > \log k, \tag{3.18}$$

where the critical value k is such that

$$\mathbb{P}_0\left(\frac{\prod_{t=1}^n f_1(y_t)}{\prod_{t=1}^n f_0(y_t)} > k \right) = \alpha.$$

Due to the optimality property of the likelihood ratio test, its power,

$$\mathbb{P}_1\left(\frac{\prod_{t=1}^n f_1(y_t)}{\prod_{t=1}^n f_0(y_t)} > k \right), \tag{3.19}$$

gives a theoretical upper bound on the power of any test of H_0 versus H_1. Furthermore, the optimality result is robust, in the technical sense that minor misspecifications of either H_0 or H_1, as quantified by the Kullback-Leibler divergence, lead to minor loss of power only (Eguchi and Copas, 2006).

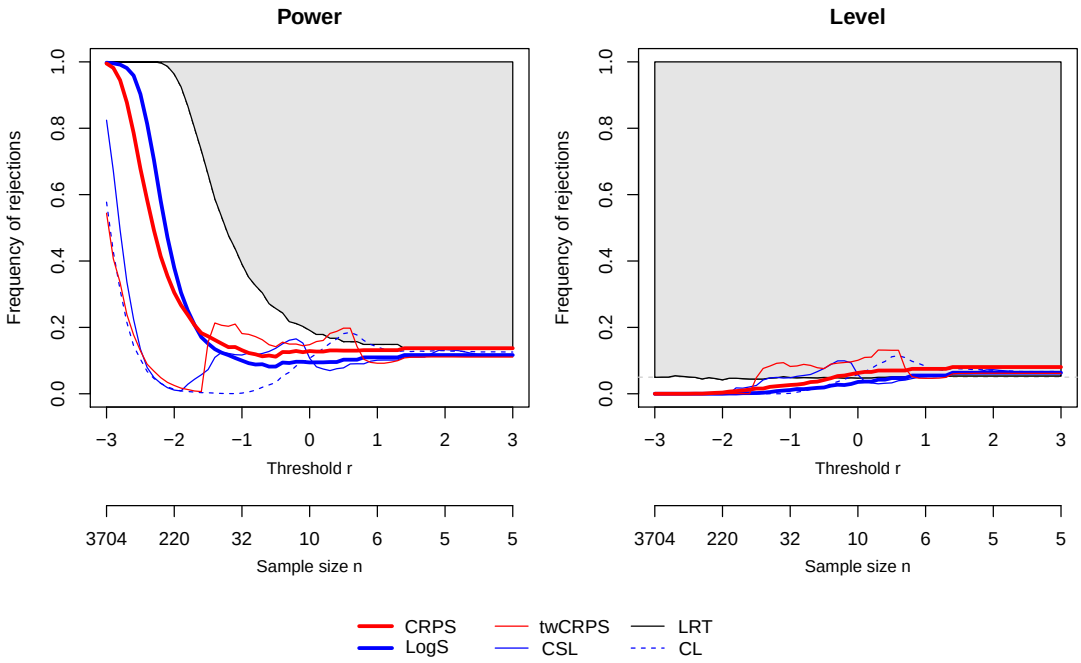


Figure 3.6: Power (left) and level (right) of the likelihood ratio test (LRT) and one-sided Diebold-Mariano tests in the first variant of the simulation setting described in Section 3.3.2. The sample size n for the tests depends on the threshold r in the indicator weight function $w(z) = \mathbb{1}\{z \leq r\}$ for the twCRPS, CL, and CSL scoring rules such that under the standard normal distribution there are five expected observations in the relevant interval $(-\infty, r]$. In the panel for power, the shaded area above the curve for the LRT corresponds to theoretically unattainable values for a test with nominal level. In the panel for level, the dashed line indicates the nominal level.

We now compare to the one-sided Diebold-Mariano test based on the logarithmic score (3.6). This test uses the statistic (3.15) and rejects H_0 if

$$\sum_{t=1}^{n} \log f_1(y_t) - \sum_{t=1}^{n} \log f_0(y_t) > \sqrt{n}\,\hat{\sigma}_n z_{1-\alpha}, \tag{3.20}$$

where $z_{1-\alpha}$ is a standard normal quantile and $\hat{\sigma}_n^2$ is given by (3.16) or (3.17). Comparing with (3.18), we see that the one-sided Diebold-Mariano test that is based on the LogS has the same type of rejection region as the likelihood ratio test. However, the Diebold-Mariano test uses an estimated critical value, which may lead to a level less or greater than the nominal level, α, whereas the likelihood ratio test uses the (in the practice of forecasting unavailable) critical value that guarantees the desired nominal level α.

In this light, it is not surprising that the one-sided Diebold-Mariano test based

on the LogS has power close to the theoretical optimum in (3.19). We illustrate this in Figure 3.6, where we plot the power and size of the likelihood ratio test and one-sided Diebold-Mariano tests based on the CRPS, twCRPS, LogS, CL, and CSL in the first variant of the setting of the previous section. In this variant, the employed one-sided Diebold-Mariano test is a test of the simple null hypothesis H_0 of a standard normal population against the simple alternative H_1 of a Student t population. For small threshold values, the Diebold-Mariano test based on the unweighted LogS has much higher power than tests based on the weighted scores, even though it does not reach the power of the likelihood ratio test, which can be explained by the use of an estimated critical value and incorrect size properties. The theoretical upper bound on the power is violated by Diebold-Mariano tests based on the twCRPS and CL for threshold values between 0 and 1. However, the level of these tests exceeds the nominal level of $\alpha = 0.05$ with too frequent rejections of H_0.

Corresponding results for the second variant of the simulation setting are shown in Figure 3.7. Here, the simple null hypothesis H_0 of a Student t distribution is tested against the simple alternative H_1 of a standard normal distribution. As before, the one-sided Diebold-Mariano test based on the unweighted LogS has higher power compared to tests based on the weighted scores, except for the CL scoring rule and values of the threshold r between -2 and -1. Interestingly, the power of tests based on the CL scoring rule decreases rapidly for increasingly extreme threshold values even though the sample size increases. This behavior will be discussed further in Section 3.3.4. The theoretical upper bound on the power is again violated by tests based on the CL scoring rule, however, tests based on all weighted scoring rules show incorrect size properties with too frequent rejections of H_0.

In the setting of two-sided tests, the connection to the Neyman-Pearson lemma is less straightforward, but the general principles remain valid and provide a partial explanation of the behavior seen in Figures 3.4 and 3.5.

3.3.4 Power of Diebold-Mariano tests: Further experiments

In the simulation experiments just reported, Diebold-Mariano tests based on proper weighted scoring rules generally are unable to outperform tests based on traditionally used, unweighted scoring rules. Several potential reasons come to mind. As we have just seen, when the true data generating process is given by one of the competing forecast distributions, the Neyman-Pearson lemma points at the superiority of tests based on the unweighted LogS. Furthermore, in the simulation setting considered thus far, the distributions considered differ both in the center, the left tail, and the right tail, and the test sample size varied with the threshold for the weight function in a peculiar way.

Therefore, we now consider a revised simulation setting, where we compare two forecast distributions neither of which corresponds to the true sampling distribution, where the forecast distributions only differ on the positive half-axis, and where the test sample size is fixed at $n = 100$. The three candidate distributions

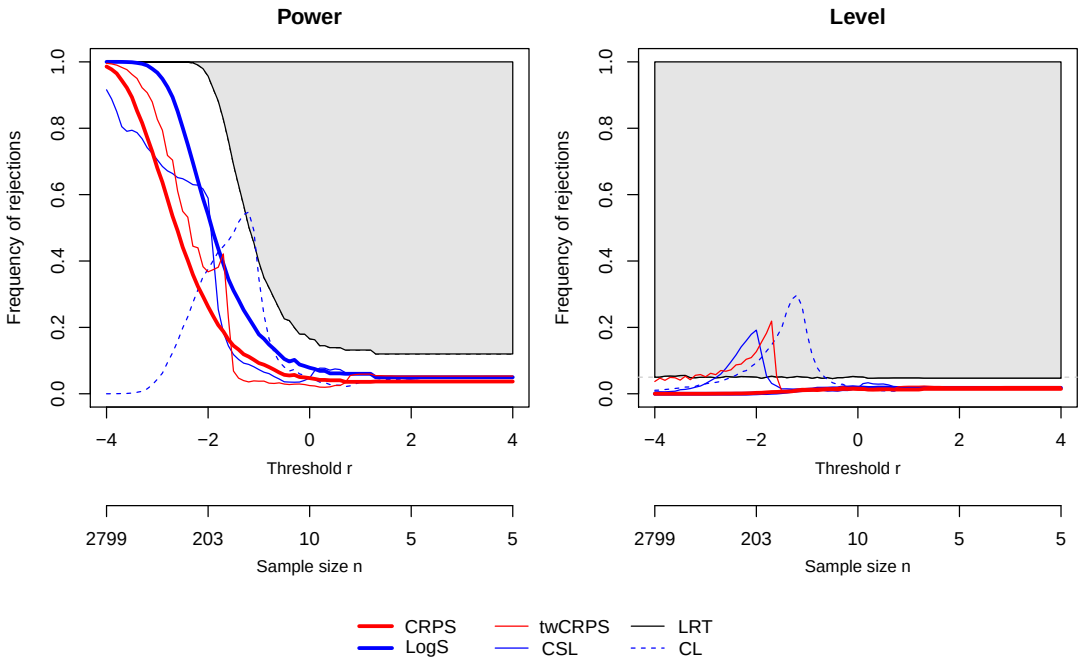


Figure 3.7: Power (left) and level (right) of the likelihood ratio test (LRT) and one-sided Diebold-Mariano tests in the second variant of the simulation setting described in Section 3.3.2. Compared to Figure 3.6 the null hypothesis H_0 and the alternative H_1 are interchanged. The sample size n for the tests depends on the threshold r in the indicator weight function $w(z) = \mathbb{1}\{z \leq r\}$ for the twCRPS, CL, and CSL scoring rules such that under the Student t distribution there are five expected observations in the relevant interval $(-\infty, r]$. In the panel for power, the shaded area above the curve for the LRT corresponds to theoretically unattainable values for a test with nominal level. In the panel for level, the dashed line indicates the nominal level.

are given by Φ, a standard normal distribution with density φ, by a heavy-tailed distribution H with density[3]

$$h(x) = \mathbb{1}\{x \leq 0\}\, \varphi(x) + \mathbb{1}\{x > 0\}\, \frac{3}{8} \left(1 + \frac{x^2}{4}\right)^{-5/2},$$

and by an equally weighted mixture F of Φ and H, with density

$$f(x) = \frac{1}{2}\left(\varphi(x) + h(x)\right).$$

The three forecast distributions are illustrated in Figure 3.8. We perform two-sided Diebold-Mariano tests of equal predictive performance based on the CRPS,

[3] On the positive half axis H coincides with a Student t distribution with 4 degrees of freedom, but compared to the simulation setting in Section 3.3.2 we forgo the standardization.

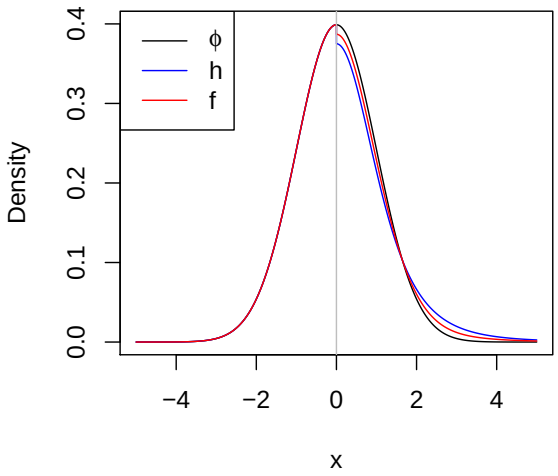


Figure 3.8: The three densities in the simulation setting described in Section 3.3.4. All densities coincide on the negative half axis.

twCRPS, LogS, CL, and CSL.

In Scenario A, the data are a sample from the standard normal distribution Φ, and we compare the forecasts F and H, respectively. In Scenario B, we interchange the roles of Φ and H, that is, the data are a sample from H, and we compare the forecasts F and Φ. The Neyman-Pearson lemma does not apply in this setting. However, the definition of F as a weighted mixture of the true distribution and a misspecified competitor lets us expect that F is to be preferred over the latter. Indeed, by Proposition 3 of Nau (1985), if $F = w\,G + (1-w)\,H$ with $w \in [0,1]$ is a convex combination of G and H, then

$$\mathbb{E}_{Y\sim G}\, S(G,Y) \leq \mathbb{E}_{Y\sim G}\, S(F,Y) \leq \mathbb{E}_{Y\sim G}\, S(H,Y)$$

for any proper scoring rule S. As any utility function induces a proper scoring rule via the respective Bayes act, this implies that under G any rational decision maker favors F over H (Dawid, 2007; Gneiting and Raftery, 2007).

We estimate the frequencies of rejections of the null hypothesis of equal predictive performance at level $\alpha = 0.05$. The choice of the estimator for the asymptotic variance of the score difference in the Diebold-Mariano test statistic (3.15) does not have a recognizable effect in this setting, and so we show results under the estimator (3.16) with $k = 1$ only.

Figure 3.9 shows rejection rates under Scenario A in favor of F and H, respectively, as a function of the threshold r in the indicator weight function $w(z) = \mathbb{1}\{z \geq r\}$ for the weighted scoring rules. The frequency of the desired rejections in favor of F increases with larger thresholds for tests based on the twCRPS and CSL, thereby suggesting an improved discrimination ability at high threshold values. Under the CL scoring rule, the rejection rate decreases rapidly for larger threshold values. This can be explained by the fact that the weight function is a multiplicative component of the CL score in (3.10). As r becomes

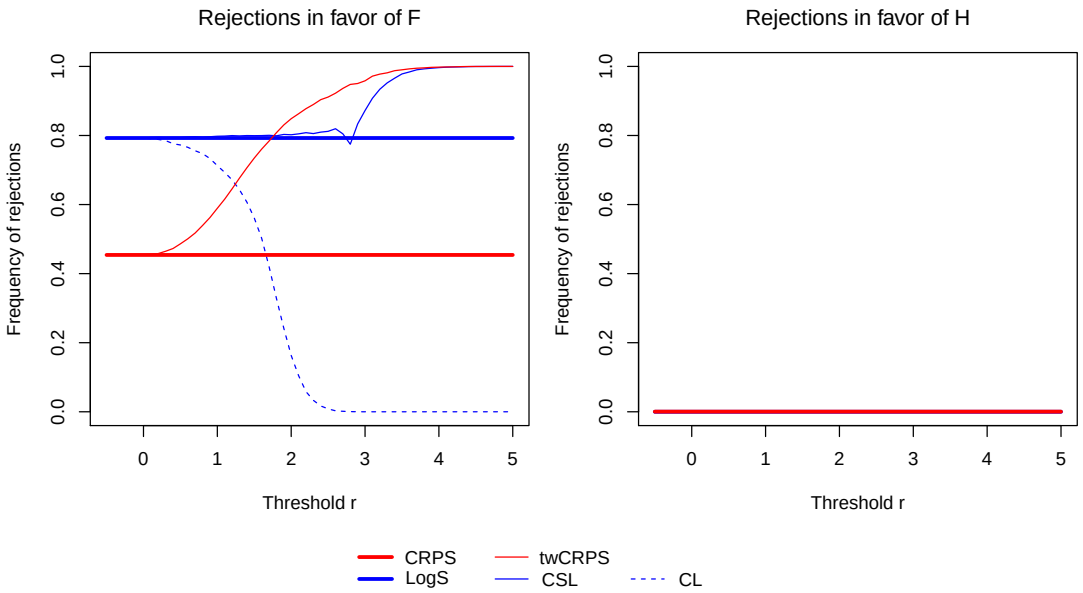


Figure 3.9: Scenario A in Section 3.3.4. The null hypothesis of equal predictive performance of F and H is tested under a standard normal population. The panels show the frequency of rejections in two-sided Diebold-Mariano tests in favor of either F (desired, left) or H (misguided, right). The tests under the twCRPS, CL, and CSL scoring rules use the weight function $w(z) = \mathbb{1}\{z \geq r\}$, and the sample size is fixed at $n = 100$.

larger and larger, none of the 100 observations in the test sample exceed the threshold, and so the mean scores under both forecasts vanish. This can also be observed in Figure 3.4, where, however, the effect is partially concealed by the increase of the sample size for more extreme threshold values. The effect is more pronounced and better visible in Figure 3.5.

Interestingly, an issue very similar to that for the CL scoring rule arises in the assessment of deterministic forecasts of rare and extreme binary events, where performance measures based on contingency tables have been developed and standard measures degenerate to trivial values as events become rarer (Marzban, 1998; Stephenson et al., 2008), posing a challenge that has been addressed by Ferro and Stephenson (2011). The proposed performance measures are discussed in more detail in Appendix 3.A.

Figure 3.10 shows the respective rejection rates under Scenario B, where the sample is generated from the heavy-tailed distribution H, and the forecasts F and Φ are compared. In contrast to the previous examples the Diebold-Mariano test based on the CRPS shows a higher frequency of the desired rejections in favor of F than the test based on the LogS. However, for the tests based on proper weighted scoring rules, the frequency of the desired rejections in favor of F decays to zero with increasing threshold value, and for the tests based on the twCRPS

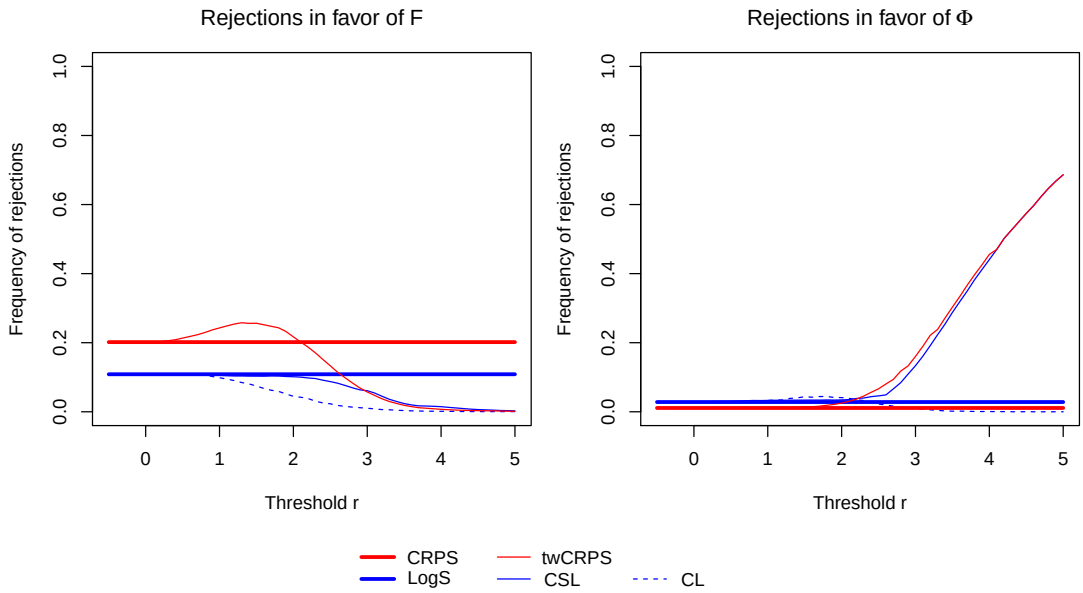


Figure 3.10: Scenario B in Section 3.3.4. The null hypothesis of equal predictive performance of F and Φ is tested under a Student t population. The panels show the frequency of rejections in two-sided Diebold-Mariano tests in favor of either F (desired, left) or Φ (misguided, right). The tests under the twCRPS, CL, and CSL scoring rules use the weight function $w(z) = \mathbb{1}\{z \geq r\}$, and the sample size is fixed at $n = 100$.

and CSL, the frequency of the undesired rejections in favor of Φ rises for larger threshold values.

This seemingly counterintuitive observation can be explained by the tail behavior of the forecast distributions, as follows. Consider the twCRPS and CSL with the indicator weight function $w(z) = \mathbb{1}\{z \geq r\}$ and a threshold r that exceeds the maximum of the given sample. In this case, the scores do not depend on the observations, and are solely determined by the respective tail probabilities, with the lighter tailed forecast distribution receiving the better score. In a nutshell, when the emphasis lies on a low-probability region with few or no observations, the forecaster assigning smaller probability to this region will be preferred. Analytical results are provided in Appendix 3.B. The traditionally used unweighted scoring rules do not depend on a threshold and thus do not suffer from this deficiency.

In comparisons of the mixture distribution F and the lighter-tailed forecast distribution Φ this leads to a loss of finite sample discrimination ability of the proper weighted scoring rules as the threshold r increases. This observation also suggests that any favorable finite sample behavior of the Diebold-Mariano tests based on weighted scoring rules in Scenario A might be governed by rejections due to the lighter tails of F compared to H.

In summary, even though the simulation setting at hand was specifically tai-

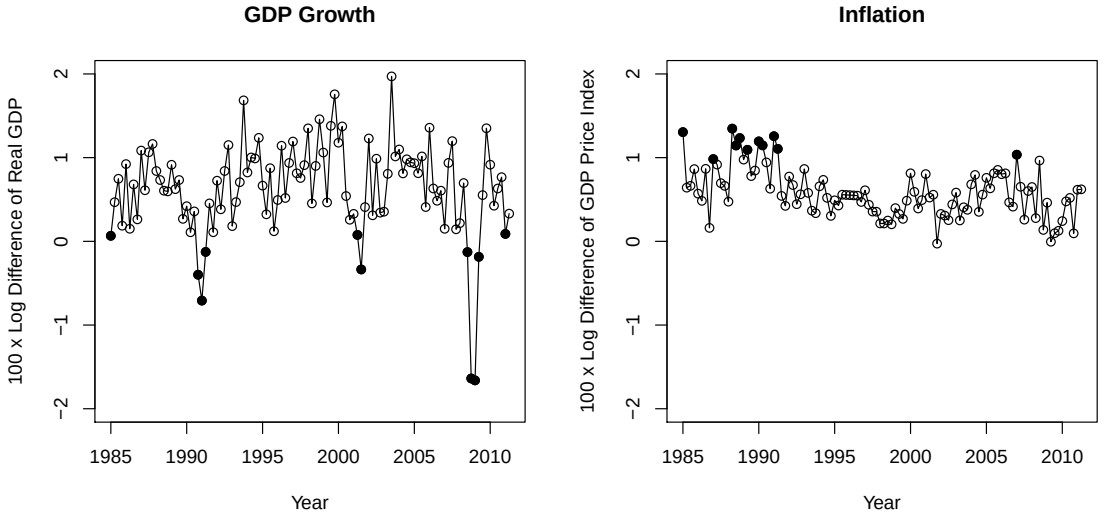


Figure 3.11: Observations of GDP growth and inflation in the U.S. from the first quarter of 1985 to the second quarter of 2011. Solid circles indicate observations considered here as extreme events.

lored to benefit proper weighted scoring rules, these do not consistently perform better in terms of statistical power when compared to their unweighted counterparts. Any advantages vanish at increasingly extreme threshold values in case the actually superior distribution has heavier tails.

3.4 Case study

Based on the work of Clark and Ravazzolo (2015), we compare probabilistic forecasting models for key macroeconomic variables for the United States, serving to demonstrate the forecaster's dilemma and the use of proper weighted scoring rules in an application setting.

3.4.1 Data

We consider time series of quarterly gross domestic product (GDP) growth, computed as 100 times the log difference of real GDP, and inflation in the GDP price index (henceforth *inflation*), computed as 100 times the log difference of the GDP price index, over an evaluation period from the first quarter of 1985 to the second quarter of 2011, as illustrated in Figure 3.11. The data are available from the Federal Reserve Bank of Philadelphia's real time dataset.[4]

For each quarter t in the evaluation period, we use the real-time data vintage t to estimate the forecasting models and construct forecasts for period t and beyond. The data vintage t includes information up to time $t-1$. The one-quarter ahead forecast is thus a current quarter (t) forecast, while the two-quarter ahead

[4] `http://www.phil.frb.org/research-and-data/real-time-center/real-time-data/`

forecast is a next quarter $(t+1)$ forecast, and so forth (Clark and Ravazzolo, 2015). Here we focus on forecast horizons of one and four quarters ahead.

As the GDP data are continually revised, it is not immediate which revision should be used as the realized observation. We follow Romer and Romer (2000) and Faust and Wright (2009) who use the second available estimates as the actual data. Specifically, suppose a forecast for quarter $t+k$ is issued based on the vintage t data ending in quarter $t-1$. The corresponding realized observation is then taken from the vintage $t+k+2$ data set. This approach may entail structural breaks in case of benchmark revisions, but is comparable to real-world forecasting situations where noisy early vintages are used to estimate predictive models (Faust and Wright, 2009).

3.4.2 Forecasting models

We consider autoregressive (AR) and vector autoregressive (VAR) models, the specifications of which are given now. For further details and a discussion of alternative models, see Clark and Ravazzolo (2015).

Our baseline model is an AR(p) scheme with constant shock variance. Under this model, the conditional distribution of Y_t is given by

$$Y_t \,|\, \mathbf{y}_{<t}, b_0, \ldots, b_p, \sigma \sim \mathcal{N}\left(b_0 + \sum_{i=1}^{p} b_i y_{t-i}, \sigma^2\right), \tag{3.21}$$

where $p = 2$ for GDP growth and $p = 4$ for inflation. Here, $\mathbf{y}_{<t}$ denotes the vector of the realized values of the variable Y prior to time t. We estimate the model parameters $b_0, \ldots, b_p$ and σ in a Bayesian fashion using Markov chain Monte Carlo under a recursive estimation scheme, where the data sample $\mathbf{y}_{<t}$ is expanded as forecasting moves forward in time. The predictive distribution then is the Gaussian variance-mean mixture

$$\frac{1}{m}\sum_{j=1}^{m} \mathcal{N}\left(b_0^{(j)} + \sum_{i=1}^{p} b_i^{(j)} y_{t-i},\ \sigma^{2(j)}\right), \tag{3.22}$$

where $m = 5\,000$ and $(b_0^{(1)}, \ldots, b_p^{(1)}, \sigma^{(1)}), \ldots, (b_0^{(m)}, \ldots, b_p^{(m)}, \sigma^{(m)})$ is a sample from the posterior distribution of the model parameters. For the other forecasting models, we proceed analogously.

A more flexible approach is the Bayesian AR model with time-varying parameters and stochastic specification of the volatility (AR-TVP-SV) proposed by Cogley and Sargent (2005), which has the hierarchical structure given by

$$\begin{aligned} Y_t \,|\, \mathbf{y}_{<t}, b_{0,t}, \ldots, b_{p,t}, \lambda_t &\sim \mathcal{N}\left(b_{0,t} + \sum_{i=1}^{p} b_{i,t} y_{t-i},\ \lambda_t\right), \\ b_{i,t} \,|\, b_{i,t-1}, \tau &\sim \mathcal{N}\left(b_{i,t-1}, \tau^2\right), \quad i = 0, \ldots, p, \\ \log \lambda_t \,|\, \lambda_{t-1}, \sigma &\sim \mathcal{N}\left(\log \lambda_{t-1}, \sigma^2\right). \end{aligned} \tag{3.23}$$

Again, we set $p = 2$ for GDP growth and $p = 4$ for inflation.

In a multivariate extension of the AR models, we consider VAR schemes where GDP growth, inflation, unemployment rate, and three-month government bill rate are modeled jointly. Specifically, the conditional distribution of the four-dimensional vector $\mathbf{Y}_t$ is given by the multivariate normal distribution

$$\mathbf{Y}_t \,|\, \mathbf{Y}_{<t}, \mathbf{b}_0, \mathbf{B}_1, \ldots, \mathbf{B}_p, \boldsymbol{\Sigma} \sim \mathcal{N}_4\left(\mathbf{b}_0 + \sum_{i=1}^{p} \mathbf{B}_i \mathbf{y}_{t-1}, \boldsymbol{\Sigma}\right), \tag{3.24}$$

where $\mathbf{Y}_{<t}$ denotes the data prior to time t, $\boldsymbol{\Sigma}$ is a 4×4 covariance matrix, $\mathbf{b}_0$ is a vector of intercepts, and $\mathbf{B}_i$ is a 4×4 matrix of lag i coefficients, where $i = 1, \ldots, p$. Here we take $p = 4$. The univariate predictive distributions for GDP growth and inflation arise as the respective margins of the multivariate posterior predictive distribution.

Finally, we consider a VAR model with time-varying parameters and stochastic specification of the volatility (VAR-TVP-SV), which is a multivariate extension of the AR-TVP-SV model (Cogley and Sargent, 2005). Let $\boldsymbol{\beta}_t$ denote the vector of size $4(4p + 1)$ comprising the parameters $\mathbf{b}_{0,t}$ and $\mathbf{B}_{1,t}, \ldots, \mathbf{B}_{p,t}$ at time t, set $\boldsymbol{\Lambda}_t = \operatorname{diag}(\lambda_{1,t}, \ldots, \lambda_{4,t})$ and let $\mathbf{A}$ be a lower triangular matrix with ones on the diagonal and non-zero random coefficients below the diagonal. The VAR-TVP-SV model takes the hierarchical form

$$\begin{aligned}
\mathbf{Y}_t \,|\, \mathbf{Y}_{<t}, \boldsymbol{\beta}_t, \boldsymbol{\Lambda}_t, \mathbf{A} &\sim \mathcal{N}_4\left(\mathbf{b}_{0,t} + \sum_{i=1}^{p} \mathbf{B}_{i,t} \mathbf{y}_{t-1}, \mathbf{A}^{-1} \boldsymbol{\Lambda}_t (\mathbf{A}^{-1})^\top\right), \\
\boldsymbol{\beta}_t \,|\, \boldsymbol{\beta}_{t-1}, \mathbf{Q} &\sim \mathcal{N}_{4(4p+1)}\big(\boldsymbol{\beta}_{t-1}, \mathbf{Q}\big), \\
\log \lambda_{i,t} \,|\, \lambda_{i,t-1}, \sigma_i &\sim \mathcal{N}\big(\log \lambda_{i,t-1}, \sigma_i^2\big), \quad i = 1, \ldots, 4.
\end{aligned} \tag{3.25}$$

We set $p = 2$ and refer to Clark and Ravazzolo (2015) for further details of the notation, the model, and its estimation.

Figure 3.12 shows one-quarter ahead forecasts of GDP growth over the evaluation period. The baseline models with constant volatility generally exhibit wider prediction intervals, while the TVP-SV models show more pronounced fluctuations both in the median forecast and the associated uncertainty. In 1992 and 1996, the Bureau of Economic Analysis performed benchmark data revisions, which causes the prediction uncertainty of the baseline models to increase substantially. The more flexible TVP-SV models seem less sensitive to these revisions.

3.4.3 Results

To compare the predictive performance of the four forecasting models, Table 3.5 shows the mean CRPS and LogS over the evaluation period. For the LogS, we follow extant practice in the economic literature and employ the quadratic approximation proposed by Adolfson et al. (2007). Specifically, we find the mean, $\hat{\mu}_F$, and variance, $\hat{\sigma}_F^2$, of a sample $\hat{x}_1, \ldots, \hat{x}_m$, where $\hat{x}_i$ is a random number

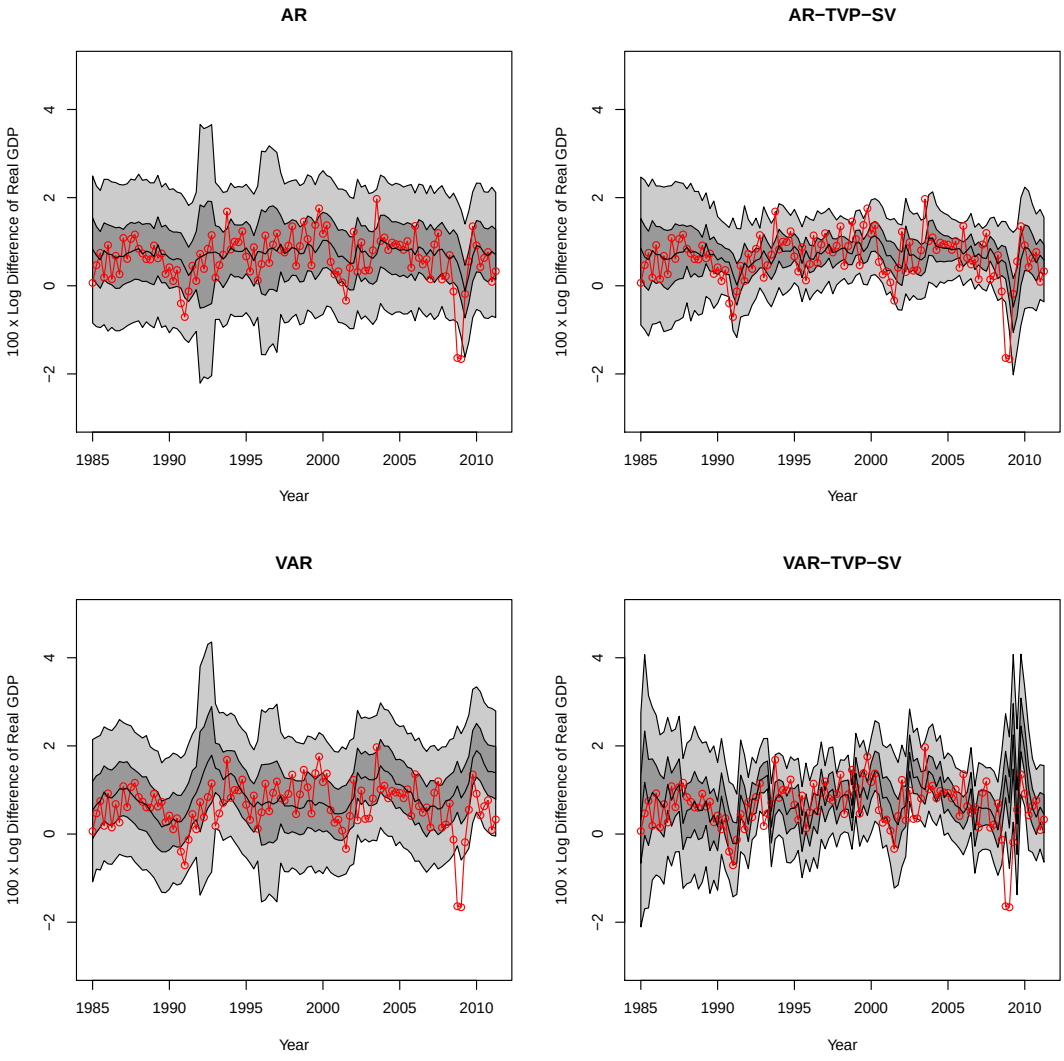


Figure 3.12: One-quarter ahead forecasts of U.S. GDP growth generated by the AR, AR-TVP-SV, VAR, and VAR-TVP-SV models. The median of the predictive distribution is shown in the black solid line, and the central 50% and 90% prediction intervals are shaded in dark and light gray, respectively. The red line shows the corresponding observations.

Table 3.5: Mean CRPS and mean DSS for probabilistic forecasts of GDP growth and inflation in the U.S. at prediction horizons of $k = 1$ and $k = 4$ quarters, respectively, for the first quarter of 1985 to the second quarter of 2011. For each variable and column, the lowest value is in bold.

	CRPS		DSS	
	$k = 1$	$k = 4$	$k = 1$	$k = 4$
GDP Growth				
AR	0.330	0.359	1.044	1.120
AR-TVP-SV	**0.292**	**0.329**	**0.833**	**1.019**
VAR	0.385	0.402	1.118	1.163
VAR-TVP-SV	0.359	0.420	0.997	1.257
Inflation				
AR	0.167	0.187	0.224	0.374
AR-TVP-SV	**0.143**	**0.156**	**0.047**	**0.175**
VAR	0.170	0.198	0.235	0.428
VAR-TVP-SV	0.162	0.201	0.179	0.552

drawn from the ith mixture component of the posterior predictive distribution (3.22), and compute the logarithmic score under the assumption of a normal predictive distribution with mean $\hat{\mu}_F$ and variance $\hat{\sigma}_F^2$. There are more efficient and theoretically principled ways of approximating the LogS in Bayesian settings which will be investigated in Chapter 4. For the moment we use the quadratic approximation based on a sample. This very nearly corresponds to replacing the LogS by the proper Dawid-Sebastiani score (DSS; Dawid and Sebastiani, 1999; Gneiting and Raftery, 2007), which for a predictive distribution F with mean μ_F and finite variance σ_F^2 is given by

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

To highlight the employed approximation method, we refer to the computed score values as Dawid-Sebastiani scores, and refer to Section 4.3 for a detailed discussion of this topic.

The quadratic approximation is infeasible for the CL and CSL scoring rules, as it then leads to improper scoring rules, see Appendix 3.C. This issue and other disadvantages of such quadratic approximations will be discussed in more detail in Chapter 4. Alternative theoretically principled approximations of the CL and CSL scoring rules are not readily available, we therefore focus on the threshold-weighted CRPS. To compute the CRPS and the threshold-weighted CRPS, we use the numerical methods proposed by Gneiting and Ranjan (2011).

The relative predictive performance of the forecasting models is consistent across the two variables and the two proper scoring rules. The AR-TVP-SV

Table 3.6: Mean restricted CRPS (rCRPS) and restricted DSS (rDSS) for probabilistic forecasts of GDP growth and inflation in the U.S. at prediction horizons of $k = 1$ and $k = 4$ quarters, respectively, for the first quarter of 1985 to the second quarter of 2011. The means are computed on instances when the observation is smaller than 0.10 (GDP) or larger than 0.98 (inflation) only. For each variable and column, the lowest value is shown in bold.

	rCRPS		rDSS	
	$k = 1$	$k = 4$	$k = 1$	$k = 4$
GDP Growth				
AR	**0.654**	**0.870**	**1.626**	**2.010**
AR-TVP-SV	0.659	0.970	2.016	3.323
VAR	0.827	0.924	2.072	2.270
VAR-TVP-SV	0.798	0.978	2.031	2.409
Inflation				
AR	0.214	0.157	0.484	**0.296**
AR-TVP-SV	0.236	0.179	0.619	0.327
VAR	**0.203**	**0.147**	**0.424**	0.317
VAR-TVP-SV	0.302	0.247	0.950	0.849

model has the best predictive performance and outperforms the baseline AR model. The p-values for the respective two-sided Diebold-Mariano tests range from 0.00 to 0.06, except for the DSS for GDP growth at a prediction horizon of $k = 4$ quarters, where the p-value is 0.37, see Appendix 3.D for details. However, the VAR models fail to outperform the simpler AR models. As we do not impose sparsity constraints on the parameters of the VAR models, this is likely due to overly complex forecasting models and overfitting, in line with results of Holzmann and Eulert (2014) and Clark and Ravazzolo (2015) in related economic and financial case studies.

To relate to the forecaster's dilemma, we restrict attention to extremes events. For GDP growth, we consider quarters with observed growth less than $r = 0.1$ only. For inflation, we restrict attention to high values in excess of $r = 0.98$. In either case, this corresponds to using about 10% of the observations. Table 3.6 shows the results of restricting the computation of the mean CRPS and the mean DSS to these observations only. For both GDP growth and inflation, the baseline AR model is considered best, and the AR-TVP-SV model appears to perform poorly. These restricted scores thus result in substantially different rankings than the proper scoring rules in Table 3.5, thereby illustrating the forecaster's dilemma. Strikingly, under the restricted assessment all four models seem less skillful at predicting inflation in the current quarter than four quarters ahead. This is a counterintuitive result that illustrates the dangers of conditioning on

Table 3.7: Mean threshold-weighted CRPS for probabilistic forecasts of GDP growth and inflation in the U.S. at prediction horizons of $k = 1$ and $k = 4$ quarters, respectively, under distinct weight functions, for the first quarter of 1985 to the second quarter of 2011. For each variable and column, the lowest value is shown in bold.

	twCRPS			
	$k = 1$	$k = 4$	$k = 1$	$k = 4$
GDP Growth	$w_{\mathrm{I}}(z) = \mathbb{1}\{z \leq 0.1\}$		$w_{\mathrm{G}} = 1 - \Phi(z \,\vert\, 0.1, 1)$	
AR	0.062	0.068	0.111	0.120
AR-TVP-SV	**0.052**	**0.062**	**0.101**	**0.115**
VAR	0.062	**0.062**	0.119	0.119
VAR-TVP-SV	0.059	0.080	0.115	0.135
Inflation	$w_{\mathrm{I}}(z) = \mathbb{1}\{z \geq 0.98\}$		$w_{\mathrm{G}} = \Phi(z \,\vert\, 0.98, 1)$	
AR	0.026	0.032	0.063	0.071
AR-TVP-SV	**0.018**	**0.018**	**0.052**	**0.056**
VAR	0.026	0.033	0.062	0.074
VAR-TVP-SV	0.022	0.037	0.060	0.077

outcomes and should be viewed as a further manifestation of the forecaster's dilemma.

In Table 3.7 we show results for the proper twCRPS under weight functions that emphasize the respective region of interest. For both variables, this yields rankings that are similar to those in Table 3.5. However, the p-values for binary comparisons with two-sided Diebold-Mariano tests generally are larger than those under the unweighted CRPS. The AR-TVP-SV model is predominantly the best, and the current quarter forecasts are deemed more skillful than those four quarters ahead. Detailed results for all pairwise comparisons are provided in Appendix 3.D.

3.5 Discussion

We have studied the dilemma that occurs when forecast evaluation is restricted to cases with extreme observations, a procedure that appears to be common practice in public discussions of forecast quality. As we have seen, under this practice even the most skillful forecasts available are bound to be discredited when the signal-to-noise ratio in the data generating process is low. Key examples might include macroeconomic and seismological predictions. In such settings it is important for forecasters, decision makers, journalists, and the general public to be aware of the forecaster's dilemma. Otherwise, charlatans might be given undue attention and recognition, and critical societal decisions could be based on

misguided predictions.

We have offered two complementary explanations of the forecaster's dilemma. From the joint distribution perspective of Section 3.2.1 stratifying by, and conditioning on, the realized value of the outcome is problematic in forecast evaluation, as theoretical guidance for the interpretation and assessment of the resulting conditional distributions is unavailable. In contrast stratifying by, and conditioning on, the forecast is unproblematic. From the perspective of proper scoring rules in 3.2.2, restricting the outcome space corresponds to the multiplication of the scoring rule by an indicator weight function, which renders any proper score improper, with an explicit hedging strategy being available.

Arguably the only remedy is to consider all available cases when evaluating predictive performance. Proper weighted scoring rules emphasize specific regions of interest and facilitate interpretation. Interestingly, however, the Neyman-Pearson lemma and our simulation studies suggest that in general the benefits of using proper weighted scoring rules in terms of power are rather limited, as compared to using standard, unweighted scoring rules. Any potential advantages vanish under weight functions with increasingly extreme threshold values, where the finite sample behavior of Diebold-Mariano tests depends on the tail properties of the forecast distributions only.

When evaluating probabilistic forecasts with emphasis on extremes, one could also consider functionals of the predictive distributions, such as the induced probability forecasts for binary tail events, as utilized in a recent comparative study by Williams et al. (2014). Another option is to consider the induced quantile forecasts, or related point summaries of the (tails of the) predictive distributions, at low or high levels, say $\alpha = 0.975$ or $\alpha = 0.99$, as is common practice in financial risk management, both for regulatory purposes and internally at financial institutions (McNeil et al., 2015). In this context, Holzmann and Eulert (2014) studied the power of Diebold-Mariano tests for quantile forecasts at extreme levels, and Fissler et al. (2016) raise the option of comparative backtests of Diebold-Mariano type in banking regulation. Ehm et al. (2016) propose decision theoretically principled, novel ways of evaluating quantile and expectile forecasts.

Variants of the forecaster's dilemma have been discussed in various strands of literature. Centuries ago, Bernoulli (1713) argued that even the most foolish prediction might attract praise when a rare event happens to materialize, referring to lyrics by Owen (1607) that are quoted in the preface of this chapter.

Tetlock (2005) investigated the quality of probability forecasts made by human experts for U.S. and world events. He observed that while forecast quality is largely independent of an expert's political views, it is strongly influenced by how a forecaster thinks. Forecasters who "know one big thing" tend to state overly extreme predictions and, therefore, tend to be outperformed by forecasters who "know many little things". Furthermore, Tetlock (2005) found an inverse relationship between the media attention received by the experts and the accuracy of their predictions, and offered psychological explanations for the attractiveness of extreme predictions for both forecasters and forecast consumers. Media attention might thus not only be centered around extreme events, but also around less

skillful forecasters with a tendency towards misguided predictions.

Denrell and Fang (2010) reported similar observations in the context of managers and entrepreneurs predicting the success of a new product. They also studied data from the Wall Street Journal Survey of Economic Forecasts, found a negative correlation between the predictive performance on a subset of cases with extreme observations and measures of general predictive performance based on all cases, and argued that accurately predicting a rare and extreme event actually is a sign of poor judgment. Their discussion was limited to point forecasts, and the suggested solution was to take into account all available observations, much in line with the findings and recommendations presented here.

Appendix 3.A Evaluation of deterministic forecasts for extreme events

Here we investigate connections to performance measures for deterministic forecasts of rare and extreme events, and explore whether these verification techniques can be extended to probabilistic forecasts.

The assessment of binary forecasts and observations of extreme events is traditionally based on 2×2 *contingency tables* which contain counts of all four possible forecast-observation pairs of events and non-events.

Table 3.8: Contingency table for the evaluation of binary forecasts of rare and extreme events.

	Event observed	Non-event observed
Event forecasted	a	b
Non-event forecasted	c	d

Table 3.8 provides a generic example. The entries a, b, c and d denote the counts of the respective forecast-observation pairs such that $a + b + c + d = n$, where n is the number of forecast cases. Extensions to deterministic forecasts and observations are straightforward as the required binary counterparts can be easily obtained by choosing thresholds that identify events.

Various measures of forecast quality based on contingency tables have been proposed, see Hogan and Mason (2012) for an extensive overview. Simple examples are the positively oriented *hit rate* (HR),

$$\mathrm{HR} = \frac{a}{a+c},$$

i.e., the frequency of correctly predicted events, and the negatively oriented *false alarm rate* (FAR),

$$\mathrm{FAR} = \frac{b}{b+d},$$

i.e., the frequency of correctly predicted non-events. However, HR and FAR are easy to hedge, e.g., an optimal HR of 1 can be obtained by always predicting an event. A variety of performance measures defined as functions of HR and FAR are less prone to such hedging strategies. However, many of these measures of forecast quality depend on the frequency of observed events and degenerate to trivial values as events become rarer (Doswell et al., 1990; Marzban, 1998). This behavior is often referred to as *base rate dependence*. In Section 3.3.4 we have observed a similar phenomenon in the evaluation of probabilistic forecasts based on mean values of the conditional likelihood scoring rule. For increasingly extreme threshold values, the mean CL scores converge to 0 as the weight function is a multiplicative component in (3.10). Therefore, the power of the corresponding Diebold-Mariano tests based on the CL score rapidly decreases for large threshold values.

Explorations of the problematic asymptotic behavior of traditional measures of forecast quality based on contingency tables have lead to the development of non-degenerating base rate independent performance measures (Stephenson et al., 2008). Ferro and Stephenson (2011) propose the *symmetric extremal dependence index* (SEDI)

$$\text{SEDI} = -\frac{\log(\text{FAR}) - \log(\text{HR}) - \log(1-\text{FAR}) + \log(1-\text{HR})}{\log(\text{FAR}) + \log(\text{HR}) + \log(1-\text{FAR}) + \log(1-\text{HR})} \tag{3.26}$$

which we define in negative orientation. The SEDI does not degenerate to trivial values for vanishing base rates, is not straightforward to hedge, and satisfies further desirable properties, see Ferro and Stephenson (2011). Recent applications of the SEDI include assessments of the quality of operational weather forecasts by the European Centre for Medium-Range Weather Forecasts (ECMWF), see Haiden et al. (2014) and Magnusson et al. (2014).

However, as argued by Ferro and Stephenson (2011), the SEDI should only be applied to *calibrated* forecasts to guarantee convergence to a meaningful limit for rare events. In the context of binary or deterministic forecasts and observations which are evaluated based on contingency tables forecasts are called calibrated if the number of predicted events, $a + b$, equals the number of observed events, $a + c$. Note that this concept of calibration is not directly related to the notions of calibration of probabilistic forecasts introduced in Section 3.2.1.[5]

Typically, forecasts thus have to be re-calibrated by adjusting the event-defining thresholds to yield a calibrated forecasting system and allow for applying the SEDI. Different re-calibration schemes are available. In the simple *quantile re-calibration* approach (Casati et al., 2004), the event-defining thresholds for forecasts and observations are selected as the same fixed quantiles of the set of all forecasts and observations, respectively. For example, if events are defined as observations in the upper 10th percentile of all observed values, all forecasts above the 90th percentile of all predictions are considered as forecasts of events. By contrast, the *numerical optimization-based re-calibration* approach (Ferro and Stephenson, 2011) determines the event-defining threshold for the predictions by minimizing the absolute difference $|b - c|$ of the entries of the corresponding contingency tables based on a fixed event-defining threshold for the observations.

From a theoretical perspective the required re-calibration implies that the SEDI assesses potential rather than actual skill of a forecasting system. Further, rankings of forecasters based on the SEDI obtained with different re-calibration schemes might disagree, and re-calibrating the forecasts may be impossible or undesirable in practical applications (Ferro and Stephenson, 2011).

The SEDI and other performance measures based on contingency tables are not straightforward to generalize towards probabilistic forecasts in the form of full distributions over future quantities or events. In order to apply these measures

[5]This overlap in notation in the literature is unfortunate, however, which definition of calibration applies is typically immediate from the context. In particular, within this thesis the notion of calibration based on contingency tables will be exclusively used in the present Appendix 3.A at hand.

of forecast quality, the probabilistic forecasts first have to be transformed into binary forecasts. Because of the required re-calibration, this can be achieved by first obtaining deterministic forecast, e.g., by computing functionals of the predictive distributions such as mean or median values. Clearly, any uncertainty information contained in the probabilistic forecast will be lost in this step, and in the light of Section 2.4 it is not obvious which functional should be chosen such that the SEDI is a consistent scoring function after the subsequent re-calibration step.

The notions of propriety of scoring rules and consistency of scoring functions do not immediately apply to performance measures based on contingency tables. Instead of computing a single real-valued score from a single forecast-observation pair, the measure of forecast quality here depends on the joint distribution of all forecasts and observations. Judging the adequacy of performance measures such as the SEDI from a decision theoretical perspective would require the development of a new, generalized notion of propriety, see, e.g., Byrne (2016) for recent seminal work in this direction.

We illustrate the above considerations in a case study on wind speed forecasts. Modern weather forecasts are typically given in the form of *ensemble predictions* which are obtained by multiple runs of numerical weather prediction models with varying initial conditions and/or model physics. Ensemble forecasts are generally biased and lack calibration (in the usual sense defined in Section 2.2), they thus require some form of statistical postprocessing. Ensemble forecasts and post-processing techniques will be discussed in detail in Chapter 5. Here, we focus on forecasts and observations of wind speed at an observation station located at Frankfurt airport, Germany, from 2002–2014. The forecasts are 3-day ahead predictions of the global 50-member ensemble of the ECMWF. For a detailed description of the data, see Hemri et al. (2014). Postprocessed forecasts are obtained in a non-homogeneous regression approach based on truncated normal (TN) distributions. This approach was proposed by Thorarinsdottir and Gneiting (2010) will be introduced and discussed in detail in Chapter 5. The parameters of the predictive distributions are estimated by minimizing the mean CRPS over a rolling training period consisting of the preceding 365 days. Deterministic point forecasts of wind speed are obtained as median values of the 50 ensemble members and the TN forecast distributions. Results from the extensive literature on post-processing of ensemble forecasts suggest that a superior predictive performance of the postprocessed forecasts compared to the raw ensemble predictions is to be expected.

Figure 3.13 shows the SEDI values obtained with different re-calibration approaches as functions of the threshold above which an observation is considered an (extreme) event, given in terms of quantiles of the distribution of wind speed observations at Frankfurt airport. It can be observed that the SEDI values do not degenerate to a trivial value as the threshold value increases and the base rate decreases. However, the ranking of the competing forecasts obtained by the two re-calibration approaches is generally not consistent and differs for various ranges of threshold values.

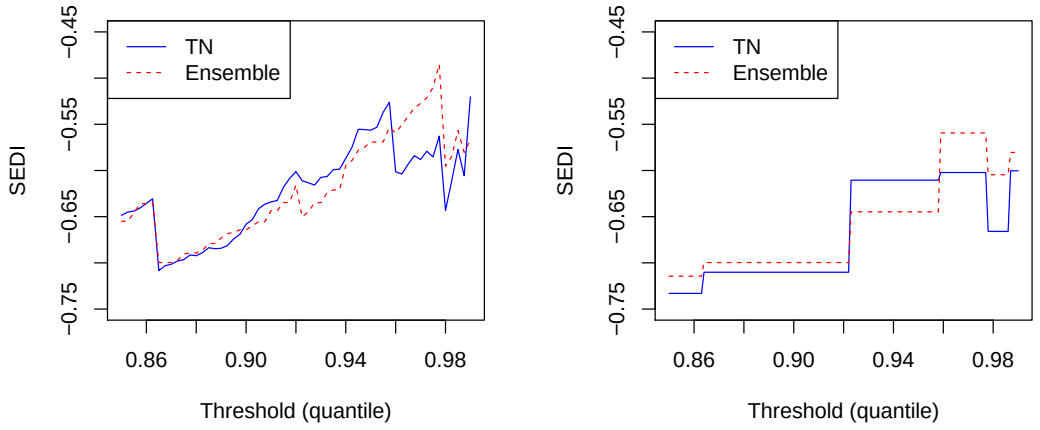


Figure 3.13: SEDI as function of the threshold above which observations are considered an event in terms of quantiles of the distribution of wind speed observations. Deterministic forecasts are obtained as median value of the ensemble (red dashed line) and the postprocessed forecasts following a truncated normal distribution (blue solid line). In the left panel the corresponding thresholds for the forecasts are obtained by quantile re-calibration, in the right panel they are obtained by numerical minimization of $|b - c|$.

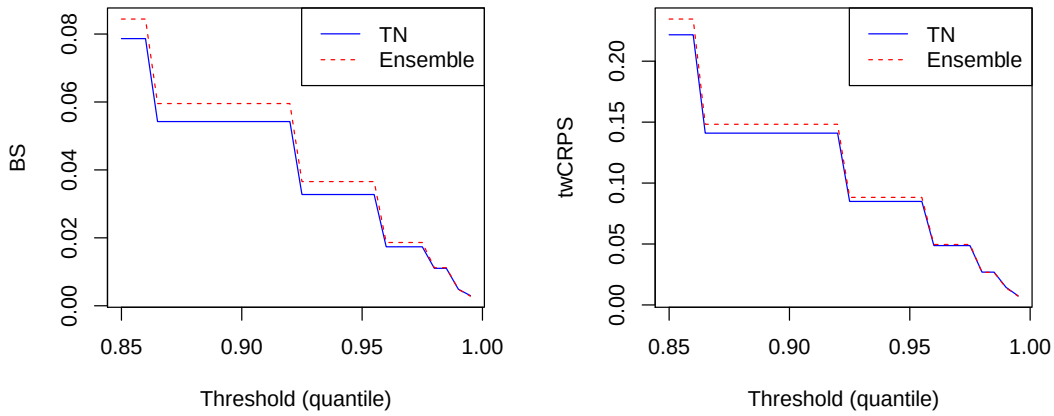


Figure 3.14: Mean Brier score (left) and threshold-weighted CRPS (right) as functions of the threshold value r in the definition of the Brier score and the indicator weight function, given in terms of quantiles of the distribution of wind speed observations at Frankfurt airport.

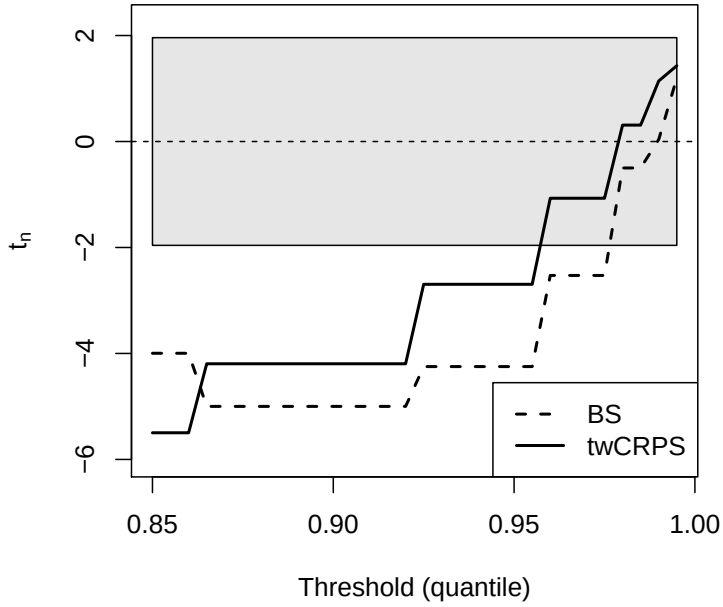


Figure 3.15: Test statistic t_n of two-sided Diebold-Mariano tests of equal predictive performance of the raw ECMWF ensemble predictions and the postprocessed TN forecasts based on the Brier score and the twCRPS as functions of the threshold value in terms of quantiles of the distribution of wind speed observations. Negative values indicate a superior predictive performance of the postprocessed TN forecasts. Values outside of the shaded area are significant under the null hypothesis of equal predictive performance at the 5% level.

Returning to the corresponding probabilistic forecasts, Figure 3.14 shows mean values of the Brier score

$$\mathrm{BS}_r(F, y) = (F(r) - \mathbb{1}\{y \leq r\})^2 = ((1 - F(r)) - \mathbb{1}\{y > r\})^2$$

of the induced binary probability forecasts, and of the threshold-weighted CRPS with an indicator weight function $w_r(z) = \mathbb{1}\{y \geq r\}$ as functions of the threshold value r in terms of quantiles of the distribution of wind speed observations. It can be observed that the rankings of the two competing forecasts obtained by the two proper scoring rules coincide for almost all threshold values. By contrast to the SEDI values shown in Figure 3.13, the postprocessed TN forecasts are generally preferred over the ensemble forecasts, except for very high threshold values.

A further difference is that both the mean BS_r and the mean twCRPS converge to 0 for increasingly extreme threshold values and thus share a deficiency of base rate dependent performance based on contingency tables. However, if the aim is a comparative assessment of predictive performance this issue can be alleviated by employing Diebold-Mariano tests. Figure 3.15 shows the test statistic (3.15) of two-sided Diebold-Mariano tests of equal predictive performance of the raw ensemble and the postprocessed TN forecasts based on the BS_r and the

twCRPS as functions of the threshold value r given in terms of quantiles of the distribution of wind speed observations at Frankfurt airport. The observed score differences are significant at the 5% level up to a threshold value around the 95th percentile of the observations. Figure 3.15 shows that the raw ensemble predictions are preferred over the postprocessed forecasts for very high threshold values. As the predictive distributions of ensemble forecasts are finite, this potentially constitutes an example of the discussed problematic finite-sample dependence of weighted proper scoring rules on the tail behavior of the forecast distributions, see also Appendix 3.B.

To conclude, we note that the SEDI and related performance measures based on contingency tables can provide valuable tools for evaluating binary and deterministic forecasts, however, the required re-calibration might lead to undesired results, and the generalization to probabilistic forecasts proves difficult. Instead, the above discussion of proper weighted scoring rules in Chapter 3 provides at least a partial answer to the call for probabilistic forecast verification for extreme events by Haiden et al. (2014, p. 33) and Magnusson et al. (2014, p. 26).

Appendix 3.B Tail dependence of proper weighted scoring rules

Here we provide analytical results which illustrate the problematic finite sample dependence of proper weighted scoring rules on the tail behavior of the forecast distributions in case of extreme threshold values. Consider an indicator weight function $w_r(z) = \mathbb{1}\{z \geq r\}$ and a threshold r that exceeds the maximum of the given sample $y_1, \ldots, y_n$, i.e., $y_i < r$ for all $i = 1, \ldots, n$. The values of all discussed proper weighted scoring rules then do not depend on the observations, and are solely determined by the respective tail probabilities.

The conditional likelihood scoring rule is 0 for all observations $y_1, \ldots, y_n$ and any forecast distribution as the weight function is a multiplicative component in (3.10). The null hypothesis of equal predictive performance of any two forecasts can thus not be rejected which causes the decrease in power of the corresponding Diebold-Mariano tests observed in Figures 3.9 and 3.10.

In the above situation, the censored likelihood scoring rule (3.11) reduces to the corresponding discrete logarithmic score at threshold r, i.e., $LS_r(F, y) = -\log F(r)$. The average score difference of two arbitrary forecast distributions F_1 and F_2 is therefore given by

$$\overline{\mathrm{CSL}}_n^{F_1} - \overline{\mathrm{CSL}}_n^{F_2} = \log F_2(r) - \log F_1(r),$$

and thus favors the forecast distribution with the lighter tails, e.g., a negative score difference favoring F_1 is obtained if $F_1(r) < F_2(r)$.

In case of the threshold-weighted CRPS (3.12), the average score difference reduces to

$$\overline{\mathrm{twCRPS}}_n^{F_1} - \overline{\mathrm{twCRPS}}_n^{F_2} = \int_r^\infty (F_1(z) - 1)^2 - (F_2(z) - 1)^2 \, \mathrm{d}z.$$

Clearly, the forecast with the lighter tail again receives the better score, irrespectively of the true distribution.

Appendix 3.C Impropriety of quadratic approximations

Here we show that naive quadratic approximations of the CL and CSL scoring rules result are improper. Let F be a predictive distribution with mean μ_F and standard deviation σ_F. As regards the conditional likelihood (CL) score (3.10), the quadratic approximation is given by

$$\mathrm{CL}^q(F, y) = -w(y) \log\left(\frac{\phi(y|F)}{\int w(x)\phi(x|F)\,\mathrm{d}x} \right),$$

where $\phi(\cdot|F)$ denotes a normal density with mean μ_F and standard deviation σ_F, respectively. Let

$$c_F = \int w(x)\phi(x|F)\,\mathrm{d}x, \qquad c_G = \int w(x)\phi(x|G)\,\mathrm{d}x, \qquad c_g = \int w(x)g(x)\,\mathrm{d}x,$$

and recall that the Kullback-Leibler divergence between two probability densities u and v is given by

$$\mathrm{KL}(u, v) = \int u(x) \log\left(\frac{u(x)}{v(x)} \right) \mathrm{d}x.$$

Assuming that CL^q is proper, it is true that

$$\begin{aligned} &\mathbb{E}_G(\mathrm{CL}^q(F, Y) - \mathrm{CL}^q(G, Y)) \\ &\quad = c_g \left[\mathrm{KL}\left(\frac{w(y)g(y)}{c_g}, \frac{w(y)\phi(y|F)}{c_F} \right) - \mathrm{KL}\left(\frac{w(y)g(y)}{c_g}, \frac{w(y)\phi(y|G)}{c_G} \right) \right] \end{aligned}$$

is non-negative. Let G be uniform on $[-\sqrt{3}, \sqrt{3}]$ so that $\mu_G = 0$ and $\sigma_G = 1$, and let $w(y) = \mathbb{1}\{y \geq 1\}$. Denoting the cumulative distribution function of the standard normal distribution by Φ, we find that

$$\begin{aligned} &\mathrm{KL}\left(\frac{w(y)g(y)}{c_g}, \frac{w(y)\phi(y|F)}{c_F} \right) - \mathrm{KL}\left(\frac{w(y)g(y)}{c_g}, \frac{w(y)\phi(y|G)}{c_G} \right) \\ &\quad = \log\left(\sigma_F \frac{1 - \Phi((1 - \mu_F)/\sigma_F)}{1 - \Phi(1)} \right) + \frac{3(\sqrt{3} - 1)\mu_F^2 - 6\mu_F + (3\sqrt{3} - 1)(1 - \sigma_F^2)}{6(\sqrt{3} - 1)\sigma_F^2}, \end{aligned}$$

which is strictly negative in a neighborhood of $\mu_F = 1.314$ and $\sigma_F = 0.252$, for the desired contradiction. Therefore, CL^q is not a proper scoring rule.

As regards the censored likelihood (CSL) score (3.11), the quadratic approximation is

$$\mathrm{CSL}^q(F, y) = -w(y) \log(\phi(y|F)) - (1 - w(y)) \log\left(1 - \int w(z)\phi(z|F)\,\mathrm{d}z \right).$$

Under the same choice of w, F, and G as before, we find that

$$\begin{aligned}
&\mathbb{E}_G(\mathrm{CSL}^q(F,Y) - \mathrm{CSL}^q(G,Y)) \\
&\quad = \frac{\sqrt{3}-1}{2\sqrt{3}} \log \sigma_F - \frac{\sqrt{3}+1}{2\sqrt{3}} \log\left(\frac{\Phi((1-\mu_F)/\sigma_F)}{\Phi(1)}\right) \\
&\qquad + \frac{3(\sqrt{3}-1)\mu_F^2 - 6\mu_F + (3\sqrt{3}-1)(1-\sigma_F^2)}{12\sqrt{3}\,\sigma_F^2},
\end{aligned}$$

which is strictly negative in a neighborhood of $\mu_F = 0.540$ and $\sigma_F = 0.589$. Therefore, CSL^q is not a proper scoring rule.

Appendix 3.D Detailed results for the case study

In order to investigate the predictive performance of the various models of GDP growth and inflation in more detail we perform pairwise two-sided Diebold-Mariano tests (3.15) based on the CRPS, DSS and twCRPS. Tables 3.9–3.12 show the corresponding values of the test statistics t_n for all possible binary comparisons, rounded to 3 digits. The asymptotic variance is estimated following Gneiting and Ranjan (2011), see equation (3.16). In any of these comparisons, the forecast of the model in the leftmost column takes the role of F in (3.15), and the model in one of the four columns on the right-hand side takes the role of G. Negative values of t_n thus indicate a superior predictive performance of the model in the left-most column whereas the model on the right-hand side produced better forecasts in case of positive values of t_n. Clearly, the tables are symmetric except for the reversed sign. All results discussed below are in line with the observations of Clark and Ravazzolo (2015).

Generally, the AR-TVP-SV model consistently performs better than all competitors for both variables and at both forecast horizons. The only exception is the comparison of 4 quarter ahead forecasts of GPD growth with those of the VAR model in terms of the two variants of the twCRPS. Further, the respective comparisons of the AR-TVP-SV model and all competitors show the largest portion of significant score difference in terms of all employed proper scoring rules. In particular, it always performs better than the simple baseline AR model. By contrast, the VAR-TVP-SV model frequently fails to outperform the simple baseline VAR model.

Turning now to a comparison of the autoregressive and vector autoregressive versions of the baseline and TVP-SV models, it can be observed that the AR models are typically preferred over their vector-autoregressive counterparts. As discussed in Section 3.4.3 this might be caused by a lack of sparsity constraints on the parameters of the VAR models leading to overfitting.

In general, the observed score differences between the models tend to be more significant for $k = 1$ quarter ahead forecasts than for forecast horizons of $k = 4$ quarters ahead. An exception is given by the forecasts for inflation evaluated by the DSS. Comparing the employed proper scoring rules from a broader perspective

Table 3.9: Values of the test statistic t_n for all binary comparisons via two-sided Diebold-Mariano tests of equal predictive performance based on the CRPS in the case study presented in Section 3.4.3. Negative values of t_n indicate a superior predictive performance of the model in the leftmost column, in case of positive values the model in the respective column on the right-hand side is preferred. Values of t_n that are significant at the 5% level are marked in bold.

	AR	AR-TVP-SV	VAR	VAR-TVP-SV
GDP growth, $k = 1$				
AR		**4.337**	**−3.126**	−1.272
AR-TVP-SV	**−4.337**		**−4.309**	**−2.914**
VAR	**3.126**	**4.309**		1.213
VAR-TVP-SV	1.272	**2.914**	−1.213	
GDP growth, $k = 4$				
AR		1.916	−1.069	−1.751
AR-TVP-SV	−1.916		−1.634	**−2.500**
VAR	1.069	1.634		−0.787
VAR-TVP-SV	1.751	**2.500**	0.787	
Inflation, $k = 1$				
AR		**3.155**	−0.880	0.501
AR-TVP-SV	**−3.155**		**−3.402**	**−2.764**
VAR	0.880	**3.402**		0.873
VAR-TVP-SV	−0.501	**2.764**	−0.873	
Inflation, $k = 4$				
AR		1.914	−1.695	−0.685
AR-TVP-SV	−1.914		**−2.319**	**−2.672**
VAR	1.695	**2.319**		−0.165
VAR-TVP-SV	0.685	**2.672**	0.165	

Table 3.10: Values of the test statistic t_n for all binary comparisons via two-sided Diebold-Mariano tests of equal predictive performance based on the DSS in the case study presented in Section 3.4.3. Negative values of t_n indicate a superior predictive performance of the model in the leftmost column, in case of positive values the model in the respective column on the right-hand side is preferred. Values of t_n that are significant at the 5% level are marked in bold.

	AR	AR-TVP-SV	VAR	VAR-TVP-SV
GDP growth, $k = 1$				
AR		**3.854**	**−2.101**	0.767
AR-TVP-SV	**−3.854**		**−4.896**	**−3.284**
VAR	**2.101**	**4.896**		**2.338**
VAR-TVP-SV	−0.767	**3.284**	**−2.338**	
GDP growth, $k = 4$				
AR		0.893	−0.633	−1.799
AR-TVP-SV	−0.893		−1.293	**−2.294**
VAR	0.633	1.293		−1.467
VAR-TVP-SV	1.799	**2.294**	1.467	
Inflation, $k = 1$				
AR		**3.084**	−0.514	0.696
AR-TVP-SV	**−3.084**		**−3.216**	**−2.627**
VAR	0.514	**3.216**		0.894
VAR-TVP-SV	−0.696	**2.627**	−0.894	
Inflation, $k = 4$				
AR		**2.188**	**−2.068**	−1.875
AR-TVP-SV	**−2.188**		**−2.558**	**−3.396**
VAR	**2.068**	**2.558**		−1.480
VAR-TVP-SV	1.875	**3.396**	1.480	

Table 3.11: Values of the test statistic t_n for all binary comparisons via two-sided Diebold-Mariano tests of equal predictive performance based on the the twCRPS with an indicator weight function $w_I(z)$ in the case study presented in Section 3.4.3. In the case of GDP growth, $w_I(z) = \mathbb{1}\{z \leq 0.1\}$, and for inflation, $w_I(z) = \mathbb{1}\{z \geq 0.98\}$. Negative values of t_n indicate a superior predictive performance of the model in the leftmost column, in case of positive values the model in the respective column on the right-hand side is preferred. Values of t_n that are significant at the 5% level are marked in bold.

	AR	AR-TVP-SV	VAR	VAR-TVP-SV
GDP growth, $k = 1$				
AR		**2.926**	−0.227	0.542
AR-TVP-SV	**−2.926**		**−2.337**	−1.451
VAR	0.227	**2.337**		1.027
VAR-TVP-SV	−0.542	1.451	−1.027	
GDP growth, $k = 4$				
AR		0.922	1.249	−1.363
AR-TVP-SV	−0.922		0.015	−1.876
VAR	−1.249	−0.015		**−2.214**
VAR-TVP-SV	1.363	1.876	**2.214**	
Inflation, $k = 1$				
AR		**2.949**	0.846	1.617
AR-TVP-SV	**−2.949**		**−3.395**	**−3.187**
VAR	−0.846	**3.395**		1.743
VAR-TVP-SV	−1.617	**3.187**	−1.743	
Inflation, $k = 4$				
AR		1.748	−0.385	−0.522
AR-TVP-SV	−1.748		**−2.390**	**−2.518**
VAR	0.385	**2.390**		−0.465
VAR-TVP-SV	0.522	**2.518**	0.465	

Table 3.12: Values of the test statistic t_n for all binary comparisons via two-sided Diebold-Mariano tests of equal predictive performance based on the the twCRPS with a Gaussian weight function $w_G(z)$ in the case study presented in Section 3.4.3. In the case of GDP growth, $w_G(z) = 1 - \Phi(z|0.1, 1)$, and for inflation, $w_G(z) = \Phi(z|0.98, 1)$. Negative values of t_n indicate a superior predictive performance of the model in the leftmost column, in case of positive values the model in the respective column on the right-hand side is preferred. Values of t_n that are significant at the 5% level are marked in bold.

	AR	AR-TVP-SV	VAR	VAR-TVP-SV
GDP growth, $k = 1$				
AR		**3.311**	−1.504	−0.657
AR-TVP-SV	**−3.311**		**−3.114**	**−2.202**
VAR	1.504	**3.114**		0.653
VAR-TVP-SV	0.657	**2.202**	−0.653	
GDP growth, $k = 4$				
AR		0.831	0.241	−1.924
AR-TVP-SV	−0.831		0.473	**−2.381**
VAR	−0.241	−0.473		−1.944
VAR-TVP-SV	1.924	**2.381**	1.944	
Inflation, $k = 1$				
AR		**3.084**	0.498	0.799
AR-TVP-SV	**−3.084**		**−3.339**	**−3.187**
VAR	−0.498	**3.339**		0.725
VAR-TVP-SV	−0.799	**3.187**	−0.725	
Inflation, $k = 4$				
AR		1.907	−0.693	−0.539
AR-TVP-SV	−1.907		**−2.366**	**−2.586**
VAR	0.693	**2.366**		−0.309
VAR-TVP-SV	0.539	**2.586**	0.309	

it can be observed that any binary comparison with significant CRPS differences also shows significant differences in terms of the DSS. On the other hand, some binary comparisons show significant differences in terms of the DSS, but not in terms of the CRPS, in particular for $k = 4$ quarter ahead forecasts of inflation.

The unweighted and threshold-weighted CRPS generally prefer the same models, however, the corresponding p-values for the binary comparisons tend to be slightly larger if the Diebold-Mariano tests are based on the threshold-weighted CRPS, in particular for inflation. Tables 3.11 and 3.12 indicate that the choice of the weight function in the threshold-weighted CRPS only has a negligible effect on the sign and magnitude of the corresponding observed score differences.

4 Probabilistic forecasting and comparative model assessment based on MCMC output

> Although this may seem a paradox, all exact science is dominated by the idea of approximation.[1]
>
> Bertrand Russell, 1931

In the preceding considerations, we have typically assumed the forecast distribution to be known explicitly in analytical form. However, in many applications such as the use of Bayesian methods for probabilistic forecasting, the predictive distribution comes as a simulated sample. In this chapter, we conduct a systematic analysis of how to make and evaluate probabilistic forecasts based on the output of Markov chain Monte Carlo (MCMC) algorithms. Utilizing the mathematical framework provided by the theory of proper scoring rules, we develop a notion of consistency that allows for assessing the adequacy of methods for estimating the stationary distribution underlying the MCMC output. We then review asymptotic results that account for the salient features of Bayesian posterior simulators, and derive conditions under which choices from the literature satisfy this notion of consistency. As we will see below, these conditions depend on the scoring rule being used, such that the choices of approximation method and scoring rule are intertwined.

4.1 Introduction

A rapidly growing literature uses Bayesian methods to produce probabilistic forecasts in a wide range of applications including meteorological, economic, and financial problems. Thereby, the posterior predictive distribution of interest takes the form of a simulated sample, typically generated by an MCMC algorithm. Following Rubin (1984), Little (2006), and others, it now seems widely accepted that the simulated sample should be evaluated using frequentist principles, i.e., without prior information entering the model evaluation stage. We analyze this topic, focusing on comparative assessments of two or more models via proper scoring rules.

[1] Russell (2001, p. 45), first edition published 1931.

Table 4.1: Summary of approximation methods (rows) and proper scoring rules (columns) used in recently published studies using Bayesian probabilistic forecasts based on MCMC output. Each cell indicates the number of studies, and lists an illustrative reference. See Appendix 4.A for the list of all 39 included studies and details on the literature review. The approximation methods are defined and discussed in detail in Section 4.3.

	LogS	CRPS
Mixture of parameters	$n = 13$ Zhou et al. (2015)	$n = 3$ Kallache et al. (2010)
Kernel density estimation	$n = 4$ Carriero et al. (2015c)	$n = 1$ Krüger and Nolte (2015)
Gaussian approximation	$n = 6$ Clark (2011)	$n = 2$ Rodrigues et al. (2014)
Empirical CDF	not applicable	$n = 2$ Smith and Vahey (2015)
Kernel representation	not applicable	$n = 13$ Sigrist et al. (2015)

In Bayesian statistics, a model's posterior predictive distribution is of the generic form

$$F_0(y) = \int_\Theta F_c(y|\theta)\,\mathrm{d}P_{post}(\theta), \tag{4.1}$$

where $P_{post}(\theta)$ is the posterior distribution of a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^d$, and $F_c(y|\theta)$ is the predictive distribution conditional on θ. Since the posterior predictive distribution in (4.1) is typically not known in closed form, it must be estimated in some way. One approach, which we will call the *mixture-of-parameters estimator*, is to draw a sequence $\{\theta_i\}_{i=1}^m$ from P_{post}, and set $\hat{F}_m(y) = \frac{1}{m}\sum_{i=1}^m F_c(y|\theta_i)$. An alternative route is to produce draws $\{X_i\}_{i=1}^m$, where $X_i \sim F_c(\cdot|\theta_i)$, and estimate F_0 from these draws, either parametrically or nonparametrically.

Table 4.1 summarizes 39 recently published studies on probabilistic forecasting using MCMC methods. Information on the literature survey methodology and the full list of all 39 references are provided in Appendix 4.A. The studies use five different ways to approximate F_0 in equation (4.1). Furthermore, they consider different scoring rules to evaluate the performance of the probabilistic forecasts, most prominently the LogS and CRPS. We will argue that the choices of approximation method and scoring rule are intertwined, and should be made jointly.

As the table demonstrates, several combinations of approximation method and

scoring rules are used in current practice. To date there are few guidelines to support any of these choices, and it is not clear how they affect the outcome of the model comparison. The present chapter provides a systematic analysis of this topic. We focus on the following questions. First, what defines "reasonable" choices of approximation method and scoring rule? Second, under what conditions do extant choices from the literature satisfy this definition? Third, for a given scoring rule, how accurate are alternative approximation methods in practically relevant scenarios?

The answer to the first question is necessarily subjective. Our aim here is to propose a definition which relates to traditional asymptotic statistical concepts. For a given scoring rule, we hence introduce the concept of a *consistent* approximation method. This definition formalizes the idea that, as the size of the simulated sample becomes infinite, the approximation should perform equally well as the unknown true forecast distribution. Thereby, performance is measured in terms of proper scoring rules. In order to tackle the second question, we provide a succinct overview of asymptotic results which are appropriate in order to evaluate methodological choices like the ones in Table 4.1. The results we survey account for the salient features of Bayesian posterior simulators, in particular time series type dependence among the MCMC draws (see, e.g., Geweke, 2005, Section 4). Regarding the third question, we provide a simulation study which is motivated by practical applications of MCMC. Furthermore, we study the behavior of various approximation methods and scoring rules in a case study of a popular model for economic time series.

On the whole, our analysis suggests the use of a *mixture-of-parameters* estimator in order to approximate the posterior predictive distribution of interest. The theoretical legitimacy of this estimator (based on our notion of a consistent approximation) can be derived with results from empirical process theory, see Section 4.3. In our simulation and empirical analysis, the estimator outperforms its competitors without exception. This finding can be explained by the fact that it efficiently exploits the parametric structure of the Bayesian model. By contrast, other approaches either impose restrictive additional assumptions, thus leading to bias, or fail to exploit what is known about the structure of the model, thus leading to unnecessary sampling variability.

A further implication of our analysis is that the choices of approximation method and scoring rule are heavily intertwined. Under the popular logarithmic score, which is sensitive to tail events, fully nonparametric kernel density estimation techniques are problematic. For other scoring rules such as the CRPS, which are less sensitive, such approaches perform similarly well as the mixture-of-parameters estimator mentioned above.

The remainder of this chapter is organized as follows. Section 4.2 introduces the notion of a consistent approximation method. Section 4.3 surveys theoretical justifications of various approximation methods encountered in the literature. Sections 4.4 and 4.5 present simulation and empirical evidence on the relative performance of these methods, and Section 4.6 concludes with a discussion. The chapter is based on Krüger et al. (2016).

4.2 Formal setup

In this section, we formalize the posterior predictive distribution in Bayesian forecasting, and introduce the concept of a consistent approximation method based on MCMC output.

4.2.1 Posterior predictive distribution

As discussed earlier, the posterior predictive distribution of a Bayesian forecasting model is given by

$$F_0(y) = \int_\Theta F_c(y|\theta)\,\mathrm{d}P_{post}(\theta),$$

where $\theta \in \Theta \subseteq \mathbb{R}^d$ is the model's parameter vector, $P_{post}(\theta)$ is the posterior distribution of the parameters (see below), and $F_c(z|\theta)$ is the predictive distribution *conditional* on a particular vector θ, see, e.g., Greenberg (2013, p. 33) or Gelman et al. (2014a, p. 7). MCMC algorithms designed to sample from F_0 can be sketched as follows:

- Fix $\theta_0 \in \Theta$ at some arbitrary value.
- For iteration $i = 1, \ldots, m$:
  - Draw $\theta_i \sim \mathcal{K}(\theta_i|\theta_{i-1})$, where $\mathcal{K}$ is a transition kernel that specifies the conditional distribution of θ_i given θ_{i-1}.
  - Draw $X_i \sim F_c(\cdot|\theta_i)$.[2]

Throughout this chapter, we assume that the transition kernel $\mathcal{K}$ is such that the sequence $\{\theta_i\}_{i=1}^m$ is ergodic, with stationary distribution $P_{post}(\cdot)$. These assumptions can be expected to hold widely in applications (see, e.g., Craiu and Rosenthal, 2014, Sections 8.1 and 8.2).

The MCMC algorithm sketched above leaves two options for estimating the posterior predictive distribution F_0 in equation (4.1),

- Option 1: output $\{\theta_i\}_{i=1}^m$
- Option 2: output $\{X_i\}_{i=1}^m$.

In the case of Option 1, the sequence $\{\theta_i\}_{i=1}^m$, together with $F_c(\cdot|\theta)$, yields a natural estimator of the posterior predictive distribution F_0,

$$\hat{F}_m(z) = \frac{1}{m}\sum_{i=1}^{m} F_c(z|\theta_i), \tag{4.2}$$

[2] Alternatively, studies such as Möller et al. (2015) and Krüger et al. (2015) use $X_{ij} \sim F_c(\cdot|\theta_i)$, where $j = 1, \ldots, J$. The number J is sometimes called the *oversampling factor*, representing the number of forecast draws for each parameter draw.

see, e.g., Gschlößl and Czado (2007, Section 3.2), Hooten and Hobbs (2015, eq. 17), and Gelman et al. (2014b, eq. 5). We provide a detailed analysis of this *mixture-of-parameters* estimator in Section 4.3.1.

Alternatively, many authors consider a sample $\{X_i\}_{i=1}^m$ from the posterior predictive distribution (Option 2). All entries in Table 4.1 except for the mixture-of-parameters estimator in the first row follow this approach. Importantly, ergodicity of $\{\theta_i\}_{i=1}^m$ implies that $\{X_i\}_{i=1}^m$ is ergodic as well (Geweke, 2005, Theorem 4.5.2). To see that X_i converges to its stationary distribution F_0 for m large enough, note that for any $z \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}(X_i \leq z) &= \frac{1}{m}\sum_{i=1}^{m} F_c(z|\theta_i) \\ &\overset{a.s.}{\to} \int_\Theta F_c(z|\theta)\ dP_{post}(\theta) \\ &= F_0(z)\end{aligned}$$

Strictly speaking, Option 2 often implies "more randomness than necessary", in that the simulation step used to draw X_i can be avoided by using Option 1 above. On the positive side, the sequence $\{X_i\}_{i=1}^m$ can be used without specific knowledge of the statistical model that generated it, i.e., without knowledge of $F_c(\cdot|\theta)$. This may simplify the communication of the results if the forecaster and the forecast user are not the same person. The random sample $X_1, \ldots, X_m \sim F_0$ can be used as an input for an approximation procedure $A(X_1, \ldots, X_m)$ which generates an alternative estimate of F_0,

$$\hat{F}_m = A(X_1, \ldots, X_m).$$

Depending on the employed approximation procedure, the output can be an estimate of the predictive CDF F_0, or of the corresponding predictive density f_0. Examples for approximation procedures are kernel density estimation, the empirical cumulative distribution function, or the use of a fixed parametric family with parameters estimated from $X_1, \ldots, X_m$.

4.2.2 Proper scoring rules and score divergences

Here, we briefly review relevant theory that has been introduced in detail in Chapter 2. Recall that a scoring rule $S : \mathcal{F} \times \Omega_Y \longrightarrow \mathbb{R} \cup \{\infty\}$ is proper if

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

for all probability distributions $F, G \in \mathcal{F}$. We typically set $\Omega_Y = \mathbb{R}$, but will occasionally restrict our attention to compact subsets of $\mathbb{R}$. With the expected score of a probabilistic forecast F under the true distribution G denoted by

$$\mathcal{S}(F, G) = \mathbb{E}_{Y\sim G} S(F, Y),$$

Table 4.2: Examples of popular scoring rules and associated score divergences. See Section 2.3.1 for explanations of the abbreviations and details on classes $\mathcal{F}$ of probabilistic forecasts relative to which the scoring rules are proper.

Scoring rule	$S(F, y)$	$d_S(F, G)$
CRPS	$\int_{\mathbb{R}} (F(z) - \mathbb{1}\{y \leq z\})^2 \, \mathrm{d}z$	$\int_{\mathbb{R}} (F(z) - G(z))^2 \, \mathrm{d}z$
LogS	$-\log f(y)$	$\int_{\mathbb{R}} g(z) \, \log\left(\frac{g(z)}{f(z)}\right) \, \mathrm{d}z$
QS	$\|f\|_2^2 - 2f(y)$	$\int_{\mathbb{R}} (f(z) - g(z))^2 \, \mathrm{d}z$
DSS	$\log \sigma_F^2 + \frac{(y-\mu_F)^2}{\sigma_F^2}$	$\frac{\sigma_G^2}{\sigma_F^2} - \log \frac{\sigma_G^2}{\sigma_F^2} - \frac{(\mu_F - \mu_G)^2}{\sigma_F^2} - 1$
HS	$2 \, \frac{f''(y)}{f(y)} - \left(\frac{f'(y)}{f(y)}\right)^2$	$\int_{\mathbb{R}} \left(\frac{g'(z)}{g(z)} - \frac{f'(z)}{f(z)}\right)^2 g(z) dz$

we have introduced the *score divergence* associated with the scoring rule S given by

$$d_S(F, G) = \mathcal{S}(F, G) - \mathcal{S}(G, G).$$

Clearly, $d_S(F, G) \geq 0$ for all $F, G \in \mathcal{F}$ is equivalent to propriety of the scoring rule S. Table 4.2 gives an overview of popular scoring rules and their associated divergence functions. Note that we continue to use the symbol F to denote both the probabilistic forecast $F \in \mathcal{F}$ as well as the cumulative distribution function corresponding to F. For details on score divergences and the relation to the concept of Bregman divergences in convex analysis, see Section 2.3.2.

Relating to the estimator $\hat{F}_m$ in (4.2), we emphasize that

$$S\left(\frac{1}{m}\sum_{i=1}^{m} F_c(\cdot|\theta_i), y\right) \neq \frac{1}{m}\sum_{i=1}^{m} S\left(F_c(\cdot|\theta_i), y\right). \tag{4.3}$$

As detailed below, the term on the left-hand side defines a legitimate and efficient estimator. By contrast, the term on the right-hand side has no particular meaning; it is sometimes erroneously used in the literature (for example, Risser and Calder, 2015, Section 4.2.2). Indeed, for many popular scoring rules S, the left-hand side of (4.3) is less than the right-hand side by construction (Krüger, 2014), so that using the right-hand side leads to a systematic overestimation of the true predictive score.

To highlight this point, consider a toy MCMC sampler with iterations $i = 1, \ldots, m$, where $y_i \sim \mathcal{N}(\mu_i, 1)$, $\mu_i = 0.5\,\mu_{i-1} + \varepsilon_i$, and ε_i iid standard normal. In this example, we know that F_0 is Gaussian with mean zero and variance $7/3$. For a hypothetical realization $y = 0$, we obtain a logarithmic score $\mathrm{LogS}(F_0, 0) = 1.343$. Based on $m = 10^7$ simulated draws, we find that $\mathrm{LogS}\left(\frac{1}{m}\sum_{i=1}^{m} F_c(\cdot|\mu_i), 0\right) = 1.343$. Therefore, the estimator on the left-hand side of (4.3) approximately recovers the true score. By contrast, using the expression on the right-hand side yields a flawed result, since $\frac{1}{m}\sum_{i=1}^{m} \mathrm{LogS}\left(F_c(\cdot|\theta_i), 0\right) = 1.586 \neq 1.343$.

4.2.3 Consistent approximations

To discuss the features of any estimate $\hat{F}_m$ based on either $\{\theta_i\}_{i=1}^m$ or $\{X_i\}_{i=1}^m$, we introduce the notion of a consistent approximation procedure.

Definition: Consistent approximation procedure. Let $\mathcal{F}$ denote a fixed convex class of probability measures and let S denote a scoring rule which is proper relative to $\mathcal{F}$. Given a sample $\theta_1, \dots, \theta_m$ or $X_1, \dots, X_m$ generated as described in Section 4.2.1, an *approximation method* A produces an estimate $\hat{F}_m = A(\theta_1, \dots, \theta_m)$ or $\hat{F}_m = A(X_1, \dots, X_m)$ of F_0, and is *consistent* relative to S and $\mathcal{F}$, if $\hat{F}_m \in \mathcal{F}$ for all m, and

$$d_S(\hat{F}_m, F_0) \longrightarrow 0 \tag{4.4}$$

or, equivalently, $\mathcal{S}(\hat{F}_m, F_0) \to \mathcal{S}(F_0, F_0)$ almost surely as $m \to \infty$ for all $F_0 \in \mathcal{F}$.

Note that $\hat{F}_m$ is a random element that depends on the random sample $\{\theta_i\}_{i=1}^m$ or $\{X_i\}_{i=1}^m$. The specific form of the divergence $d_S(\hat{F}_m, F_0)$ depends on the scoring rule S, see Table 4.2. For the logarithmic score, it is given by the Kullback-Leibler divergence between $\hat{f}_m$ and f_0. For the CRPS, it is given by the integrated quadratic difference of $\hat{F}_m$ and F_0. Generally, each scoring rule thus demands convergence of a different functional of the estimate ($\hat{f}_m$ or $\hat{F}_m$) and the true posterior predictive distribution (f_0 or F_0) to satisfy the conditions for consistency. As we will argue below, this aspect has important implications for the choice of scoring rule and approximation method in applied work. To simplify notation, we will typically not explicitly specify $\mathcal{F}$ and assume that it is immediate from the context in that $\mathcal{F}$ is appropriately chosen such that the scoring rule of interest is proper relative to $\mathcal{F}$ and the score divergence is well-defined.

Note that our concept of a consistent approximation procedure is independent of the question of how well a forecast model approximates the true distribution. The definition thus allows to separate the question of interest (how to find a good approximation $\hat{F}_m$ of F_0) from the distinct task of finding a good model F_0 for G. One could even think of cases in which an inconsistent approximation of F_0 is closer to G than a consistent one. For example, suppose $G = \mathcal{N}(\frac{1}{2}, \frac{1}{12})$, $F_0 = \mathcal{U}(0, 1)$, and consider a Gaussian approximation with $\hat{F}_m = \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$, see Section 4.3.2. If the parameters are estimated using maximum likelihood estimation, $\hat{F}_m \to \mathcal{N}(\frac{1}{2}, \frac{1}{12}) = G \neq F_0$. Clearly, the Gaussian approximation is inconsistent, as it does not converge to F_0. Nevertheless, it attains a *better* score than F_0 as it happens to reproduce the true distribution G in this example. Direct counterexamples can also be derived for specific scoring rules, see for example Appendix 3.C on weighted versions of the logarithmic score. In our experience, such cases often point to severe misspecification, which can be detected via visual inspection of PIT histograms. This suggests to improve the model specification F_0, rather than attempting to "save" a flawed model by using an inconsistent approximation.

We further emphasize that we study convergence in the number of simulation draws, m, given a fixed number of observations used to fit the model (say, T).

Our analysis is thus distinct from traditional Bayesian asymptotic analyses which study convergence of the posterior distribution as $T \to \infty$, see, e.g., Gelman et al. (2014a, Section 4).

4.3 Consistency results

Building on results from classical asymptotic statistics, we now investigate sufficient conditions for consistency of the various popular approximation methods summarized in Table 4.1. As discussed above, consistency will always require convergence of some functional of the estimate $\hat{F}_m$ and the true predictive distribution F_0. We will demonstrate below that the specific conditions on the properties of $\hat{F}_m, F_0$ and the dependence in the MCMC output strongly depend on the scoring rule of interest.

4.3.1 Approximation based on parameter draws

As discussed in Section 4.2, F_0 can frequently be estimated as

$$\hat{F}_m(z) = \frac{1}{m} \sum_{i=1}^{m} F_c(z|\theta_i). \tag{4.5}$$

This estimator is often called a *conditional*, or *Rao-Blackwellized*, (kernel density) estimator, and was first proposed by Gelfand and Smith (1990, Section 2.2), based on earlier work by Tanner and Wong (1987). For independent samples, the Rao-Blackwell theorem implies that exploiting the full conditional distributions (given θ_i) leads to variance reduction (see, e.g., Geweke, 2005, Section 4.4.1). However, this theoretical motivation does not easily extend to the more realistic case of dependent MCMC samples (Geyer, 1995; Chen and Shao, 1997). Therefore, we follow suggestions of Geyer (1995, p. 152) and avoid the term Rao-Blackwellization. Instead, we refer to (4.5) as *mixture-of-parameters* estimator (MPE).

From the ergodicity of the sequence $\theta_1, \ldots, \theta_m$, it follows that for any $z \in \mathbb{R}$,

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^{m} F_c(z|\theta_i) &\longrightarrow \int_{\Theta} F_c(z|\theta) \, \mathrm{d}P_{post}(\theta) \\ &= F_0(z), \end{aligned}$$

almost surely as $m \to \infty$, i.e., $\hat{F}_m(z)$ converges pointwise to $F_0(z)$. Consistency of the mixture-of-parameters estimator relative to the CRPS and the logarithmic score in the sense of equation (4.4) is summarized in the following proposition.

Proposition 4.1 (Consistency of the mixture-of-parameters estimator)
Assume that

(A1) The sequence $\theta_1, \ldots, \theta_m$ is stationary and ergodic.

(A2) Ω_Y *is a compact subset of* $\mathbb{R}$.

(A3) The predictive density $f_c(z|\theta)$ *conditional on* θ *is Lipschitz-continuous in* z *for all* $\theta \in \Theta$.

(A4) The true predictive density f_0 *is continuous and positive on* Ω_Y.

Under (A1) and (A2), the MPE is consistent relative to the CRPS. Under (A1)–(A4), it is consistent relative to the LogS.

A proof of Proposition 4.1 is provided in Appendix 4.B. (A1) is a standard assumption in Bayesian statistics that can be expected to hold widely (see, e.g., Craiu and Rosenthal, 2014, Sections 8.1 and 8.2). (A2) excludes the case of $\Omega_Y = \mathbb{R}$, however, practical applications often require a truncation of the support for numerical reasons, see Section 4.4.4. The validity of assumptions (A3) and (A4) depends on the specific Bayesian forecasting model at hand.

From a practical perspective, application of the mixture-of-parameters estimator requires the knowledge of the full model specification as the conditional distribution $F_c(\cdot|\theta)$ must be known to compute (4.5). There may be situations where this is restrictive, for example, if the goal is to predict a complex transformation of the predictand for which the distribution is not available in a closed analytical form.

4.3.2 Approximations based on simulated samples from the posterior predictive distribution

We next discuss various approximations based on a simulated sample $X_1, \ldots, X_m$ (Option 2 in Section 4.2) and corresponding conditions for consistency relative to several proper scoring rules.

Parametric/Gaussian approximation

The most simplistic approximation method based on a sample $X_1, \ldots, X_m$ is the use of a fixed parametric family of distributions $\mathcal{F}_\gamma, \gamma \in \Gamma \subseteq \mathbb{R}^{d'}$ to approximate F_0, i.e.,

$$A^{\mathcal{F}_\gamma}(X_1, \ldots, X_m) = \hat{F}_m = F_{\hat{\gamma}_m} \in \mathcal{F}_\gamma,$$

where $\hat{\gamma}_m$ is a parameter estimate based on $X_1, \ldots, X_m$. The infinite-dimensional problem of estimating an unknown distribution F_0 is then reduced to a finite-dimensional parameter estimation problem in a parametric family. However, from the definition of consistency in (4.4) it is clear that this approximation scheme can only be consistent if $F_0 \in \mathcal{F}_\gamma$ and a suitable parameter estimator is used, such that, under regularity conditions, $\hat{F}_m \to F_0$ and $S(\hat{F}_m, F_0) \to S(F_0, F_0)$ almost surely.

From an applied perspective, the most important special case is the *quadratic* or *Gaussian approximation* which assumes a Gaussian distribution for the unknown F_0. The parameters $\gamma = (\mu, \sigma^2)$ are estimated via the usual (maximum

likelihood) estimates $\hat{\gamma}_m = (\hat{\mu}, \hat{\sigma}^2) = \left(\frac{1}{m}\sum_{i=1}^m X_i, \frac{1}{m}\sum_{i=1}^m (X_i - \hat{\mu})^2\right)$, the Gaussian approximation is thus given by

$$A^{\mathcal{N}}(X_1, \ldots, X_m) = \hat{F}_m = \mathcal{N}(\hat{\mu}, \hat{\sigma}^2). \tag{4.6}$$

If f_0 is unimodal and symmetric, the quadratic approximation can be motivated by Taylor series expansion of the log predictive density at the mode similar to Gaussian approximations of posterior distributions in large-sample Bayesian inference, see, for example, Gelman et al. (2014a, Chapter 4).

Note that the logarithmic score computed for the quadratic approximation corresponds to the Dawid-Sebastiani score (Dawid and Sebastiani, 1999) up to an affine transformation. We have highlighted this connection in Section 3.4.3 where we used the quadratic approximation to compare the LogS of Bayesian macroeconomic forecasting models. For the particular case of the LogS, the quadratic approximation is not necessarily consistent in the sense of (4.4), but can be unproblematic in comparative model assessment as the LogS is replaced by another proper scoring rule. However, we emphasize that this approach corresponds to computing the DSS rather than the LogS, and therefore explicitly refer to the employed scoring rule as DSS in Section 3.4.3.

For the DSS and other moment-based proper scoring rules such as the error-spread score (2.7), it suffices to estimate the relevant moments $\mathbb{E}X^r$, typically by setting

$$\hat{\mathbb{E}}_m(X^r) = \frac{1}{m}\sum_{i=1}^m X_i^r.$$

Consistency of such approximations follows directly from ergodicity of $\{X_i\}_{i=1}^m$. Note, however, that the DSS and ESS are strictly proper only with respect to restrictive classes of distributions, for example, the DSS is only strictly proper if the class of probability measures $\mathcal{F}$ of interest is fully characterized by the first two moments. Employing the quadratic approximation in conjunction with the LogS might therefore be undesirable in this regard. Further, it should be ensured that the same quadratic approximation is applied for all models as otherwise, values of different scoring rules are compared. For example, if we compare a Bayesian model where we employ the quadratic approximation for the unknown F_0 and a traditional parametric model where we compute the LogS in the usual form (2.2), then this corresponds to comparing the DSS of the Bayesian model and the LogS of the parametric model.

Here, our goal is to compare approximation methods for various proper scoring rules. In this light, the quadratic approximation, and more broadly, any approximation based on a fixed parametric family is typically not consistent as the sufficient condition $F_0 \in \mathcal{F}_\gamma$ is generally unlikely to hold. Therefore, the use of a fixed parametric family is generally not a feasible approximation method despite the apparent popularity in the applied literature. The quadratic approximation (4.6) is used in 6 of the 23 studies employing the LogS summarized in Table 4.1, see Appendix 4.A for details. As we will see below, the other approximation methods discussed here can be considered more suitable alternatives.

Empirical CDF

A natural choice for approximating the unknown predictive CDF corresponding to F_0 is the *empirical cumulative distribution function* (ECDF), i.e.,

$$A^{\text{ECDF}}(X_1, \ldots, X_m) = \hat{F}_m(z) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{X_i \leq z\}, \tag{4.7}$$

where $\mathbb{1}\{X_i \leq z\}$ denotes the indicator function which is 1 if $X_i \leq z$, and 0 otherwise.

If $X_1, \ldots, X_m$ is an independent sample from F_0, the classical Glivenko-Cantelli theorem (e.g., van der Vaart, 2000, Theorem 19.1) states that

$$\left\| \hat{F}_m - F_0 \right\|_\infty = \sup_{z \in \Omega_Y} \left| \hat{F}_m(z) - F_0(z) \right| \longrightarrow 0$$

almost surely as $m \to \infty$. In the context of MCMC output, the independence assumption on $X_1, \ldots, X_m$ is too strong. However, the Glivenko-Cantelli theorem has been generalized to allow for time series dependence, see, e.g., Dehling and Philipp (2002, Theorem 1.1). In particular, it suffices that the sequence is stationary and ergodic with stationary distribution F_0. If Ω_Y is a compact subset of $\mathbb{R}$, this directly implies that the ECDF is a consistent approximation method relative to the CRPS. Consistency relative to the CRPS for the more general case $\Omega_Y = \mathbb{R}$ can be obtained under the additional assumption of a finite first moment as follows.

Proposition 4.2 (Consistency of approximations based on the ECDF)
Assume that

(A1) The sequence $X_1, \ldots, X_m$ is stationary and ergodic.

(A2) $\mathbb{E}_{F_0}|X_1| < \infty$.

Under (A1) and (A2), the empirical CDF is a consistent approximation relative to the CRPS.

A proof is provided in Appendix 4.C. The conditions for consistency of approximations based on the empirical CDF relative to the CRPS are generally not restrictive and can be assumed to hold widely in applications. (A1) is a standard assumption in Bayesian statistics (see, e.g., Craiu and Rosenthal, 2014, Sections 8.1 and 8.2), and (A2) is implicitly assumed anyway as the CRPS is only strictly proper relative to the class of probability distributions with finite first moment.

From a practical perspective, computing the CRPS of the ECDF-based approximation (4.7) is straightforward as

$$\text{CRPS}\left(\hat{F}_m, y\right) = \frac{1}{m} \sum_{i=1}^{m} |X_i - y| - \frac{1}{2m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} |X_i - X_j|. \tag{4.8}$$

The CRPS of the empirical CDF $\hat{F}_m$ can thus be computed directly from the sample $X_1, \ldots, X_m$, see Grimit et al. (2006). Equation (4.8) is often referred to as *kernel representation*.

Hersbach (2000) suggests the following computationally efficient implementation of (4.8). Let $\tilde{X}_1, \ldots, \tilde{X}_m$ denote the sample $X_1, \ldots, X_m$ after sorting in ascending order, i.e., $\tilde{X}_i \leq \tilde{X}_j$ if $i < j$. The CRPS of the ECDF-based approximation can then be computed as

$$\mathrm{CRPS}\left(\hat{F}_m, y\right) = \sum_{i=0}^{m} \alpha_i \left(\frac{i}{m}\right)^2 + \beta_i \left(1 - \frac{i}{m}\right)^2, \tag{4.9}$$

where

$$\alpha_i = \begin{cases} 0, & y < \tilde{X}_i, \\ y - \tilde{X}_i, & \tilde{X}_i < y < \tilde{X}_{i+1}, \\ \tilde{X}_{i+1} - \tilde{X}_i, & \tilde{X}_{i+1} < y, \end{cases} \qquad \beta_i = \begin{cases} \tilde{X}_{i+1} - \tilde{X}_i, & y < \tilde{X}_i, \\ \tilde{X}_{i+1} - y, & \tilde{X}_i < y < \tilde{X}_{i+1}, \\ 0, & \tilde{X}_{i+1} < y, \end{cases}$$

for $i = 1, \ldots, m-1$, and

$$\alpha_i = \begin{cases} 0, & y < \tilde{X}_1, \\ y - \tilde{X}_m, & \tilde{X}_m < y, \end{cases} \qquad \beta_i = \begin{cases} \tilde{X}_1 - y, & y < \tilde{X}_1, \\ 0, & \tilde{X}_m < y, \end{cases}$$

for $i = 0$ and $i = m$. Representations (4.8) and (4.9) are algebraically equivalent, however, evaluation of (4.9) is more efficient as only $\mathcal{O}(m \log m)$ operations are required compared to $\mathcal{O}(m^2)$ operations for evaluating (4.8) (Gneiting and Raftery, 2007).

The Székely and Rizzo (2005) formula

Equation (17) of Székely and Rizzo (2005) implies that for any distribution F_0 with finite first moment,

$$\mathrm{CRPS}(F_0, y) = \mathbb{E}_{F_0}|X - y| - \frac{1}{2}\mathbb{E}_{F_0}|X - X'|, \tag{4.10}$$

where $\mathbb{E}_{F_0}$ denotes expectation with respect to F_0, and X and X' are two independent copies drawn from F_0. In Section 2.3.1, equation (4.10) was introduced as kernel representation of the CRPS.

When comparing (4.8) and (4.10), it becomes clear that for the ECDF to be consistent relative to the CRPS, it must hold that almost surely

$$\frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} |X_i - X_j| \longrightarrow \mathbb{E}_{F_0}|X - X'| \tag{4.11}$$

as $m \to \infty$. Crucially, the terms $\{X_i\}_{i=1}^m$ occurring on the left-hand side are *dependent* draws with stationary distribution F_0, whereas the expectation on the right-hand side concerns *independent* copies X, X' with distribution F_0. Our

discussion above implies that the convergence relation in (4.11) holds true if the process for $\{X_i\}_{i=1}^m$ is stationary and ergodic, with finite first moment. Under these assumptions, (4.11) also follows from general limit theorems for dependent U-statistics, see Aaronson et al. (1996, Theorem U), Dehling and Philipp (2002, Section 5.2), and Székely and Rizzo (2013).

Given the dependence issues just discussed, it is not obvious how to design an empirical analogue to (4.10), other than (4.8). While measures such as thinning or rearranging the MCMC sequence may reduce autocorrelation in practice, they are no rigorous justification for claiming independence. We hence advise against such methods, and propose to use the consistent estimator (4.8) in the computationally efficient representation (4.9) instead.

Kernel density estimation

Thus far, we have demonstrated that approximating the unknown posterior predictive CDF F_0 by the empirical CDF of a sample $X_1, \dots, X_m$ is consistent relative to the CRPS under weak regularity conditions that will generally be valid in most applications. However, other proper scoring rules such as the logarithmic score require an estimate of the corresponding predictive density f_0 for which there is no immediate analogue to the empirical CDF.

A classical approach to nonparametric estimation of a probability density function is kernel density estimation (KDE, Rosenblatt, 1956). The *kernel density estimate* of f_0 based on the sample $X_1, \dots, X_m$ is given by

$$\hat{f}_m(z) = \hat{f}_{\mathrm{KDE}}(z) = \frac{1}{m h_m} \sum_{i=1}^{m} K\left(\frac{z - X_i}{h_m}\right), \tag{4.12}$$

where $h_m > 0, m \in \mathbb{N}$ is a sequence of bandwidths and K is a kernel, i.e., a nonnegative symmetric bounded density. Typical choices for kernel functions are the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$ or the Epanechnikov kernel $K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}\{|u| \leq 1\}$.

Here, we discuss conditions for the consistency of the kernel density estimator relative to the logarithmic score. As we will argue below, these conditions are typically much stronger than the conditions needed for consistency relative to the CRPS, in particular compared to the ECDF-based approximation, see Proposition 4.2.

Recall that the score divergence associated with the LogS is the Kullback-Leibler divergence

$$\mathrm{KL}\left(\hat{f}_m, f_0\right) = \int_{-\infty}^{\infty} f_0(z) \log\left(\frac{f_0(z)}{\hat{f}_m(z)}\right) \mathrm{d}z,$$

see Section 2.3.2. Consistency of an approximation method that produces an estimate $\hat{f}_m$ of the true forecast density f_0 relative to the LogS thus requires

almost sure convergence of $\mathrm{KL}(\hat{f}_m, f_0) \to 0$ as $m \to \infty$ which is implied by

$$\left\| 1 - \frac{\hat{f}_m}{f_0} \right\|_\infty \longrightarrow 0, \tag{4.13}$$

as $m \to \infty$, see Ikeda (1960). Both convergence of $\mathrm{KL}(\hat{f}_m, f_0) \to 0$ as well as convergence of (4.13) are difficult to establish for typical approximation procedures. Instead, we investigate conditions for almost sure strong uniform consistency, i.e.,

$$\left\| \hat{f}_m - f_0 \right\|_\infty \longrightarrow 0, \tag{4.14}$$

as $m \to \infty$ under further regularity assumptions. The convergence in (4.14) implies (4.13), and thereby convergence of the Kullback-Leibler divergence and consistency relative to the LogS, if f_0 is bounded away from 0 from below. For simplicity, we restrict our attention to continuous densities supported on compact subsets of $\mathbb{R}$. Due to the sensitivity of the Kullback-Leibler divergence to the tails of $\hat{f}_m$ and f_0, generalizations to densities with arbitrary support are not straightforward and for example require that the tail properties of the chosen kernel K and of the true density are carefully matched, see Hall (1987) and Wasserman (2006, p. 57).

We now discuss sufficient conditions for almost sure strong uniform consistency (4.14) of the kernel density estimator (4.12) which in turn implies consistency relative to the LogS in the sense of (4.4).

Almost sure uniform convergence of the kernel density estimator for independent samples was first proved by Nadaraya (1965). Seminal work of Roussas (1969) and Rosenblatt (1970) initiated studies of asymptotic properties of kernel density estimators for dependent sequences. Naturally, convergence and general asymptotic properties of kernel density estimators critically depend on the employed dependence model, see Györfi et al. (1989) and Wied and Weißbach (2012) for surveys.

Here, we are interested in samples $X_1, \ldots, X_m$ obtained as output of MCMC algorithms, and thus focus on mixing conditions for Markov chains that have been considered in the literature. We omit the technical definitions of the involved mixing conditions, and instead relate to practically relevant concepts encountered earlier. For a detailed study of the various mixing conditions, see Bradley (2005).

Stationary and ergodic Markov chains are known to be strongly (or α-) mixing (Rosenblatt, 1971). For stationary Markov chains, the stronger notion of absolute regularity (or β-mixing) is equivalent to the Markov chain being further Harris-recurrent and aperiodic (Bradley, 2005, Corollary 3.6). β-mixing implies α-mixing. Typical Markov chains obtained as MCMC output in practical applications will thus usually be at least α-mixing, while β-mixing can be established for many algorithms such as the Metropolis-Hastings algorithm (Tierney, 1994; Robert and Casella, 2004, Section 7.3).

Roussas (1988) proves almost sure strong uniform consistency (4.14) of the kernel density estimator on expanding compact subsets of $\mathbb{R}$ if the kernel is a

Lipschitz-continuous bounded density with $\int |z|K(z)\,\mathrm{d}x < \infty$, the true density f_0 is Lipschitz-continuous, the sequence $X_1, \ldots, X_m$ is stationary and α-mixing, and further technical conditions on the bandwidths h_m are satisfied. Specifically, admissible choices of θ for bandwidths of the form $h_m = m^{-\theta}$ depend on the exact mixing properties of the sequence $X_1, \ldots, X_m$, see Roussas (1988, Theorem 3.1) for details. Similar results have been derived by Györfi et al. (1989), see also Yu (1993) for generalizations and optimal minimax convergence rates under β-mixing conditions.

Under these assumptions, consistency of KDE relative to the LogS follows from Theorem 3.1 of Roussas (1988) and the relation of convergence of the KL divergence and convergence in (4.14). Clearly, these regularity conditions are much more stringent compared to those in Propositions 4.1 and 4.2, and are difficult to check in practice. For example, the exact mixing coefficients will typically be not available in applications.

As the mixing coefficients are generally unknown, it is challenging to find choices of bandwidths that satisfy the conditions for consistency of the KDE. We will therefore focus on the practically relevant case of standard bandwidth selection methods which at least comply with the basic condition that $h_m \to 0$ as $m \to \infty$. While the presence of dependence in the sample calls for suitably adapted data-driven variants, it has been demonstrated that some standard bandwidth selection algorithms developed for independent data are robust to moderate amounts of dependence in the data (Hart and Vieu, 1990), and can even be considered a good choice for some strongly dependent sequences (Hall et al., 1995). Alternatively, Sköld and Roberts (2003) present a bandwidth selection rule which is tailored to the properties of a Metropolis-Hastings algorithm.

By integrating (4.12), we can hypothetically also use KDE to obtain an estimate of the posterior predictive CDF F_0. Consistency of such a KDE-based approximation relative to the CRPS can be established under somewhat weaker conditions than those required for the LogS. For the CRPS, note that

$$\left\| \hat{F}_m - F_0 \right\|_\infty \leq \int_{\mathbb{R}} \left| \hat{f}_m(z) - f_0(z) \right| \mathrm{d}z, \tag{4.15}$$

see the proof of Proposition 4.1 in Appendix 4.B for details. Hence, $\mathcal{L}^1$ consistency of the estimated densities already implies strong uniform consistency of the corresponding CDFs. Tran (1989) studies convergence of (4.15) for the kernel density estimator (4.12) under α-mixing assumptions. However, we emphasize that the empirical CDF (4.7) provides a simpler and more suitable approximation procedure for which consistency relative to the CRPS can be established under much weaker assumptions which will generally hold in most applications, see Proposition 4.2.

4.4 Simulation study

We now investigate the various approximation methods in a simulation study that is designed to emulate realistic MCMC behavior. Here, the posterior pre-

dictive distribution F_0 is known by construction, we can therefore compare the different approximations to the true forecast distribution. We do so by comparing the distribution of score divergences of the different approximations methods and examining convergence and variation across replications of the simulation experiment.

4.4.1 Basic setup

In order to judge the quality of an approximation $\hat{F}_m$ of F_0 we consider the score divergence between $\hat{F}_m$ and F_0, $d_S(\hat{F}_m, F_0)$. Note that $d_S(\hat{F}_m, F_0)$ is a random variable, since $\hat{F}_m$ depends on the particular MCMC sample ($\theta_1, \ldots, \theta_m$ or $X_1, \ldots, X_m$) used to estimate it. In our results below, we therefore consider the distribution of $d_S(\hat{F}_m, F_0)$ across repeated simulation runs. For a generic approximation method producing an estimate $\hat{F}_m$, we proceed as follows:

- For simulation run $k = 1, \ldots, K$:
  - Draw the random "MCMC samples" $\theta^{(k)} = \left\{\theta_1^{(k)}, \ldots, \theta_m^{(k)}\right\}$ and $X^{(k)} = \left\{X_1^{(k)}, \ldots, X_m^{(k)}\right\}$.
  - Compute the approximation $\hat{F}_m^{(k)}$ based on either $\theta^{(k)}$ or $X^{(k)}$.
  - Compute the divergence $d_S\left(\hat{F}_m^{(k)}, F_0\right)$ (via numerical integration, see Section 4.4.4).
- Summarize the sample $d_S\left(\hat{F}_m^{(1)}, F_0\right), \ldots, d_S\left(\hat{F}_m^{(K)}, F_0\right)$

In order to simplify notation, we will typically suppress the superscript (k) identifying the Monte Carlo iteration. The results presented below are based on $K = 1\,000$ replications of the simulation experiment.

4.4.2 Description of the data generating process

We simulate data $X_1, \ldots, X_m$ from a compounded Gaussian distribution,

$$f_0(z) = \int_{\sigma^2 \in \mathbb{R}_+} \mathcal{N}\left(z|\mu = 0, \sigma^2\right) \, \mathrm{d}H_0(\sigma^2),$$

where $H_0(\sigma^2)$ denotes the stationary distribution of σ^2.

To mimic a realistic MCMC scenario, the draws $\{\theta_i\}_{i=1}^m$ and $\{X_i\}_{i=1}^m$ should both be dependent across iterations $i = 1, \ldots, m$. We achieve this by drawing a sequence $\{\sigma_i^2\}_{i=1}^m$ from the model proposed by Fox and West (2011). This model implies autoregressive-type dependence in the draws, while still yielding an (inverse gamma) stationary distribution. Here we use the simplest (univariate)

Table 4.3: Overview of hyper-parameters for the data generating process in the simulation study, see equations (4.16) to (4.18).

Parameter	Main role	Value(s) considered
α	Persistence of σ_i^2	$\{0.1, 0.5, 0.9\}$
s	Unconditional distribution of σ_i^2	2
n	Unconditional distribution of σ_i^2	$\{12, 20\}$

variant of the model, as described in their Section 2.3. Given hyper-parameters $n > 0, s > 0, \alpha \in (-1, 1)$, we simulate

$$\sigma_i^2 = \psi_i + v_i^2 \, \sigma_{i-1}^2, \tag{4.16}$$

$$\psi_i \overset{\text{iid}}{\sim} \text{IG}\left(\frac{n+3}{2}, \frac{ns(1-\alpha^2)}{2}\right), \tag{4.17}$$

$$v_i|\psi_i \sim \mathcal{N}\left(\alpha, \frac{\psi_i}{ns}\right), \tag{4.18}$$

where IG is the inverse gamma (IG) distribution. We parametrize it such that $Z \sim \text{IG}(a, b) \Leftrightarrow 1/Z \sim \text{G}(a, b)$, where G is the gamma distribution.

For the unconditional distribution, this implies that

$$\sigma_i^2 \sim \text{IG}\left(\frac{n+2}{2}, \frac{ns}{2}\right)$$

with expected value $\mathbb{E}(\sigma_i^2) = s$ and variance $\frac{s^2}{0.5\,n-1}$. Note that the parameter α has no impact on the unconditional distribution of σ_i^2, but governs the autocorrelation across draws $i = 1, \ldots, m$. The unconditional mean of v_i^2, which can be viewed an average AR coefficient (Fox and West, 2011, Section 2.3), is given by $(n\alpha^2 + 1)/(n + 1)$.

Conditional on $\theta_i = \sigma_i^2$, we have that $X_i \sim \mathcal{N}(0, \sigma_i^2)$. Unconditionally, our setting implies that

$$f_0(z) = t\left(z \middle| 0, \; \frac{ns}{n+2}, \; n+2\right), \tag{4.19}$$

where $t\,(\cdot|a,\, b,\, c)$ denotes the density of a variable Z with the property that $\frac{Z-a}{\sqrt{b}}$ is t distributed with c degrees of freedom (see, e.g., Gneiting, 1997, for general results on compounding Gaussian distributions). In contrast to typical applications of Bayesian forecasting methods, our simulation study is thus designed such that the posterior predictive distribution F_0 is known and available in closed form.

Table 4.3 summarizes all parameters of the data generating process. α determines the persistence of σ_i^2 and $\{X_i\}_{i=1}^m$. We consider three values, aiming to mimic MCMC chains with different persistence properties. n governs the tail thickness of unconditional t distribution for X_i. This is likely to be important for the performance of normal approximations. We consider values of 12 and 20

which represent moderate degrees of tail thickness that seem realistic for macroeconomic time series like GDP growth or inflation. The parameter s represents a scale effect that does not appear particularly important, we therefore consider only one value for it.

4.4.3 Approximation methods

We consider the following approximation methods that have been introduced in Section 4.3.

1. Mixture of parameters estimator, i.e., we condition on $\theta_i = \sigma_i^2$ and obtain the forecast distribution

$$\hat{F}_m(z) = \frac{1}{m} \sum_{i=1}^{m} \Phi\left(\frac{z}{\sigma_i}\right),$$

   where σ_i is the predictive standard deviation drawn in MCMC iteration i.

2. Gaussian approximation, i.e.,

$$\hat{F}_m(z) = \Phi\left(\frac{z - \hat{\mu}_m}{\hat{\sigma}_m}\right),$$

   where Φ is the CDF of the standard normal distribution, and $\hat{\mu}_m$ and $\hat{\sigma}_m$ are the empirical mean and standard deviation of $X_1, \ldots, X_m$.

3. Nonparametric estimation using the simple empirical cumulative distribution function, i.e.,

$$\hat{F}_m(z) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{X_i \leq z\}.$$

   This estimator only allows for obtaining a predictive CDF and can thus be used for the CRPS, but not for the logarithmic score.

4. Nonparametric kernel density estimation using a Gaussian kernel and the Silverman (1986, Section 3.4) plug-in rule for bandwidth selection. Formally, this means that

$$\hat{F}_m(z) = \frac{1}{m} \sum_{i=1}^{m} \Phi\left(\frac{z - X_i}{h_m}\right),$$

   where Φ again denotes the CDF of the standard normal distribution and

$$h_m = \left(\frac{4\hat{\sigma}_m^5}{3m}\right)^{\frac{1}{5}} \approx 1.06\, \hat{\sigma}_m\, m^{-\frac{1}{5}}$$

   with $\hat{\sigma}_m$ as above is the selected bandwidth. Our choice of the bandwidth selection rule is motivated by simulation evidence in Hall et al. (1995). Using the Sheather and Jones (1991) rule, as well as biased and unbiased cross-validation methods, yields similar but slightly inferior results.

Clearly, the Gaussian approximation can not be consistent relative to the CRPS or the LogS as F_0 is not Gaussian, see equation (4.19). The conditions for consistency of the MPE and the ECDF are satisfied if we assume that Ω_Y is a compact subset of $\mathbb{R}$, see Propositions 4.1 and 4.2. This assumptions is justified here as a truncation of the support is required when estimating the score divergences due to numerical issues, see Section 4.4.4. Consistency of KDE relative to the LogS can not be verified as the exact mixing properties are unknown.

4.4.4 Estimation of the score divergence

For each Monte Carlo replication k, sample size m, and approximation method described above, we need to evaluate the score divergence $d_S(\hat{F}_m^{(k)}, F_0)$. The divergence takes the form of a univariate integral (see Table 4.2) which is typically not available in closed form for the situation here. Therefore, we estimate $d_S(\hat{F}_m^{(k)}, F_0)$ via numerical integration as implemented in the `R` function `integrate`.

This turns out to be unproblematic for all approximation methods if the scoring rule is the CRPS. For the case that $\hat{F}_m$ is the empirical distribution function, we split the integration problem into several parts, each corresponding to a jump point of $\hat{F}_m$.

For the logarithmic score, the integration is numerically more challenging as the logarithm of $\hat{f}_m(z)$ has to be computed in the tails of the density. We therefore truncate the support of the integral to the minimal and maximal values z which still yield (numerically) finite values of the integrand. We emphasize, however, that this is a purely numerical issue. Theoretically, the (Kullback-Leibler) divergence $d_{\text{LogS}}(\hat{F}_m^{(k)}, F_0)$ is finite for all approximation methods $\hat{F}_m$ considered here.

4.4.5 Results

In the interest of brevity, we restrict our attention to results for a certain set of parameters of the data generating process in Section 4.4.2, with $(\alpha, s, n) = (0.5, 2, 12)$ implying an unconditional t distribution with 14 degrees of freedom and intermediate autocorrelation of the MCMC draws. The results are robust across other parameter constellations.

Figure 4.1 illustrates the performance of the four approximation methods described in Section 4.4.3 under the LogS and the CRPS by showing the distribution of the score divergences $d_S(\hat{F}_m, F_0)$ for different sample sizes. A first striking finding is that the mixture-of-parameters estimator dominates the other methods by a wide margin for both scoring rules. The divergences are very close to zero, and show little variation across the 1 000 Monte Carlo replications indicated by very short black vertical bars in Figure 4.1.

Figure 4.2 provides further insight into the performance of the mixture-of-parameters estimator by summarizing the distribution of the obtained score divergences for a finer grid of small sample sizes. For approximating the CRPS, as little as 150 simulation draws suffice for the method to attain a smaller median

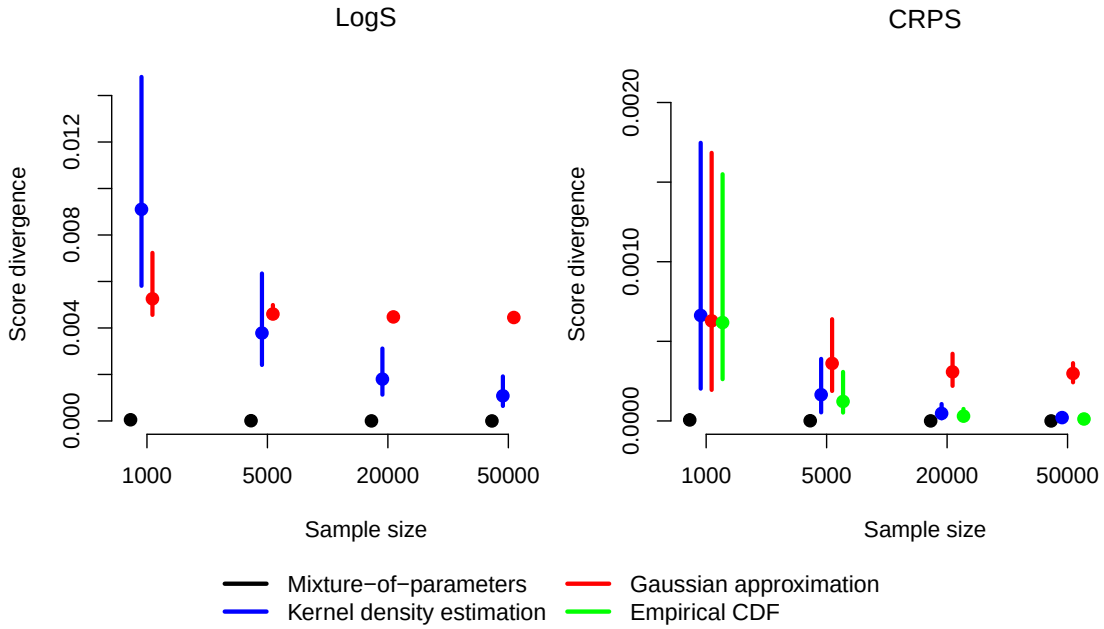


Figure 4.1: Summary of score divergences for various approximation methods in the simulation study for the data generating process in equations (4.16)–(4.18), with parameters $(\alpha, s, n) = (0.5, 2, 12)$. For a given method and sample size, the vertical bars range from the 10th to the 90th percentile of the score divergences observed across the 1 000 replications of the simulation experiment. The dots mark the median of the score divergences.

divergence than the kernel density estimator based on 20 000 simulation draws indicated by the horizontal blue line. The superiority of the mixture-of-parameters estimator is even more pronounced for the LogS, where only 50 simulation draws are required to outperform the kernel density estimator based on 20 000 draws. We have focused on a comparison with the kernel density estimator as the empirical CDF is only applicable for the CRPS, but the results are similar.

Returning to Figure 4.1, we further observe the lack of consistency of the Gaussian approximation as the score divergence does not go to zero for large sample sizes. This is a simple consequence of the fact that F_0 is a t distribution and therefore $F_0 \notin \mathcal{F}_\gamma$ violating the elementary condition for consistency of the Gaussian approximation.

For the logarithmic score, the performance of the kernel density estimator is very variable across the 1 000 Monte Carlo iterations even for large sample sizes as indicated by the long blue vertical bars. Although the conditions for consistency are satisfied for the kernel density estimator, the practical usefulness thus appears to be limited as a large sample size is required for a sufficiently small divergence. The performance of the kernel density estimator is less fragile under the CRPS where it performs comparably to the empirical cumulative distribution function. Importantly, these findings are conditional on using the Silverman (1986) rule of

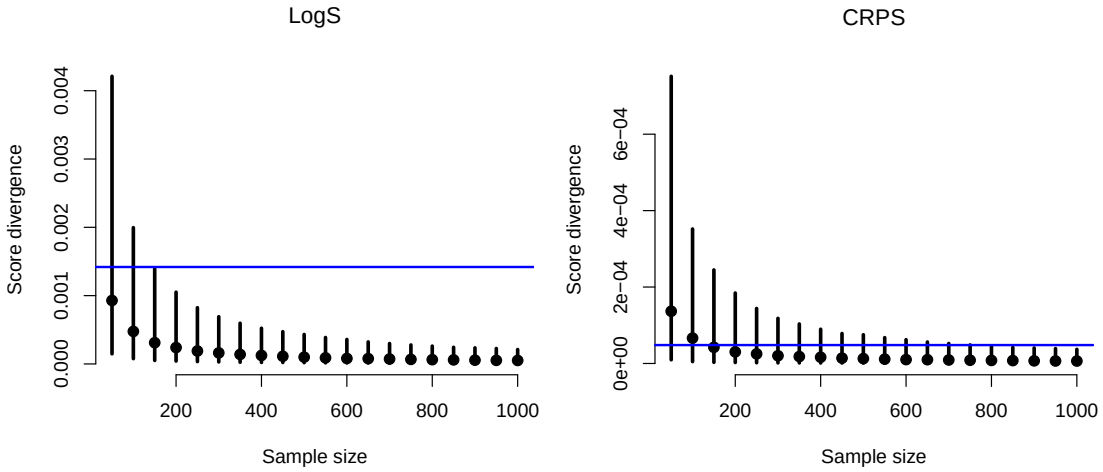


Figure 4.2: Details on the performance of the mixture-of-parameters estimator. The design corresponds to Figure 4.1, but for a different grid of sample sizes. The horizontal blue line marks the median divergence of the kernel density estimator based on 20 000 observations.

thumb for bandwidth selection as motivated above. Other bandwidth selection rules we experimented with (such as biased and unbiased cross-validation, and the Sheather and Jones (1991) rule) had a clear tendency to yield inferior results indicated by slower convergence and higher variability across replications of the simulation experiment.

4.5 Case study

In the simulation study, we investigated the efficiency of the approximation methods in a scenario that was designed to mirror realistic MCMC behavior with dependent samples. Knowing the true forecast distribution F_0 by construction allowed us to empirically assess consistency of the approximation methods by computing the score divergences $d_S(\hat{F}_m, F_0)$ and examining convergence to zero.

In practical applications of Bayesian forecasting methods, the posterior predictive distribution F_0 is typically not available in a closed analytical form. Therefore, computing or estimating the object of interest for assessing consistency, i.e., the score divergence $d_S(\hat{F}_m, F_0)$, is not possible in applications. Therefore, we compare the approximation methods via their average out-of-sample predictive performance over a verification period, and examine the variation of the mean scores across multiple Markov chains obtained by multiple runs of the forecasting model. While studying the predictive performance does not allow to assess consistency of the approximation methods, it does allow us to study the efficiency and applicability of the approximations in a practical application.

This section presents a case study on a popular Bayesian econometric model for quarterly growth rates of real GDP in the U.S. from the second quarter of 1947 to the fourth quarter of 2014, see Figure 4.3. We consider the real-time

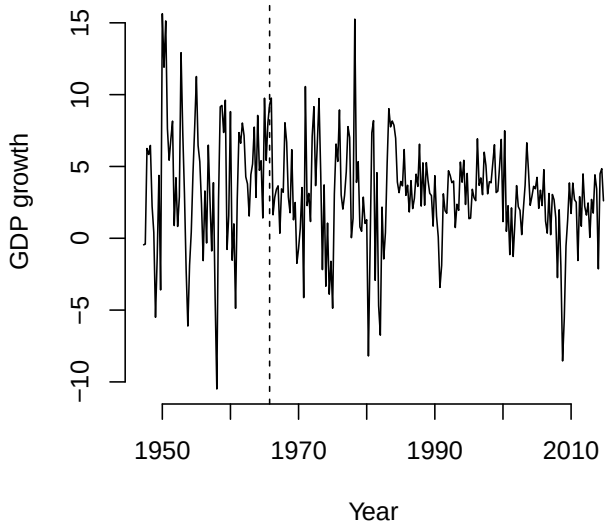


Figure 4.3: Observations of GDP growth in the U.S. from the second quarter of 1947 to the fourth quarter of 2014. The vertical dashed line marks the beginning of the verification period over which the Bayesian Markov-Switching AR(2) model (4.20) is evaluated.

data set provided by the Federal Reserve Bank of Philadelphia.[3] As argued in Section 3.4.1, the use of real-time data allows to account for data revisions and publication lags which are important in practice.

We compute current quarter forecasts from a Markov-Switching AR(2) model proposed by Hamilton (1989). The model is given by

$$Y_t = \nu_{s_t} + \alpha_{1,s_t} Y_{t-1} + \alpha_{2,s_t} Y_{t-2} + \varepsilon_t, \tag{4.20}$$

with $\varepsilon_t \sim \mathcal{N}(0, \sigma^2_{s_t})$. $s_t \in \{1, 2\}$ is a discrete state variable that switches according to a first-order Markov chain. Our Bayesian implementation follows Amisano and Giacomini (2007). In order to better identify the latent states, we assume that the residual variance is larger in the first than in the second state. The model is estimated using a Gibbs sampling algorithm, see Amisano and Giacomini (2007, Section 6.3) for details. The data sample used for model estimation is recursively expanded as forecasting moves forward in time.

Let θ_i denote the complete set of latent states and model parameters sampled at iteration i of the Gibbs sampler. Given θ_i, the forecast distribution of the model in (4.20) is Gaussian. By contrast, the model's posterior predictive distribution F_0 after integrating over θ_i is not available in closed analytical form and is potentially skewed or multimodal.

Regarding the theoretical conditions for consistency of the approximation methods summarized in Propositions 4.1 and 4.2, the assumptions for consistency of the empirical CDF relative to the CRPS are clearly valid. By contrast, consistency of the mixture-of-parameters estimator and KDE is difficult to establish

[3] `http://www.phil.frb.org/research-and-data/real-time-center/real-time-data/`

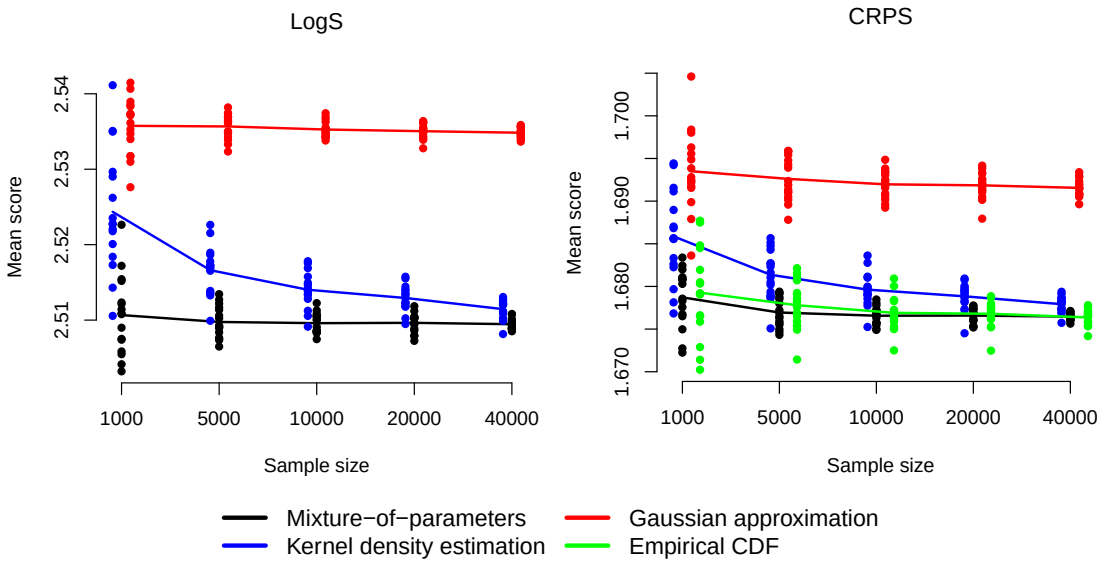


Figure 4.4: Predictive performance of current quarter forecasts of GDP growth by the Bayesian Markov-Switching AR(2) model (4.20) for the U.S. over an evaluation period from 1965:Q4 to 2014:Q3. The dots are mean scores $\bar{S}_{m,c}$ of the approximation methods for various sample sizes m, see equation (4.21). Each dot represents one of the 16 parallel MCMC chains, and the lines represent averages across chains.

as f_0 and the exact mixing properties are unknown, and the conditions on continuity, support and bandwidths thus cannot be checked. However, note that all other assumptions on the conditional densities $f_c(\cdot|\theta)$, as well as the kernel K are fulfilled.

At each forecast origin date $t = 1, \ldots, T$, we discard the first 25 000 burn-in draws, and use 40 000 draws post burn-in. We construct 16 parallel chains in this way. The forecast distributions produced by the different approximation methods are evaluated over an out-of-sample verification period from the fourth quarter of 1965 to the third quarter of 2014. The mean score of a given approximation method, based on m MCMC draws and chain index $c \in \{1, \ldots, 16\}$ is given by

$$\bar{S}_{m,c} = \frac{1}{T} \sum_{t=1}^{T} S\left(\hat{F}_{m,c,t}, y_t\right), \tag{4.21}$$

where $\hat{F}_{m,c,t}$ is the estimated forecast distribution at time t. Smaller variation of $\bar{S}_{m,c}$ across chains c indicates lower dependence on the arbitrary random seed used to generate the chain.

Figure 4.4 shows the mean scores $\bar{S}_{m,c}$ of the different approximation methods for sample sizes m between 1 000 and 40 000. For both the LogS and the CRPS, the mean scores of the mixture-of-parameters estimator, the kernel density estimator and the approximation based on the empirical CDF appear to converge to the same limit. By contrast, the mean scores of the Gaussian approximation seem

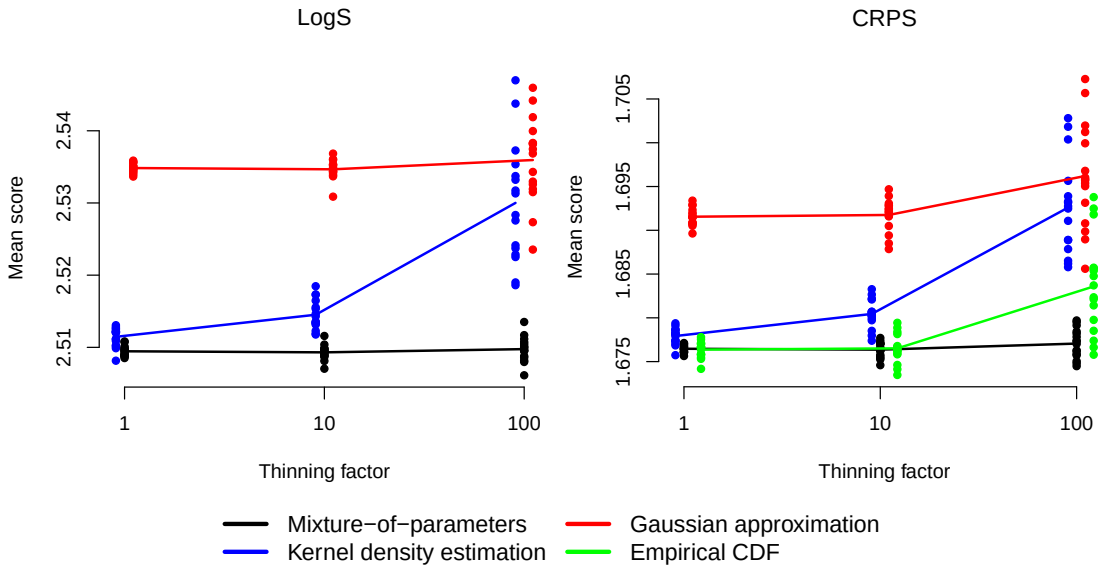


Figure 4.5: Effect of thinning on the mean scores for the different approximation methods. The results are based on the sample of size 40 000 in Figure 4.4. Each dot represents one of the 16 parallel MCMC chains, and the lines represent averages across chains.

to converge to a different limit, presumably because F_0 is not Gaussian and the parametric approximation is therefore not consistent. Although we are interested in the variation across multiple chains rather than the actual score values, it is worth noting that the predictive performance of the Gaussian approximation is worse compared to the alternatives, highlighting the likely non-Gaussian nature of the distribution of the observations of GDP growth.

For both scoring rules the mean values of the scores of the MPE across chains are almost constant in the sample size, and the variation is considerably smaller compared to the other approximation methods. When comparing the KDE and ECDF-based approximations for the CRPS, we note that the variation of the mean scores across chains is similar, with the mean values of the ECDF-based approximations showing smaller deviations from those of the mixture-of-parameters estimator. The impact of the sample size on the variation of the obtained mean scores across chains is smallest for the MPE, where a sample size of $m = 5\,000$ results in a similar (LogS), or smaller (CRPS) variation compared to KDE and ECDF-based approximations with a much larger sample size of $m = 20\,000$. These observations highlight the superior efficiency of the MPE, and are all in line with the behavior of the approximation procedures in the simulation study.

We further investigate the effect of thinning the Markov chains. Thinning a chain by a factor of τ means that only every τth simulated value is kept, and the rest is discarded. Thinning is often routinely applied in the literature with the goal of reducing autocorrelation in the simulation draws. Of the 39 articles summarized in Table 4.1, around a third explicitly report thinning of the simulation output, with factors typically ranging from $\tau = 2$ to $\tau = 100$.

Figure 4.5 illustrates the effect of thinning on the mean scores $\bar{S}_{m,c}$ of the approximation methods. Starting from samples of size $m = 40\,000$, we thin the chains by factors of 10 and 100, and compute the mean scores of the corresponding approximations. It can be observed that thinning generally degrades the efficiency of all approximation methods by increasing the mean scores and the between-chain variation of the scores. Therefore, the additional computational costs of producing, and subsequently thinning a larger chain are not justified. Note that the negative effect on both the mean score and the variation across chains is small for the MPE, again indicating the superior efficiency compared to the alternatives.

These observations are not surprising and are in line with findings from the theoretical and applied literature indicating that the greater precision of unthinned chains is a salient feature of MCMC simulations, see Geyer (1992), MacEachern and Berliner (1994), and Link and Eaton (2012). Note that historically, there existed other legitimate reasons for thinning such as limitations in computer memory and storage which are likely no longer valid with the computational resources available today.

4.6 Discussion

In this chapter, we have investigated how to make and evaluate probabilistic forecasts based on MCMC output. The formal notion of consistency in (4.4) allows us to assess the appropriateness of approximation methods from a theoretical perspective using the framework of proper scoring rules. The required conditions for consistency critically depend on the scoring rule of interest. We have demonstrated that the empirical CDF is consistent relative to the CRPS under weak regularity conditions that are standard assumptions in Bayesian statistics. Consistency of the mixture-of-parameters estimator relative to the CRPS can be established under some additional conditions which are unlikely to be restrictive in most applications. Consistency relative to the LogS generally requires more stringent regularity conditions. For example, additional continuity assumptions on the posterior predictive density and the conditional densities are required for the MPE. In particular, much more restrictive conditions are required if kernel density estimation is used in conjunction with the logarithmic score. Despite the popularity in the literature, parametric approximations which assume a fixed parametric family for F_0 are generally not consistent.

Following these theoretical considerations as well as the investigation of the efficiency of the different approximations from the practical perspective taken in the simulation and case study, we consider the following general recommendations. If the conditions for consistency can be assumed to hold, the mixture-of-parameters estimator provides an efficient approximation method that outperforms the alternative approaches in our practical examples. The simulation and case study suggest that a moderate number of draws (say, 5 000) often seems enough. For the CRPS, using the empirical CDF along with efficient implementations of the kernel representation provides an alternative that is likely consistent in most ap-

plications. Interestingly, although the mixture-of-parameters estimator is more efficient than approximations based on the empirical CDF in our simulation experiments, it appears to be rarely used in the applied literature, see Table 4.1 and Appendix 4.A for details.

The ECDF-based approximation further provides an appealing option if it is for some reason desirable to draw directly from the posterior predictive distribution F_0. For example, Krüger et al. (2015) consider a postprocessing method (entropic tilting) which operates on draws of the forecast distribution, and would be difficult to apply to the mixture-of-parameters approximation. Utilizing the LogS based on such a sample $X_1, \dots, X_m$ proves more problematic as the kernel density estimator requires more stringent regularity conditions and appears to be less efficient in the simulation and case studies.

The recommendations based on the theoretical and applied considerations demonstrated here are implemented in the `scoringRules` package for `R`, that has been developed in joint work with Alexander Jordan and Fabian Krüger, and was introduced in Section 2.3.4, see also Jordan et al. (2016) for details. The implementations of the LogS and CRPS for a given sample $X_1, \dots, X_m$, as well as default choices for tunning parameters follow our suggestions, and aim to provide readily applicable and efficient implementations. The mixture-of-parameters estimator based on a sample $\theta_1, \dots, \theta_m$ depends on the specific structure of the Bayesian forecasting model and can therefore not be included in a general form. However, the large number of implemented analytical solutions of the CRPS and LogS allow for a straightforward and efficient computation.

In Section 4.3, we have derived conditions for consistency for popular scoring rules. As discussed above, these conditions critically depend on the respective scoring rule of interest, and might be much more involved and difficult to formulate for more complex scoring rules. For example, consistency relative to the Hyvärinen score (HS, compare for Table 4.2) will require the existence and convergence of derivatives of $\hat{f}_m$ and f_0. While the specific sufficient conditions for such scoring rules may be difficult to derive, it might be interesting to study such conditions from a broader perspective. The connections to convex analysis introduced in Section 2.3.2 can potentially be leveraged by studying the convergence of Bregman divergences. Results of Bauschke et al. (2001, for example, Lemma 7.3(x)) may provide helpful starting points in this direction.

We have focused on procedures which approximate the score values by estimating the unknown underlying predictive distribution F_0 from the given sample. An alternative approach is to interpret the sample as a set of discrete predictions, and to use these forecasts directly to calculate the score value (Weigel, 2012). Fricker et al. (2013) propose the notion of a fair scoring rule for ensemble forecasts. A scoring rule is called *fair* if it is optimized for samples with members that behave as though they and the verifying observation were sampled from the same distribution. It can be demonstrated that the CRPS as defined in Table 4.2 is not fair in this sense. However, fair adjusted versions of the Brier score and the CRPS can be constructed by introducing terms that correct for the sample size m (Ferro et al., 2008). While certainly relevant in the context of meteorological forecast

ensembles where m is typically between 10 and 50, these considerations seem less helpful in the context of MCMC output. First, the sample size m is on the order of a few thousand and can be increased at low cost, rendering small sample corrections unimportant. Second, the proposed adjustments and the characterization of fair scores derived by Ferro (2014) only hold for independent samples, an assumption which is thoroughly violated in the case of MCMC.

In the chapter at hand, we are interested in evaluating forecasts produced via MCMC. This means that performance of a model during the out-of-sample (or test sample, or evaluation sample) period is used to estimate its forecast performance on future occasions. Information criteria (Spiegelhalter et al., 2002, 2014; Watanabe, 2010; Gelman et al., 2014b) suggest a different route towards estimating forecast performance. They consider a method's in-sample performance, and account for model complexity via a penalty term. The exact way of doing so has been the issue of much debate in the past, without clear implications for applied work. Our analysis does not concern in-sample comparisons, and thus does not provide evidence on whether these are more or less effective than out-of-sample comparisons. However, our results and observations indicate that out-of-sample comparisons yield robust results across a range of "reasonable" implementation choices and might therefore seem preferable in this regard.

Appendix 4.A Literature survey methodology and full list of references

To survey how probabilistic forecasts based on MCMC output are evaluated in the literature, we have attempted to conduct a systematic review. Note that efforts to do so are made difficult by various hurdles. Due to the relatively recent popularity of Bayesian forecasting methods, the literature notably lacks unified terminology and notation. Not only the employed implementation choices, but also the verification approaches and standard references vary a lot across different strands of applied literature and scientific disciplines. The literature survey presented below and the list of references in Table 4.5 should therefore not be viewed as necessarily objective or complete, but aim to provide a thorough overview of popular approaches.

In order to obtain a broad set of candidate articles for further consideration, we have performed Web of Science[4] and Google Scholar[5] searches for combinations of the terms "(probabilistic) forecast", "proper scoring rule", "CRPS" and "logarithmic score" with either "Bayesian", "MCMC", or combinations thereof. As Gneiting and Raftery (2007) has become a standard reference for the evaluation of probabilistic forecast and proper scoring rules, we additionally applied the above search queries to the 1 260 articles citing Gneiting and Raftery (2007) listed by Google Scholar[6] (as of February 25, 2016). This exploratory approach left us with approximately 200 articles for further review.

We applied the following selection criteria to the this preliminary set of candidate articles.

- We only consider articles published in scientific journals or books. In particular, working papers and preprints are excluded from the analysis.

- We only retain studies where forecasts based on Bayesian MCMC methods are produced and evaluated. Articles with a lack of formal forecast evaluation are excluded from the survey.

- Further, we restrict our attention to studies of real-valued data, and exclude articles that deal with binary and categorical observations.

- As we are interested in full probabilistic forecasts based on MCMC output, we disregard articles where only functionals of the forecast distributions such as mean or median values are evaluated. More specifically, we only keep articles where probabilistic forecasts are evaluated with proper scoring rules. Papers that are excluded by this criterion include Di Narzo and Cocchi (2010) where forecast evaluation is limited to the visual inspection of PIT histograms.

[4] `http://webofscience.com`
[5] `https://scholar.google.com/`
[6] `https://scholar.google.com/scholar?cites=11120728558307529279`

Table 4.4: Explanation of abbreviations used in Table 4.5

Abbreviation	Meaning
CRPS computation	
ECDF	Numerical integration of (2.3) where the empirical CDF takes the role of F
KDE	Kernel density estimation of the predictive density, combined with integration to obtain the predictive CDF
KR	Kernel representation, see equation (4.8)
MPE	Mixture of parameters estimator (4.2)
$\mathcal{N}$	Gaussian approximation (4.6)
LogS computation	
KDE	Kernel density estimation of the predictive density, see equation (4.12)
MPE	Mixture of parameters estimator (4.2)
$\mathcal{N}$	Gaussian approximation (4.6)

- The posterior predictive distribution is generally not available in a closed analytical form. Therefore, it is not obvious how to compute the value of proper scoring rules, and some approximation has to be applied. As we aim to understand how forecasts based on MCMC output are evaluated in the literature, we only retain studies where the computation of the scoring rules is explained in sufficient detail.

- Finally, we only consider the CRPS and LogS, as well as the weighted version thereof which were introduced in Chapter 3. Note that only very few articles apply scoring rules other than the CRPS and LogS. For example, Riebler et al. (2012) explicitly use the DSS, and Friederichs and Thorarinsdottir (2012) also consider the quadratic score.

Retaining only articles that meet the above selection criteria leaves us with the studies listed in Table 4.5. The abbreviations denoting the various approximation methods are explained in Table 4.4.

Table 4.5: Full list of 39 recently published studies using Bayesian probabilistic forecasts. For explanations of the abbreviations, see Table 4.4.

Reference	Scoring rule	Approx.
Adolfson et al. (2007, p. 323–325)	LogS	$\mathcal{N}$
Amisano and Giacomini (2007, p. 184)	LogS	MPE
Bauwens et al. (2014, p. 607)	LogS	KDE
Berg and Henzel (2015, p. 1078, 1084, 1089)	CRPS	Other
	LogS	KDE

Berrocal et al. (2014, p. 286)	CRPS	KR
Brandt et al. (2014, p. 949, 954)	CRPS	$\mathcal{N}$
Carriero et al. (2015a, p. 50)	LogS	$\mathcal{N}$
Carriero et al. (2015b, p. 20, 26)	LogS	$\mathcal{N}$
Carriero et al. (2015c, p. 848)	LogS	KDE
Carriero et al. (2015d, p. 331–332, 346)	LogS	KDE
Clark (2011, p. 331, 337)	LogS	$\mathcal{N}$
Clark and Ravazzolo (2015, p. 561)	CRPS	KR
	LogS	$\mathcal{N}$
De la Cruz and Branco (2009, p. 598, 602)	CRPS	KR
Delatola and Griffin (2011, p. 910–911)	LogS	MPE
Friederichs and Thorarinsdottir (2012, p. 581, 585)	CRPS	KR
Geweke and Amisano (2010, p. 219)	LogS	MPE
Geweke and Amisano (2011, p. 14, 20–21)	LogS	MPE
Giannone et al. (2015, p. 442)	LogS	$\mathcal{N}$
Groen et al. (2013, p. 34, 37)	CRPS	KR
Gschlößl and Czado (2007, p. 210, 214)	CRPS	KR
	LogS	MPE
Kallache et al. (2010, p. 5427, 5434)	CRPS	MPE
	LogS	MPE
Koop (2013, p. 180, 185, 199)	LogS	MPE
Krüger and Nolte (2015, p. 13–14)	CRPS	KDE
Krüger et al. (2015, p. 9, 19)	CRPS	ECDF
Leininger et al. (2013, p. 323–324)	CRPS	KR
Li et al. (2010, p. 3644, 3647)	LogS	MPE
Lopes et al. (2008, p. 767, 770, 775)	CRPS	KR
	LogS	MPE
Maneesoonthorn et al. (2012, p. 223, 227)	LogS	MPE
Panagiotelis and Smith (2008, p. 719)	CRPS	KR
Risser and Calder (2015, p. 292–293)	CRPS	MPE
	LogS	MPE
Rodrigues et al. (2014, p. 7914)	CRPS	$\mathcal{N}$
Sahu et al. (2015, p. 270)	CRPS	KR
Salazar et al. (2011, p. 594–595)	CRPS	KR
Sigrist et al. (2012, p. 1470)	CRPS	KR
Sigrist et al. (2015, p. 24)	CRPS	KR
Smith and Vahey (2015, p. 28, 36)	CRPS	ECDF
	twCRPS	ECDF
Tran et al. (2016, p. 371–372)	CSL	KDE
Trombe et al. (2012, p. 639, 650)	CRPS	MPE
Zhou et al. (2015, p. 10, 12–13)	LogS	MPE

Appendix 4.B Proof of consistency of the MPE

Here, we provide a proof of Proposition 4.1. Assume that

(A1) The sequence $\theta_1, \ldots, \theta_m$ is stationary and ergodic.

(A2) Ω_Y is a compact subset of $\mathbb{R}$.

(A3) The predictive density $f_c(z|\theta)$ conditional on θ is Lipschitz-continuous in z for all $\theta \in \Theta$.

(A4) The true predictive density f_0 is continuous and positive on Ω_Y.

Under (A1) and (A2), the mixture-of-parameters estimator (4.5) is consistent relative to the CRPS. Under (A1)–(A4), it is consistent relative to the LogS.

Proof. Regarding the CRPS, note that

$$\begin{aligned}\left\|\hat{F}_m - F_0\right\|_\infty &= \sup_{z\in\Omega_Y} \left| \int_{-\infty}^{z} \hat{f}_m(x) - f_0(x)\,\mathrm{d}x \right| \\ &\leq \sup_{z\in\Omega_Y} \int_{-\infty}^{z} \left| \hat{f}_m(x) - f_0(x) \right| \mathrm{d}x \\ &\leq \int_{\Omega_Y} \left| \hat{f}_m(x) - f_0(x) \right| \mathrm{d}x.\end{aligned}$$

The last term converges to zero by pointwise convergence of $\hat{f}_m(z)$ towards $f_0(z)$ and Scheffés Lemma. By (A2), this implies the desired convergence $\|\hat{F}_m - F_0\|_2 \to 0$ almost surely as $m \to \infty$.

Furthermore, note that

$$\begin{aligned}\sup_{z\in\Omega_Y} \left| \hat{f}_m(z) - f_0(z) \right| &= \sup_{z\in\Omega_Y} \left| \frac{1}{m}\sum_{i=1}^{m} f_c(z|\theta_i) - \int_\Theta f_c(z|\theta)\,\mathrm{d}P_{post}(\theta) \right| \\ &= \sup_{z\in\Omega_Y} \left| \int_\Theta f_c(z|\theta)\,[\mathrm{d}P_m(\theta) - \mathrm{d}P_{post}(\theta)] \right|\end{aligned}$$

where P_m denotes the empirical measure of $\{\theta_i\}_{i=1}^m$. Results from empirical process theory can be used together with suitably general versions of the strong law of large numbers to show that the latter term converges to zero under (A1) and (A3), see van der Vaart (2000, Chapter 19) for details on the involved Glivenko-Cantelli classes. By (A2) and (A4), and the connection between convergence of (4.14) and convergence of the Kullback-Leibler divergence (see Section 4.3.2), this implies consistency relative to the logarithmic score. □

Appendix 4.C Proof of consistency of ECDF-based approximations

Here, we provide a proof of Proposition 4.2. Assume that

(A1) The sequence $X_1, \ldots, X_m$ is stationary and ergodic

(A2) $\mathbb{E}_{F_0}|X_1| < \infty$.

Under (A1) and (A2), the empirical CDF $\hat{F}_m(z) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{X_1 \leq z\}$ is a consistent approximation relative to the CRPS, i.e.,

$$\int_{\mathbb{R}} \left(\hat{F}_m(z) - F_0(z)\right)^2 \mathrm{d}z \to 0$$

almost surely as $m \to \infty$.

Proof. We show the stronger result $\int_{\mathbb{R}} |\hat{F}_m(z) - F_0(z)| \, \mathrm{d}z \to 0$ almost surely as $m \to \infty$. With fixed $N \in \mathbb{N}$,

$$\begin{aligned}
\int_{\mathbb{R}} \left|\hat{F}_m(z) - F_0(z)\right| \mathrm{d}z &\leq \int_{-N}^{N} \left|\hat{F}_m(z) - F_0(z)\right| \mathrm{d}z + \int_{N}^{\infty} |1 - F_0(z)| + |1 - \hat{F}_m(z)| \mathrm{d}z \\
&\quad + \int_{N}^{\infty} |\hat{F}_m(-z)| + |F_0(-z)| \mathrm{d}z \\
&= \int_{-N}^{N} \left|\hat{F}_m(z) - F_0(z)\right| \mathrm{d}z + \int_{N}^{\infty} (1 - F_0(z)) + F_0(-z) \, \mathrm{d}z \\
&\quad + \int_{N}^{\infty} (1 - \hat{F}_m(z)) + \hat{F}_m(-z) \, \mathrm{d}z.
\end{aligned}$$

To simplify notation, let

$$\begin{aligned}
H(z) &= (1 - F_0(z)) + F_0(-z), \text{ and} \\
H_m(z) &= (1 - \hat{F}_m(z)) + \hat{F}_m(-z).
\end{aligned}$$

By the generalized Glivenko-Cantelli theorem (Dehling and Philipp, 2002, Theorem 1.1) and (A1),

$$\begin{aligned}
\limsup_{m\to\infty} \int_{\mathbb{R}} \left|\hat{F}_m(z) - F_0(z)\right| \, \mathrm{d}z &\leq \underbrace{\limsup_{m\to\infty} \int_{-N}^{N} \left|\hat{F}_m(z) - F_0(z)\right| \, \mathrm{d}z}_{\to 0 \text{ a.s. as } m\to\infty} \\
&\quad + \int_{N}^{\infty} H(z) \, \mathrm{d}z + \limsup_{m\to\infty} \int_{N}^{\infty} H_m(z) \, \mathrm{d}z
\end{aligned}$$

almost surely.

Note that $\int_N^\infty H(z) \, \mathrm{d}z = \mathbb{E}\left[(|X_1| - N)\mathbb{1}\{|X_1| \geq N\}\right]$, and by the ergodic theorem

$$\int_{N}^{\infty} H_m(z) \, \mathrm{d}z = \frac{1}{m}\sum_{i=1}^{m} (|X_i| - N)\mathbb{1}\{|X_i| \geq N\} \to \mathbb{E}\left[(|X_1| - N)\mathbb{1}\{|X_1| \geq N\}\right]$$

almost surely as $m \to \infty$, which along with (A2) implies that

$$\limsup_{m\to\infty} \int_{\mathbb{R}} \left|\hat{F}_m(z) - F_0(z)\right| \, \mathrm{d}z \leq 2 \underbrace{\mathbb{E}\left[(|X_1| - N)\mathbb{1}\{|X_1| \geq N\}\right]}_{\to 0 \text{ as } N\to\infty}$$

almost surely. □

5 Probabilistic wind speed forecasting based on ensembles

> Perhaps some day in the dim future it will be possible to advance the computations faster than the weather advances and at a cost less than the saving to mankind due to the information gained. But that is a dream.[1]
>
> Lewis Fry Richardson, 1922

Weather prediction is a key application of probabilistic forecasting. Many of the verification methods that have been discussed throughout this thesis originated from the meteorological literature, and ensemble prediction systems can be seen as mature and successful implementation of the paradigms of uncertainty quantification and probabilistic forecasting (Gneiting and Katzfuss, 2014). In this chapter, we investigate techniques for statistical postprocessing of forecast ensembles in order to correct for bias and dispersion errors. Proper scoring rules, and the methods and results from the preceding chapters thereby serve as valuable tools for parameter estimation and forecast evaluation. With a focus on wind speed, we investigate the choice of suitable parametric models.

5.1 Introduction

Reliable forecasts of wind speed are a necessity in a diverse number of applications such as agriculture, most modern means of transportation and wind energy production. Wind power, as a renewable and emissions free alternative to fossil fuels, has been growing rapidly over the last decade. In Europe, the wind power's share of total installed power capacity amounted to about 11.4% at the end of 2012 and it has increased five-fold since 2000 (European Wind Energy Association, 2012). For wind energy production, accurate forecasts of wind speed at different lead times are required to regulate electricity markets, to schedule maintenance and, more generally, to improve the competitiveness of wind power compared to sources of electricity which allow for dispatchable generation (Genton and Hering, 2007; Pinson et al., 2007; Lei et al., 2009). In many of these applications and for weather warnings, high wind speeds are of particular importance.

We focus on forecasts with medium-range lead times of a few days. In this setting, forecasts are usually based on outputs from numerical weather predic-

[1] Richardson (1922, p. vi)

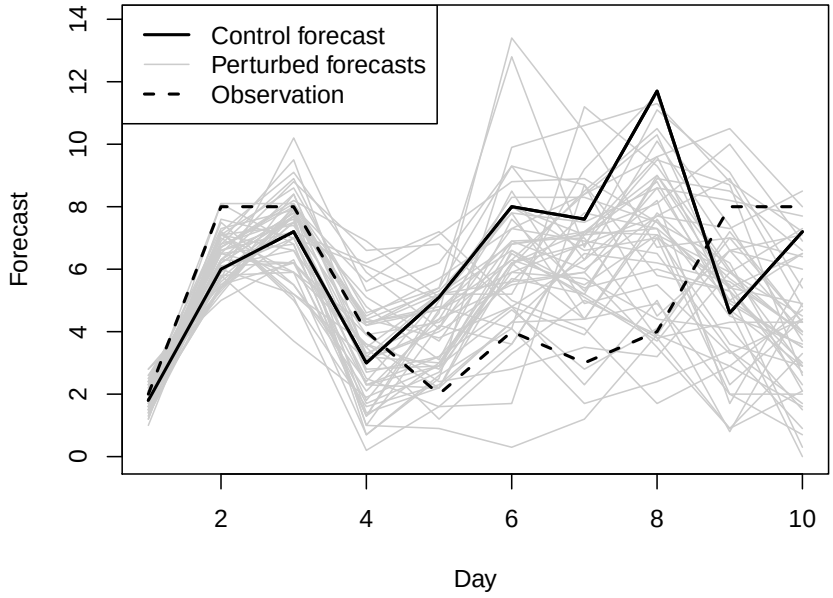


Figure 5.1: ECMWF ensemble forecasts of wind speed (in m s^{-1}) at the observation station located at Frankfurt airport, Germany, with lead times of 0–10 days, initialized on November 24, 2007. Each gray line corresponds to the forecast trajectory of one perturbed ensemble member.

tion (NWP) models which use physical descriptions of the atmosphere and oceans to propagate the state of the atmosphere forward in time based on the current weather conditions. They consist of sets of coupled hydro-thermodynamic nonlinear partial differential equations which do not have analytical solutions. NWP models are solved numerically on a spatial grid around the globe or smaller regions, discretized in time.

The basic idea of describing the atmosphere through a set of fundamental differential equations with the aim of numerical weather forecasting was developed more than a century ago and can be traced back at least to Bjerknes (1904). In the 1920s Lewis Fry Richardson put Bjerknes' ideas into practical use and produced a single six-hour ahead forecast of surface pressure at two locations by manually solving the differential equations using finite differences approaches, probably over the course of several months (Lynch, 2006). Following the rapid advancement of computer technology and the increased availability of observations, the first operational numerical weather forecasting systems were implemented in the 1950s and subsequently underwent tremendous improvement over the following decades. With the help of modern supercomputers, global weather forecasts for up to two weeks ahead are nowadays produced operationally, thereby making Richardson's dream quoted on the preface of this chapter reality. For details on the historical development of numerical weather prediction, see, e.g., Lynch (2006, 2008), Bauer et al. (2015).

Historically, single runs of NWP models with the best available initial con-

ditions were used to obtain single-valued predictions of the future state of the atmosphere. However, such deterministic, single-valued forecasts fail to account for uncertainties in the initial conditions and the numerical model. Following seminal work of Epstein (1969) and Leith (1974), and with the increased availability of computational resources, there has been a radical culture change over the last decades. NWP models are nowadays often run several times with different initial conditions and/or numerical representations of the atmosphere resulting in an *ensemble* of forecasts, as reviewed by Palmer (2002), Gneiting and Raftery (2005), and Leutbecher and Palmer (2008). Ensemble forecasts aim to provide an estimate of the uncertainty of the forecasts, and should ideally be interpretable as a random sample from the predictive distribution of future weather states (Gneiting, 2014).

Figure 5.1 shows an example of wind speed forecasts of the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble which are a part of the data used in Appendix 3.A. The control forecast is the model run with the best initial conditions, and the 50 perturbed forecasts are generated with slightly different model physics and with random perturbations in the initial conditions obtained with the singular vector method (Buizza and Palmer, 1995; Buizza et al., 1999; Gneiting, 2014). It can be observed that the forecast uncertainty represented by the spread of the ensemble predictions increases with the lead time. Although the observed wind speed trajectory clearly differs from the control forecast, it is contained within the set of possible scenarios provided by the ensemble predictions at all times.

Since the first operational implementations by the European Centre for Medium-Range Weather Forecasts (Buizza et al., 1993; Molteni et al., 1996; Buizza, 2006; Leutbecher and Palmer, 2008; ECMWF Directorate, 2012) and the National Centers for Environmental Prediction (NCEP; Toth and Kalnay, 1997), the generation of ensemble forecasts has become standard practice in meteorology. All major national meteorological services operate their own ensemble prediction system (EPS) as for example the Prévision d'Ensemble ARPege (PEARP; Descamps et al., 2014) EPS of Méteo France, or the Consortium for Small-scale Modeling (COSMO-DE; Gebhardt et al., 2011; Peralta et al., 2012) EPS of the German Meteorological Service. Other examples will be introduced in Sections 5.3.1 and 6.2. Related ensemble simulation techniques have also become popular in various other scientific disciplines, see, e.g., Adcock and McCammon (2006), Araújo and New (2007), Cloke and Pappenberger (2009), and Lozano et al. (2011).

The development of ensemble prediction systems plays a key role in the transition from deterministic to probabilistic weather forecasting and has become an established part of weather and climate prediction. As argued in Chapter 1 probabilistic forecasts are essential in many applications in that they allow for quantification of the associated prediction uncertainty. From a user perspective probabilistic forecasts further allow for optimal decision making since optimal deterministic forecasts can be obtained as functionals of the forecast distributions (Richardson, 2000; Krzysztofowicz, 2001; Gneiting, 2008, 2011). This is particularly important for applications such as wind power forecasting for auc-

tion processes in electricity markets where the optimal bidding strategy depends on permanently changing features of the market conditions, (Jeon and Taylor, 2012; Pinson et al., 2007; Pinson, 2013).

While the implementation of ensemble prediction systems is an important step in the transition from deterministic to probabilistic forecasting, ensemble forecasts are finite and do not provide full predictive distributions. Further, ensemble forecasts generally tend to be underdispersive and subject to systematic bias, and thus require some form of statistical postprocessing (Hamill and Colucci, 1997; Gneiting and Raftery, 2005).

Following seminal work of Hamill and Colucci (1997), a variety of statistical postprocessing techniques have been developed over the last two decades, for reviews and comparisons, see Wilks and Hamill (2007), Bröcker and Smith (2008), Schmeits and Kok (2010), Ruiz and Saulo (2012), Baran et al. (2014), Williams et al. (2014), and Gneiting (2014). State of the art techniques include Bayesian model averaging (BMA; Raftery et al., 2005) and ensemble model output statistics (EMOS) or non-homogeneous regression (Gneiting et al., 2005) which will be introduced in detail in Section 5.2.1. Both approaches rely on modeling the future distribution of a weather quantity through suitable parametric families of probability distributions, and thus involve the statistical estimation of parameters from training data.

The remainder of this chapter is organized as follows. Section 5.2 provides an introduction to state of the art approaches to statistical postprocessing, as well as parameter estimation approaches and verification methods for ensemble forecasts. In Section 5.3, we propose new EMOS models for postprocessing ensemble forecasts of wind speed based on generalized extreme value (GEV) and log-normal (LN) distributions as alternatives to the standard model which is based on truncated normal (TN) distributions. We further introduce new combination models that select one of the candidate distributions based on covariate information, as well as a mixture model that utilizes a weighted mixture of truncated normal and log-normal distributions. These novel EMOS models are compared in case studies based on three different ensemble prediction systems in Section 5.4, and are demonstrated to outperform the standard truncated normal model. The chapter combines three research articles (Lerch and Thorarinsdottir, 2013; Baran and Lerch, 2015, 2016). We conclude with a discussion in Section 5.5. The study of postprocessing approaches is continued in the following Chapter 6 where we propose new similarity-based semi-local approaches to estimating the parameters of the forecast distributions.

5.2 Statistical postprocessing of ensemble forecasts

The general goal of statistical postprocessing of ensemble forecasts is to correct for biases and dispersion errors in NWP model output. This section provides a general introduction to statistical postprocessing of ensemble forecasts.

5.2.1 Postprocessing approaches

Here, we review BMA and EMOS, two state of the art approaches.

5.2.1.1 Bayesian model averaging

BMA predictions are given by weighted mixtures of parametric densities or kernels each of which depends on a single ensemble member, with the mixture weights being determined by the performance of the ensemble members in the training period. Let y denote the weather variable of interest, and $x_1, \ldots, x_M$ the corresponding ensemble member forecasts. The predictive distribution in the BMA approach (Raftery et al., 2005) is a weighted mixture of the general form

$$y|x_1, \ldots, x_M \sim \sum_{i=1}^{M} w_i f(y|x_i), \tag{5.1}$$

where $f(y|x_i)$ is a suitably chosen parametric density that depends on the ensemble member x_i, and the weights $w_i \geq 0, i = 1, \ldots, M$ sum to 1. For temperature and pressure, Raftery et al. (2005) propose the use of a weighted mixture of normal distributions such that the kernel $f(y|x_i) = \phi(y|\mu, \sigma^2)$ is Gaussian with mean $\mu = a_{0i} + a_{1i}x_i$ and variance $\sigma^2 = \sigma_0^2$. The basic BMA model (5.1) can be adapted to various weather variables by choosing suitable parametric kernels $f(y|x_i)$, and by extending the link functions connecting the ensemble predictions and the parameters of the forecast density. BMA implementations are available for a variety of univariate weather variables such as precipitation (Sloughter et al., 2007), wind direction (Bao et al., 2010) or visibility (Chmielecki and Raftery, 2011), see Gneiting (2014) for an overview. `R` implementations of some basic BMA models are provided by the `ensembleBMA` package (Fraley et al., 2011, 2015).

Sloughter et al. (2010) propose a BMA model for wind speed based on gamma densities that has been applied by Baran et al. (2013) and Courtney et al. (2013), among others, whereas Baran (2014) considers truncated normal component densities. Bivariate BMA models for wind vectors have been studied by Sloughter et al. (2013).

5.2.1.2 Ensemble model output statistics

The ensemble model output statistics or non-homogeneous regression approach (Gneiting et al., 2005) is conceptually simpler. The predictive distribution is given by a single parametric distribution with parameters depending on the ensemble members, and has the general form

$$y|x_1, \ldots, x_M \sim f(y|x_1, \ldots, x_M), \tag{5.2}$$

where again, y denotes the future weather quantity of interest and $x_1, \ldots, x_M$ are the corresponding ensemble forecasts. The parametric density $f(y|x_1, \ldots, x_M)$ depends on the ensemble predictions through suitably chosen link functions. For

temperature and pressure, Gneiting et al. (2005) propose an EMOS model based on Gaussian predictive distributions where $y|x_1, \ldots, x_M \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = a_0 + \sum_{i=1}^{M} b_i x_i$ and $\sigma^2 = c + d\, S^2$, where S^2 denotes the variance of the ensemble forecasts. See also Hagedorn et al. (2008) and Kann et al. (2009) for further applications of this model.

The basic Gaussian EMOS model for temperature and pressure is implemented in the `R` package `ensembleMOS` (Fraley et al., 2011; Yuen et al., 2013). Over the last years, EMOS models for various weather variables such as precipitation (Scheuerer, 2014; Scheuerer and Hamill, 2015a) or cloud cover (Hemri et al., 2016) have been proposed, see Gneiting (2014) for an overview.

The first EMOS model for wind speed was proposed by Thorarinsdottir and Gneiting (2010) and utilizes truncated normal distributions. In Section 5.3, we will introduce this model in detail and propose several alternatives. The truncated normal model for wind speed has been extended to wind gusts by Thorarinsdottir and Johnson (2012). Pinson (2012) and Schuhen et al. (2012) study bivariate EMOS models for statistical postprocessing wind vector ensembles.

5.2.1.3 Other approaches

There exist various other approaches to postprocessing ensemble forecasts, however, many of those techniques can be viewed within the framework of BMA and EMOS (Gneiting, 2014). For example, ensemble dressing approaches proposed by Roulston and Smith (2003) and Wang and Bishop (2005) are directly related to Bayesian model averaging, and the logistic regression approach Wilks (2009) and subsequent extensions Messner et al. (2014a,b) have close connections to the EMOS framework (Roulin and Vannitsem, 2012; Scheuerer, 2014; Gneiting, 2014).

EMOS models generally are more parsimonious, whereas BMA models tend to be more flexible. The predictive performance of BMA and EMOS models is typically comparable as illustrated in various case studies. In the remainder of this chapter, we focus on the EMOS approach and investigate alternative parametric models for wind speed in Section 5.3.

In the general formulations of the BMA and EMOS models in equations (5.1) and (5.2) we have implicitly assumed that the forecast distributions only depend on the ensemble predictions $x_1, \ldots, x_M$ of the weather quantity of interest to simplify notation. It is of course possible to extend the model formulations to include other outputs of the NWP model or covariate information, see, e.g., Kleiber et al. (2011). Possible extensions in this direction will be discussed in Section 6.5.

5.2.2 Parameter estimation

Both BMA and EMOS models require the estimation of statistical parameters of the forecast distributions. For example, in the case of the Gaussian EMOS model

for temperature and pressure introduced above,

$$y|x_1, \ldots, x_M \sim \mathcal{N}\left(a_0 + \sum_{i=1}^{M} b_i x_i, c + d\, S^2\right),$$

the parameters of the EMOS model (or *EMOS coefficients*) $a_0, b_1, \ldots, b_M, c, d$ connecting the ensemble predictions and the parameters of the forecast distribution have to be estimated.

In statistical postprocessing, the parameters are typically estimated by minimizing the mean values of proper scoring rules over suitably chosen training sets of past pairs of forecasts and observations. This optimum score estimation approach was proposed by Gneiting et al. (2005), see also Section 2.3.3. In this light, classical maximum likelihood (ML) estimation of the parameters corresponds to optimum score estimation based on minimizing the mean logarithmic score (2.2) (Gneiting et al., 2005; Raftery et al., 2005; Wilks, 2011). As argued by Gneiting et al. (2005) and summarized in Section 2.3.3, optimum score estimation can be viewed within the framework of M-estimation and asymptotic results such as consistency therefore apply for a general class of proper scoring rules including the CRPS.

From an applied perspective, a key difference between parameter estimation based on the LogS and the CRPS is that the logarithmic score assigns high penalties to poor probabilistic forecasts, for example in case of outliers or extreme events. Optimum score estimation based on minimizing the CRPS leads to more robust estimation procedures and the CRPS is thus often seen as the more appropriate scoring rule in practical applications, see Gneiting et al. (2005) for a summary. Recall that the CRPS is given by

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(z) - \mathbb{1}\{y \leq z\}\right)^2 \mathrm{d}z. \tag{5.3}$$

Computationally efficient parameter estimation based on minimizing the CRPS requires that an analytical expression of the integral in (5.3) is available. Such closed form solutions of the integral have been derived for a variety of distributions (for an overview see, e.g., Jordan, 2015), many of which are implemented in the `scoringRules` package (Jordan et al., 2016). In Section 5.3.3, we will discuss different estimation approaches for EMOS models for wind speed in more detail.

The training data for estimating the parameters are given by past pairs of ensemble forecasts and corresponding observations. In the remainder of the chapter at hand, we consider observations to be physical observations made at manned or automated weather stations. Alternatives are analysis data produced by NWP models which are given on a model grid and therefore have the advantage of being available at locations with few or no weather stations and actual observations, for example over the oceans. However, analysis data are generally not independent of the forecast models and can share bias or other peculiarities, see Hagedorn (2010), Hagedorn et al. (2012), and Gneiting (2014) for discussions of these issues.

The training data are typically chosen to be rolling training periods consisting of forecasts and observations from the preceding n days. In general, shorter training periods allow for a rapid adaption to changes in environmental conditions while longer training periods reduce the statistical variability in the parameter estimation (Gneiting et al., 2005). There is no automated way to determine the optimal training period length, therefore, models are usually fitted several times with varying choices of n, and the effect on the predictive performance is examined, see Section 5.3.3.

Another important decision is the spatial composition of the training set. Two basic approaches are given by local and regional methods. In the local approach, only forecast cases from the single observation station of interest are considered for the parameter estimation, whereas in the regional approach, data from all available observation stations are composited to form a single training set for all stations. Local estimation allows to account for locally varying forecast errors and generally results in better predictive performance (see, e.g., Thorarinsdottir and Gneiting, 2010; Schuhen et al., 2012), however, is problematic if only limited amounts of training data are available. In Chapter 6 we discuss issues of local and regional parameter estimation in more detail and propose alternative similarity-based approaches where the training data for a specific station are augmented with corresponding data from stations with similar characteristics.

5.2.3 Verification of ensemble forecasts

Proper scoring rules provide summary measures of the predictive performance of probabilistic forecasts and are important measures of the forecast quality of ensemble predictions. As discussed in Chapter 4, the CRPS of an ensemble forecast can be computed by replacing the predictive CDF F with the empirical CDF of the ensemble predictions $x_1, \ldots, x_M$. Proper scoring rules based on forecast densities such as the LogS (2.2) and the Hyvärinen score (2.6) are impractical. The approximations and asymptotic considerations from Chapter 4 can not be applied here as the complex dependence structure in the sample of ensemble predictions is unknown, and the number of ensemble members M is generally small. As mentioned above, fair versions of proper scoring rules may be of interest to compare forecast ensembles of different sizes and assess the effect of the number of ensemble members on the predictive performance (Ferro et al., 2008; Fricker et al., 2013; Ferro, 2014).

Anderson (1996) and Hamill and Colucci (1997) propose verification rank (VR) histograms as a graphical tool to assess the calibration of ensemble forecasts. VR histograms show the distribution of the ranks of the observations when pooled within the ordered ensemble predictions. For a calibrated ensemble, the observations and the ensemble predictions should be exchangeable, resulting in a uniform VR histogram. VR histograms can be seen as discrete analogues of PIT histograms that have been introduced in Section 2.2. Therefore, they share the respective indications of potential reasons of miscalibration that can be derived from deviations from the desired uniform distribution.

Further, the use of average coverage and width of central prediction intervals to assess calibration and sharpness of probabilistic forecast was introduced in Section 2.2. In the case of an ensemble with M members, $\frac{M-1}{M+1}100\,\%$ central prediction intervals correspond to the nominal coverage of the ensemble and allow for a direct comparison of full probabilistic forecasts obtained via statistical postprocessing.

5.3 EMOS models for probabilistic wind speed forecasting

In this section, we introduce EMOS models for postprocessing ensemble forecasts of wind speed based on Lerch and Thorarinsdottir (2013), Baran and Lerch (2015), and Baran and Lerch (2016). The new models are investigated based on three data sets with different ensemble prediction systems and forecast domains.

Due to the importance of accurate and reliable wind speed predictions illustrated in Section 5.1, statistical modeling and probabilistic forecasting of wind speed have received considerable attention over the last decades. Hourly average wind speeds are usually modeled using log-normal, gamma (Garcia et al., 1998), Rayleigh, Weibull (Justus et al., 1978; Seguro and Lambert, 2000; Celik, 2004) or truncated normal distributions (Gneiting et al., 2006). Generalized extreme value distributions have been employed for modeling maxima of wind and gust speed over a single day (Friederichs and Thorarinsdottir, 2012) or over long return periods, typically 50 years (Palutikof et al., 1999).

The standard EMOS model for postprocessing ensemble forecasts of wind speed proposed by Thorarinsdottir and Gneiting (2010) was originally developed for daily maximum wind speed and utilizes truncated normal distributions. However, as we will demonstrate below, the TN model often fails to account for the skewness and heavy right tails of the distribution of wind speed, and therefore frequently fails to accurately predict high wind speed values. We therefore propose alternative EMOS models based on distributions with heavier right tails. In particular, we propose models that employ generalized extreme value and log-normal distributions. We further investigate flexible regime-switching combination and weighted mixture model approaches which combine advantages of lighter and heavier-tailed distributions. In three case studies on ensemble prediction systems with different properties, forecast domains, and observed wind quantities, the proposed EMOS models are able to consistently outperform the TN model and provide calibrated and skillful forecasts.

The remainder is organized as follows. In Section 5.3.1, the three data sets of ensemble forecasts and observations are introduced. Further, we introduce the notion of exchangeability in ensemble forecasting which is of importance for the formulation of postprocessing models. In Section 5.3.2, we extend the EMOS model (5.2) to wind speed and to ensemble prediction systems with exchangeable members, and review the TN model of Thorarinsdottir and Gneiting (2010). Further, we introduce new EMOS models based on GEV and LN distributions, as well as combinations and weighted mixtures with the TN model. Section 5.3.3

contains a description of the parameter estimation methods employed for the various models. The results of the three case studies are reported in Section 5.4.

5.3.1 Data

We consider three distinct data sets of ensemble forecasts and corresponding observations which differ both in the stochastic properties of the ensemble as well as the observed wind quantities. Outside of the present Section 5.3.1, we use the general term wind speed to denote the respective technical definitions of the wind quantities in the different data sets given below in order to increase readability.

An important notion which will reappear throughout the present chapter is that of exchangeable ensemble members. Ensemble members are called *exchangeable* if they differ only in random perturbations and are therefore statistically indistinguishable. For example, the 50 perturbed members of the ECMWF ensemble are generated with random perturbations in initial conditions and model physics, and can thus be regarded as exchangeable. On the other hand, if ensemble members are individually distinguishable, for example if they are generated with varying physical NWP models and thus exhibit systematic differences, they are not exchangeable. An example of an EPS with non-exchangeable members is the COSMO-DE EPS of the German Weather Service introduced above.

The presence of exchangeable ensemble members is important for specifying the link functions connecting the EMOS coefficients and the parameters of the predictive distribution as exchangeable members should share the same coefficient values. For a detailed discussion of the notion of exchangeability in ensemble forecasts, see Fraley et al. (2010), Bröcker and Kantz (2011), and Ferro (2014), among others.

5.3.1.1 ECMWF ensemble

We consider 50 ensemble member forecasts of near-surface (10-meter) wind speed obtained from the global ensemble prediction system of the ECMWF. Ensemble forecasts for lead times up to 10 days ahead are issued twice a day at 00 UTC and 12 UTC, with a horizontal resolution of about 33 km and a temporal resolution of 3-6 hours. To account for uncertainties in the initial conditions and the numerical model, the ensemble members are generated from random perturbations in initial conditions and stochastic physics parametrization (Molteni et al., 1996; Leutbecher and Palmer, 2008; Pinson and Hagedorn, 2012). The ensemble members are thus statistically indistinguishable and can be treated as exchangeable (Fraley et al., 2010). We restrict attention to the ECMWF ensemble run initialized at 00 UTC and lead times of 1 day. To obtain predictions of daily maximum wind speed, we take the daily maximum of each ensemble member at each grid point location. For instance, one day ahead forecasts are given by the maximum over lead times of $3, 6, \ldots, 24$ hours.

The forecasts are verified over a set of 228 synoptic observation stations over Germany, see Figure 5.2. All maps in this chapter were produced using the

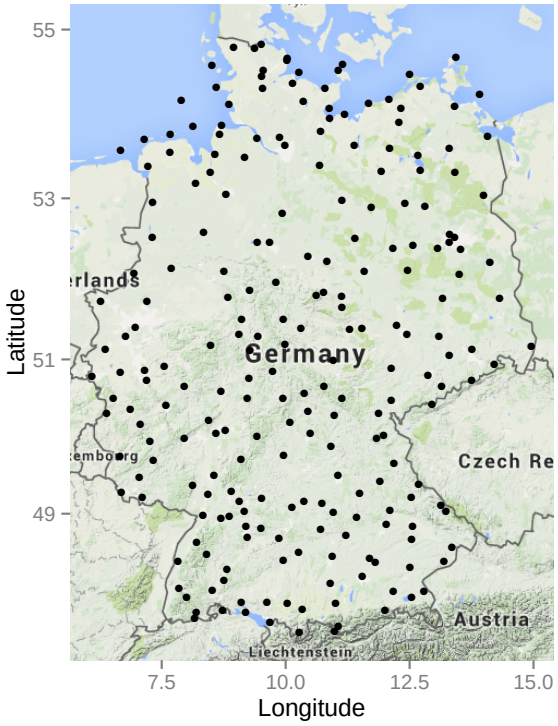


Figure 5.2: Map of Germany showing the locations of the 228 synoptic observation stations over which ECMWF ensemble forecasts are evaluated.

`ggmap` package for `R` (Kahle and Wickham, 2013). The observations are hourly observations of 10-minute average wind speed which is measured over the 10 minutes before the hour, and were provided by Michael Scheuerer. To obtain daily maximum wind speed, we take the maximum over the 24 hours corresponding to the time frame of the ensemble forecast. Ensemble forecasts at individual stations are obtained by bilinear interpolation of the gridded model output. The results presented below are based on a verification period from May 1, 2010 to April 30, 2011, consisting of 83 220 individual forecast cases. Additionally, we use data from February 1, 2010 to April 30, 2011 to obtain training periods of equal lengths for all days in the verification period and for model selection purposes.

Note that the data set used here differs from the ECMWF forecasts and observations at Frankfurt airport previously used in Sections 3.A and 5.1 which were taken from a distinct data set that also includes a control forecast and observations at weather stations around the globe for a longer time period, see Hemri et al. (2014).

5.3.1.2 ALADIN-HUNEPS ensemble

The Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System (ALADIN-HUNEPS; Horányi et al., 2006; Hagel, 2010) is the operational limited area model EPS of the Hungarian Meteorological Service. It covers large parts of continental Europe with a horizontal resolution of

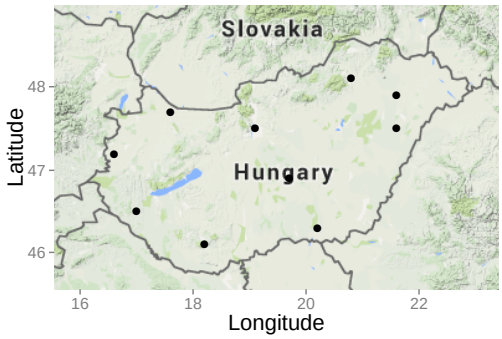


Figure 5.3: Map of Hungary showing the locations of the observation stations over which the ALADIN-HUNEPS forecasts are evaluated.

12 km and is obtained with dynamical downscaling by the ALADIN limited area model of the global ARPEGE based PEARP system of Météo France (Descamps et al., 2014). The ensemble consists of 11 members, one control forecast from the unperturbed analysis and 10 members initialized from perturbed initial conditions. The 10 members of the latter group can be regarded as exchangeable since they are generated with random perturbations in initial conditions.

The data considered here contains 42 hour ahead forecasts of 10-meter instantaneous wind speed for 10 major cities in Hungary (shown in Figure 5.3) together with the corresponding validating observations from April 1, 2012 to March 31, 2013. The validating observations were scrutinized by basic quality control algorithms including consistency checks. The forecasts are initialized at 18 UTC. Six days without available forecasts are excluded from the analysis.

The wind speed observations are considered as instantaneous values (valid at a given time), however, they are in fact mean values over the preceding 10 minutes. The ensemble forecasts of the NWP model are also considered as instantaneous wind speed values, but are representatives for a given model time step, which is 5 min in our case. The data were provided by Mihály Szűcs from the Hungarian Meteorological Service.

5.3.1.3 University of Washington Mesoscale Ensemble

The University of Washington Mesoscale Ensemble (UWME; Eckel and Mass, 2005) covers the Pacific Northwest region of western North America and has eight members which are obtained from different runs of the fifth generation Pennsylvania State University–National Center for Atmospheric Research mesoscale model with initial conditions from different sources (Grell et al., 1995).

The ensemble forecasts are initialized at 00 UTC and are given on a 12 km grid. Our data set contains ensembles of 48 hour ahead forecasts and corresponding validating observations of 10-meter maximal wind speed (maximum of the hourly instantaneous wind speeds over the previous twelve hours, given in m s^{-1}, see Sloughter et al. (2010)) for 152 stations in the Automated Surface Observing Network (National Weather Service, 1998), see Figure 5.4. The ensemble

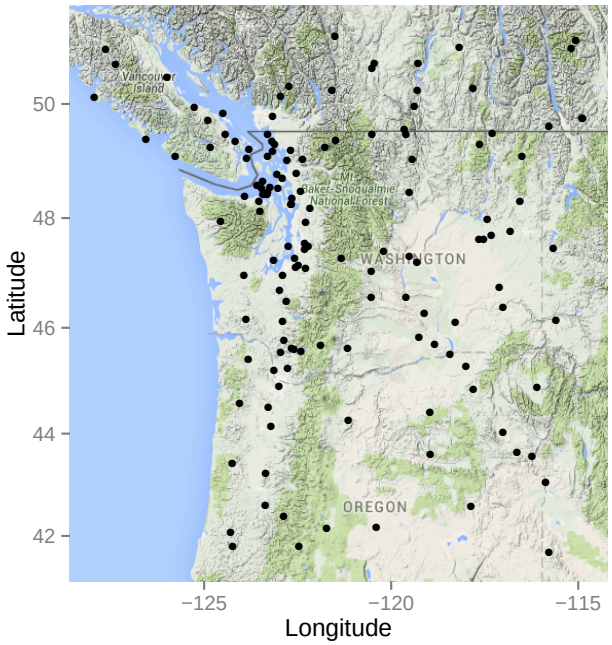


Figure 5.4: Map of the U.S. Pacific Northwest showing the locations of the observation stations over which UWME forecasts are evaluated.

members are not exchangeable as they are generated with initial conditions from different sources and thus are statistically distinguishable.

We investigate forecasts for calendar year 2008 with additional data from the last month of 2007 used for parameter estimation. After removing days and locations with missing data 101 stations remain where the number of days for which forecasts and validating observations are available varies between 160 and 291. The total number of forecast cases in the verification period is 27 481. The data were provided by Annette Möller and have also been used in Möller et al. (2013).

5.3.2 EMOS models

In what follows, we consider three different types of EMOS models for wind speed. Standard EMOS models based on a single parametric family extending the model (5.2) to wind speed are introduced in Section 5.3.2.1 where we review the EMOS model based on truncated normal distributions of Thorarinsdottir and Gneiting (2010) and propose novel models based on generalized extreme value and log-normal distributions.

In Section 5.3.2.2, we extend these basic models and introduce combination models that select one of the candidate distributions based on covariate information. As an alternative approach to combining lighter and heavier tailed distributions a weighted mixture model based on truncated normal and log-normal components is proposed in Section 5.3.2.3.

In Section 5.3.1, we have introduced the notion of exchangeable ensemble mem-

bers which effects the formulation of EMOS models as exchangeable members should share the same EMOS coefficients. To account for the different stochastic properties of the investigated ensembles, we specify suitable link functions connecting the ensemble forecasts and the parameters of the predictive distributions.

5.3.2.1 Models based on single parametric families

Truncated normal model

Let y denote wind speed and $x_1, \ldots, x_M$ the corresponding ensemble member forecasts. The predictive distribution for y is given by a truncated normal distribution with a cutoff at 0,

$$y|x_1, \ldots, x_M \sim \mathcal{N}_{[0,\infty)}(\mu, \sigma^2), \tag{5.4}$$

where the location parameter $\mu = a + b_1 x_1 + \cdots + b_M x_M$ is an affine function of the ensemble forecasts and the variance $\sigma^2 = c + dS^2$ is an affine function of the ensemble variance $S^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \bar{x})^2$ with $\bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i$. The cumulative distribution function of the TN distribution is given by

$$F_{\mathrm{TN}}(z) = \Phi\left(\frac{\mu}{\sigma}\right)^{-1} \Phi\left(\frac{z-\mu}{\sigma}\right)$$

for $z > 0$, and 0 otherwise, and the corresponding density function is

$$f_{\mathrm{TN}}(z) = \frac{\frac{1}{\sigma}\varphi\big((z-\mu)/\sigma\big)}{\Phi\big(\mu/\sigma\big)}, \quad z \geq 0,$$

and $f_{\mathrm{TN}}(z) = 0$ otherwise. Here Φ and φ denote the cumulative distribution function and density function of the standard normal distribution, respectively. The TN model was proposed by Thorarinsdottir and Gneiting (2010).

The above formulation of the link functions for the parameters μ and σ^2 assumes that the ensemble members are individually distinguishable and thereby non-exchangeable. This is the case for the UWME data where we employ this model. As the ECMWF ensemble members are exchangeable, we adapt the link function for the location parameter μ and assume that $b_1 = \cdots = b_M$, or $\mu = a + b\bar{x}$. The ALADIN-HUNEPS predictions consist of one control member x_{c} initialized with the best available initial conditions, and 10 members $x_1, \ldots, x_{10}$ which are generated with random perturbations. There thus exist two groups of ensemble members, where the members within each group are exchangeable should receive the same coefficient of the location parameter. We therefore set $\mu = a + b_c x_{\mathrm{c}} + b_1 \sum_{i=1}^{10} x_i$. More general formulations for an arbitrary number of groups of exchangeable members will be introduced in Chapter 6.

In the above link functions we have always defined the variance parameter σ^2 to be an affine function of the ensemble variance. Alternative formulations that take into account the existence of exchangeable members and groups have been investigated, but result in a reduced predictive performance.

Generalized extreme value model

As an alternative to the TN model in (5.4), we consider a model based on extreme value theory (Lerch and Thorarinsdottir, 2013). The cumulative distribution function of the generalized extreme value distribution (see, e.g., Coles, 2001) with location parameter μ, scale parameter σ and shape parameter ξ is given by

$$F_{\mathrm{GEV}}(z) = \begin{cases} \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, & \xi \neq 0 \\ \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}, & \xi = 0. \end{cases} \tag{5.5}$$

This distribution is defined on the set $\{z \in \mathbb{R} : 1+\xi(z-\mu)/\sigma > 0\}$, where the parameters satisfy $\mu, \xi \in \mathbb{R}$ and $\sigma > 0$. For $\xi > 0$, F_{GEV} is of Fréchet type with a heavy right tail and it holds that $z \in [\mu - \sigma/\xi, \infty)$. We obtain the Fréchet type in almost all of our forecast cases in the three data sets. We estimate the parameters of the model in (5.5) without any constraints on the parameter values. It is thus possible to obtain non-zero probabilities of negative wind speed. However, we find that this rarely happens in practice for the investigated data sets as we will discuss below.

The GEV model was first proposed by Lerch and Thorarinsdottir (2013), where it was applied to the ECMWF data. To link the parameters of the predictive GEV distribution to the ensemble, we apply the Bayesian covariate selection algorithm described in Friederichs and Thorarinsdottir (2012) to the data from February 1, 2010 to April 30, 2010. In this analysis, we assume a constant shape parameter ξ while the location μ and the scale σ may depend on the ensemble mean and variance,

$$\mu = \mu_0 + \kappa_1\mu_1\bar{x} + \kappa_2\mu_2 S^2, \qquad \log(\sigma) = \sigma_0 + \nu_1\sigma_1\bar{x} + \nu_2\sigma_2 S^2,$$

where $\mu_i, \sigma_i \in \mathbb{R}$ for $i = 0, 1, 2$ and $\kappa_i, \nu_i \in \{0, 1\}$ for $i = 1, 2$. For 100 000 iterations of the Metropolis within Gibbs algorithm with a burn-in period of 20 000 iterations, we obtain very high posterior inclusion probabilities for the mean ensemble forecast $\bar{x}$ while $\kappa_2 = 1$ or $\nu_2 = 1$ holds for less than 0.1% of the posterior sample for each parameter. In our subsequent predictions for the test set from May 1, 2010 to April 30, 2011, we thus set $\mu = \mu_0 + \mu_1\bar{x}$ and $\sigma = \sigma_0 + \sigma_1\bar{x}$ under the constraint that $\sigma > 0$, as the results of Friederichs and Thorarinsdottir (2012) indicate that an identity link on σ results in minimally improved performance compared to the logarithmic link for the estimation procedure described in Section 5.3.3 below.

We employ similar types of link functions in the applications of the GEV model for the ALADIN-HUNEPS and UWME data. In the case of the groups of exchangeable members in the ALADIN-HUNEPS forecasts, we set $\mu = \mu_0 + \mu_c x_c + \mu_1\bar{x}$ and $\sigma = \sigma_0 + \sigma_c x_c + \sigma_1\bar{x}$, where x_c again denotes the control forecast. The members of the UWME are non-exchangeable, we therefore set $\mu = \mu_0 + \sum_{i=1}^{M} \mu_i x_i$ and $\sigma = \sigma_0 + \sigma_1\bar{x}$. Alternative full models for the scale parameter that include the individual ensemble forecasts $x_1, \ldots, x_M$ rather than just the ensemble mean $\bar{x}$

have been investigated, but result in a reduction of predictive performance. Similarly, incorporating the ensemble variance into the link functions did not result in improvements for the ALADIN-HUNEPS and UWME data.

Log-normal model

As an alternative to the TN and GEV models, we further propose an EMOS model based on a log-normal distribution (Baran and Lerch, 2015). The LN distribution has heavier right tails than the TN distribution and is thus more appropriate to model high wind speed values. In comparison to the GEV model, the LN distribution avoids positive mass on negative wind speed values, and allows for a numerically more stable estimation of the parameters based on minimizing the CRPS, see Section 5.3.3. The LN distribution with location parameter μ and shape parameter $\sigma > 0$ has cumulative distribution function

$$F_{\mathrm{LN}}(z) = \Phi\left(\frac{\log(z) - \mu}{\sigma}\right)$$

and density

$$f_{\mathrm{LN}}(z) = \frac{1}{z\sigma}\varphi\big((\log z - \mu)/\sigma\big), \quad z \geq 0,$$

and $f_{\mathrm{LN}}(z) = 0$, otherwise, with φ and Φ denoting the PDF and CDF of the standard normal distribution. Mean m and variance v of the log-normal distribution are

$$m = \mathrm{e}^{\mu+\sigma^2/2} \qquad \text{and} \qquad v = \mathrm{e}^{2\mu+\sigma^2}\big(\mathrm{e}^{\sigma^2} - 1\big),$$

respectively. Further, since

$$\mu = \log\left(\frac{m^2}{\sqrt{v+m^2}}\right) \qquad \text{and} \qquad \sigma = \sqrt{\log\left(1 + \frac{v}{m^2}\right)},$$

the LN distribution can be readily expressed in terms of mean and variance by the above transformations.

Similar to the TN model, we take m and v to be affine functions of the ensemble forecasts and their variance, respectively. In case of distinguishable ensemble members, e.g., for the UWME data, we set $m = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_M x_M$. Exchangeable members should again share the same EMOS coefficients, we therefore set $m = \alpha_0 + \alpha_1 \bar{x}$ in case of the ECMWF data, and $m = \alpha_0 + \alpha_c x_c + \alpha_1 \bar{x}$ in case of the ALADIN-HUNEPS forecasts. For all three ensemble prediction systems, we model the variance v as an affine function of the ensemble variance, i.e., $v = \beta_0 + \beta_1 S^2$.

To illustrate the differences between the three models, Figure 5.5 shows the predictive distributions for examples from all three data sets. All models correct the bias and the underdispersion of the ensemble. Compared to the TN model, the GEV and LN densities are less symmetric and exhibit a heavy right tail.

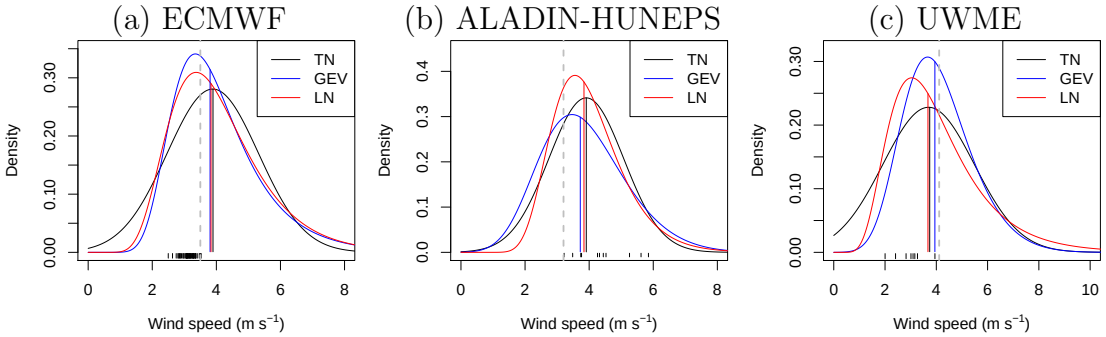


Figure 5.5: Illustration of predictive distributions for the three data sets, with observations valid at (a) Frankfurt airport on March 19, 2011; (b) Budapest on August 25, 2012, and (c) Seattle Tacoma International Airport on May 1, 2008. The ensemble forecasts are indicated by black line segments, and the observation is shown as vertical gray dashed line. The solid lines under the TN, GEV and LN predictive distributions indicate the respective median value.

5.3.2.2 Regime-switching combination models

The second type of models we consider are regime-switching methods which combine the TN model and GEV or LN models. These models aim to combine the good performance of the TN model for low and medium wind speeds with the better performance of the heavy-tailed distributions for higher wind speeds.

Conditional on the median of the ensemble predictions,

$$x^{med} = \text{median}(x_1, \ldots, x_M),$$

we either issue a TN or a heavy-tailed predictive distribution independently at each station. That is, for a model threshold $\theta \in \mathbb{R}_+$, we define the predictive distribution of the TN-GEV combination model by

$$G = \begin{cases} \mathcal{N}_{[0,\infty)}\left(\mu^{\mathcal{N}}, \sigma^{2\mathcal{N}}\right), & \text{if } x^{med} < \theta \\ \text{GEV}\left(\mu^{\text{GEV}}, \sigma^{\text{GEV}}, \xi^{\text{GEV}}\right), & \text{if } x^{med} \geq \theta. \end{cases} \tag{5.6}$$

The corresponding predictive distribution of the TN-LN combination model with threshold $\tilde{\theta}$ is given by

$$G = \begin{cases} \mathcal{N}_{[0,\infty)}\left(\mu^{\mathcal{N}}, \sigma^{2\mathcal{N}}\right), & \text{if } x^{med} < \tilde{\theta} \\ \text{LN}\left(\mu^{\text{LN}}, \sigma^{2\,\text{LN}}\right), & \text{if } x^{med} \geq \tilde{\theta}. \end{cases} \tag{5.7}$$

The parameters of the individual distributions depend on the ensemble forecasts as described above. However, we train the TN model only on training data for which it holds that $x^{med} < \theta$ or $x^{med} < \tilde{\theta}$, respectively. Similarly, the parameters of the heavy-tailed distribution are learned from data where $x^{med} \geq \theta$ or $x^{med} \geq \tilde{\theta}$, respectively. The model thresholds $\theta, \tilde{\theta}$ are selected by comparing predictive performance over a range of possible thresholds based on out-of-sample

data. In what follows, we will use the common symbol θ to denote the model selection threshold for both the TN-GEV and the TN-LN combination models to simplify notation.

As we will see below, the combination models are able to outperform the simple models based on a single parametric family. However, they suffer from the obvious drawback that a suitable covariate has to be chosen as a selection criterion. Although the ensemble median works well for all three data sets, this necessary step may limit the flexibility of the combination models in practice as the adequacy of covariates might depend on the data at hand.

5.3.2.3 Weighted mixture models

In order to combine the advantages of lighter and heavier-tailed distributions and to avoid the aforementioned problems in the process, we introduce a third type of EMOS models based on weighted mixtures of two parametric distributions. In particular, we propose to model wind speed with a weighted mixture of truncated normal and log-normal distributions with density

$$h(z) = w f_{\mathrm{TN}}(z) + (1 - w) f_{\mathrm{LN}}(z), \tag{5.8}$$

where $w \in [0, 1]$, and the parameters of the truncated normal and log-normal distributions depend on the ensemble members as specified above in Section 5.3.2.1.

Note that instead of a log-normal distribution, other non-negative distributions with heavy right tails can be incorporated in (5.8). A natural choice is the generalized Pareto distribution (GPD) used in extreme value theory (see, e.g., Frigessi et al., 2002; Bentzien and Friederichs, 2012), however, tests for the ensemble forecasts considered here indicate a worse predictive performance of the TN-GPD model compared to the TN-LN mixture.

The mixture models exhibit desirable properties from a theoretical perspective as they do not require the exclusive choice of one of multiple parametric families and are more flexible than models based on single parametric distributions. Their advantages from a practical perspective such as a significantly improved calibration will be demonstrated below. On the other hand, however, the parameter estimation based on minimizing the CRPS for such mixture models is computationally much more demanding since there are no analytical closed-form expressions of the integral in (5.3) available.

5.3.3 Estimation details

All EMOS models for wind speed introduced in Section 5.3.2 require the estimation of parameters. Following the optimum score estimation approach described above, the parameters are determined by numerically optimizing the mean value of suitable proper scoring rules over a training set. As argued by Gneiting et al. (2005) and in Section 5.2.2, parameter estimation based on minimizing the mean CRPS generally results in more robust estimation procedures and the CRPS is thus often seen as the more appropriate scoring rule compared to the LogS. Here,

we give detailed descriptions of the employed parameter estimation methods for the proposed EMOS models.

5.3.3.1 Models based on single parametric families

Thorarinsdottir and Gneiting (2010) derive a closed-form expression of the CRPS of a truncated normal distribution given by

$$\begin{aligned}\mathrm{CRPS}\left(F_{\mathrm{TN}}, y\right) = \sigma\Bigg[\frac{y-\mu}{\sigma}\Phi\left(\mu/\sigma\right)\left(2\Phi\left((x-\mu)/\sigma\right)+\Phi\left(\mu/\sigma\right)-2\right)\\ + 2\varphi\left((y-\mu)/\sigma\right)\Phi\left(\mu/\sigma\right) - \frac{1}{\sqrt{\pi}}\Phi\left(\sqrt{2}\mu/\sigma\right)\Bigg]\left[\Phi\left(\mu/\sigma\right)\right]^{-2}.\end{aligned}$$

Similar calculations for the log-normal distribution show that

$$\begin{aligned}\mathrm{CRPS}\left(F_{\mathrm{LN}}, y\right) = y\Big[2\Phi\left((\log y-\mu)/\sigma\right)-1\Big]\\ -2\mathrm{e}^{\mu+\sigma^2/2}\left[\Phi\left((\log y-\mu)/\sigma-\sigma\right)+\Phi\left(\sigma/\sqrt{2}\right)-1\right],\end{aligned}$$

where $y \geq 0$, see Baran and Lerch (2015). These analytical expressions of the CRPS allow for an efficient parameter estimation by minimizing the mean CRPS over suitably chosen training periods, and are available in the `scoringRules` package (Jordan et al., 2016).

A closed-form expression of the CRPS for the GEV distribution with shape parameter $\xi < 1$ is derived by Friederichs and Thorarinsdottir (2012). For $\xi \neq 0$,

$$\begin{aligned}\mathrm{CRPS}(F_{\mathrm{GEV}}, y) = \left(\mu - \frac{\sigma}{\xi} - y\right)\left(1 - 2F_{\mathrm{GEV}}(y)\right)\\ -\frac{\sigma}{\xi}\left(2^{\xi}\Gamma(1-\xi)\right) - 2\Gamma_l(1-\xi, -\log F_{\mathrm{GEV}}(y)),\end{aligned}$$

where Γ denotes the gamma function and Γ_l denotes the lower incomplete gamma function. For $\xi = 0$, the CRPS is given by

$$\mathrm{CRPS}(F_{\mathrm{GEV}}, y) = \mu - y + \sigma(C - \log 2) - 2\sigma Ei(\log F_{GEV}(y)),$$

where $C \approx 0.5772$ is the Euler-Mascheroni constant and $Ei(x) = \int_{-\infty}^{x} \frac{e^t}{t}\,\mathrm{d}t$. Despite this closed-form analytical solution, minimum CRPS estimation for the GEV distribution turns out to be challenging due to numerical stability issues, for example in a neighborhood of $\xi = 0$, and requires various numerical approximations, see also Scheuerer (2014). For example, the `scoringRules` package (Jordan et al., 2016) includes an implementation of the CRPS of the GEV distribution where the exponential integral $Ei(\log F_{GEV}(y))$ is computed via numerical integration. Scheuerer (2014) suggests to approximate the CRPS for $\xi = 0$ by linear interpolation of CRPS values for $\xi = -\epsilon$ and $\xi = \epsilon$ with small $\epsilon > 0$.

For the data sets investigated here, maximum likelihood estimation proved to be more parsimonious and numerically stable. There is no analytical solution of

the corresponding ML minimization problem (Coles, 2001). However, numerical approximations can be obtained using standard algorithms for any given dataset (Prescott and Walden, 1980; Hosking, 1985), see also Bücher and Segers (2016) for a recent study of asymptotic properties of the ML estimator for the GEV distribution. We therefore employ ML estimation to determine the parameters of the GEV model (5.5), and use the `R` implementation of the estimation algorithms provided by the `ismev` package (Heffernan et al., 2014).

5.3.3.2 Regime-switching combination models

For the regime-switching combination models introduced in Section 5.3.2.2, we use the respective optimum score estimation approaches described above for the individual distributional components of the model. Therefore, in case of the TN-GEV combination model (5.6), minimum CRPS estimation is applied for the parameters of the TN distribution, and ML estimation for the parameters of the GEV distribution, whereas in the case of the TN-LN combination model (5.7), the parameters of both the TN and the LN distribution are determined using minimum CRPS estimation. As discussed in Section 5.3.2.2, the training sets are partitioned according to the model selection criterion, and the parameters of the two candidate distributions are estimated separately using only past forecast cases from the respective subsets.

Both combination models require the choice of a model selection threshold θ for the ensemble median that determines whether the TN or a heavy-tailed distribution is issued as forecast distribution. We estimate θ by computing the mean CRPS for a range of threshold values over an out-of-sample training period for all three data sets, see Section 5.4 for details.

As an alternative to using a fixed model selection threshold θ, we have also investigated a more adaptive estimation procedure where the threshold parameter is re-estimated as a fixed quantile of the median values of the ensemble in the corresponding training period for each forecast date. However, this estimation approach does not result in significant improvements of predictive performance for any of the investigated ensembles or combination models, we therefore restrict our attention to the simpler use of a fixed model selection threshold described above.

5.3.3.3 Weighted mixture models

The CRPS for the weighted mixture model (5.8) is not available in a closed analytical form and has to be evaluated numerically. The required numerical integration steps lead to computationally demanding optimization procedures. Rewriting the CRPS as

$$\mathrm{CRPS}(H, y) = \int_{-\infty}^{y} H(z)^2 \, \mathrm{d}z + \int_{y}^{\infty} (1 - H(z))^2 \, \mathrm{d}z$$

leads to slightly lower computational costs and better predictive performance of the resulting models, but still requires significantly more computational resources

compared to all other investigated approaches.

Initial tests for the simple EMOS models based on single TN or LN distributions indicated that maximum likelihood estimation leads to less calibrated forecasts and lower predictive performance. However, because of the high computational costs of minimum CRPS estimation for the mixture model, we also investigate the more parsimonious ML estimation of the parameters. In figures and tables, the corresponding mixture models are denoted by TN-LN mix. (CRPS) and TN-LN mix. (ML).

5.3.3.4 Temporal extent and spatial composition of training sets

For all three types of EMOS models, the parameters are estimated over a rolling training period consisting of the forecast-observation pairs of the last n days. To determine the optimal length of the training period, we estimate the different models for training periods of various lengths, and choose the value of n that leads to the best predictive performances across all types of models. However, as we will see below in Section 5.4, the length of the training period generally has a negligible effect on the verification scores and the rankings of the different models.

Regarding the spatial composition of the training sets, the parameters are estimated regionally in that data from all stations are pooled together. Local estimation of the models leads to numerically unstable optimization algorithms and partially lower predictive performance for all three data sets, in particular for the combination and mixture models. We will investigate alternative similarity-based semi-local estimation methods in Chapter 6.

5.4 Case studies

Here, we present detailed results for the three case studies.

5.4.1 ECMWF data

To determine the optimal length of the training period, we have estimated the models for training periods of length $n = 15, 16, \ldots, 40$ days. The mean CRPS values for the corresponding EMOS models based on single parametric families are shown in Figure 5.6. The best results are obtained with training period lengths of 20 days. The CRPS and other performance scores reported below change by less than 1% for the different values of n, and, in accordance with the results of Thorarinsdottir and Gneiting (2010), we conclude that the methods are robust against changes in n.

The model selection threshold θ for the TN-GEV and TN-LN regime-switching combination models is determined in a similar fashion. We estimate the combination models over a range of possible threshold values, and investigate the influence on the predictive performance. Figure 5.7 shows the mean CRPS as a function of the threshold θ for both the TN-GEV and the TN-LN model for various choices

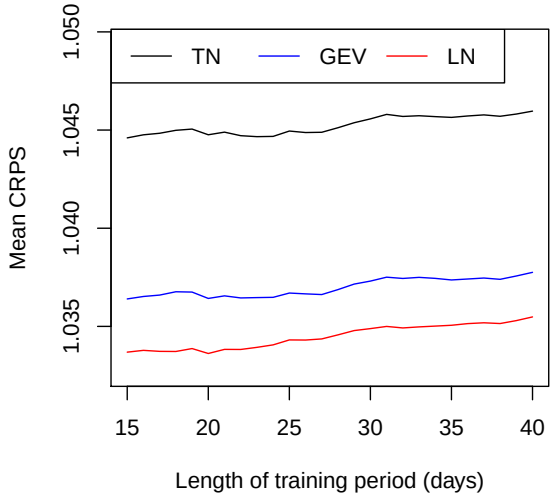


Figure 5.6: Mean CRPS of the EMOS models based on a single parametric family introduced in Section 5.3.2.1 for the ECMWF data for various training period lengths.

of n. As before, we observe that a training period length of $n = 20$ days proves optimal. The optimal model selection thresholds are $\theta = 8.0$ m s^{-1} for the TN-GEV combination model and $\theta = 7.3$ m s^{-1} for the TN-LN combination model. With these threshold values, a GEV distribution is used in around 19% of the forecast cases in the verification set, and an LN distribution is used in around 14% of the forecast cases. We use these threshold parameter values and a training period length of 20 days to estimate all models introduced in Section 5.3.2. These models are now evaluated over the out-of-sample verification period from May 1, 2010 to April 30, 2011.

We compare the ensemble postprocessing methods discussed above to the raw, unprocessed ECMWF ensemble and a climatological reference forecast. For each day, the climatological reference forecast is obtained from the observed wind speeds in the 20 day training period used for the parameter estimation of the postprocessing methods.

Verification rank and PIT histograms for the ensemble and the various postprocessing methods are shown in Figure 5.8. The ECMWF forecasts are underdispersive, with too many observations falling outside the ensemble range. This deficiency has repeatedly been observed for various ensemble prediction systems. Possible causes for the ECMWF ensemble in this case are underdispersiveness of the underlying model, unsatisfactory modeling of the uncertainty using random perturbations, and spatial and temporal interpolation and smoothing issues, see, e.g., Hamill and Colucci (1997), Palmer (2002) and Raftery et al. (2005).

All postprocessing methods significantly improve the calibration of the ensemble. While the GEV forecasts are slightly overdispersive, their PIT histogram shows smaller deviations from uniformity than that of the TN forecasts. The PIT histograms thus indicate that the GEV distributions tend to have minimally too heavy tails while the upper tails for the TN distributions seem slightly too

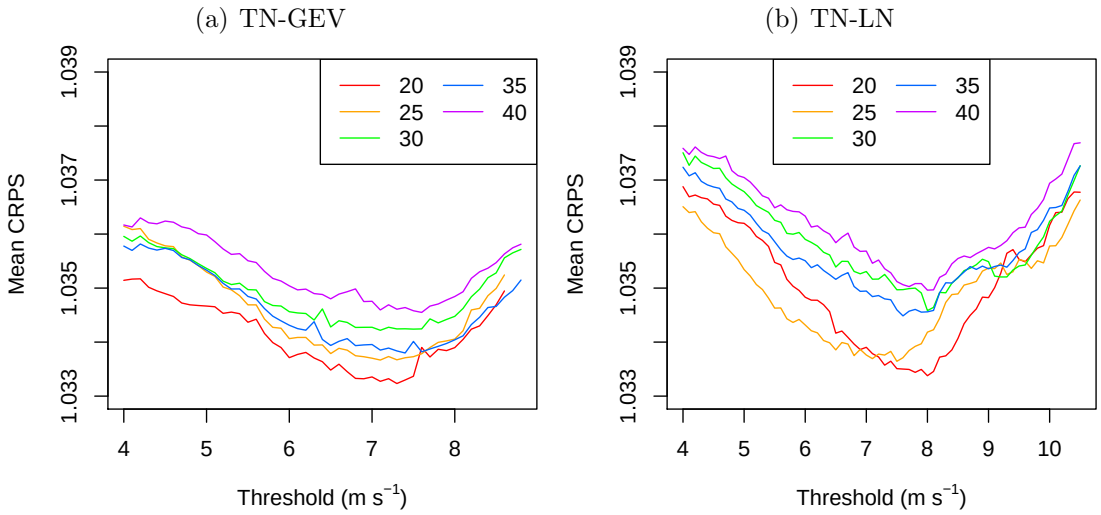


Figure 5.7: Mean CRPS of the regime-switching combination models introduced in Section 5.3.2.2 for the ECMWF data for various training period lengths (in days) and model selection thresholds.

light. The LN model shows similar behavior to the TN model, but the deviations from uniformity are considerably smaller. The PIT histograms of the combination models resemble the PIT histogram of the TN technique, with minor improvements for large PIT values. The forecast of the TN-LN mixture models (5.8) exhibit the best calibration with only minor deviations from the desired uniform distribution of the PIT values.

As discussed in Section 2.2, formal statistical test of uniformity can be used to assess calibration in addition to the visual inspection of PIT histograms. As the PIT values of multi-step ahead probabilistic forecast exhibit serial correlation (see, e.g., Diebold et al., 1998) and the probabilistic forecasts cannot be assumed to be independent in space and time, we employ a moment-based test of uniformity proposed by Knüppel (2015) which accounts for dependence in the PIT values. In particular, we use the α^0_{1234} test of Knüppel (2015) that has been demonstrated to have superior size and power properties compared to alternative choices. Due to the large sample size in case of the ECMWF and UWME data, the null hypothesis of uniformity is rejected for all postprocessing models. However, as our focus lies on the comparative assessment of calibration, we report bootstrap estimates of the rejection rates of the α^0_{1234} test based on 10 000 random samples of size 1 000 each in Table 5.1. If a model exhibits superior calibration, the null hypothesis of uniformity should be rejected in fewer cases compared to a model with inferior calibration.

For the ECMWF data it can be observed that the null hypothesis of uniformity is rejected in almost all of the cases for the TN, GEV, LN and combination models, whereas the TN-LN mixture models show much lower rejection rates. This observation is clearly in line with the visual inspection of the PIT histograms in Figure 5.8.

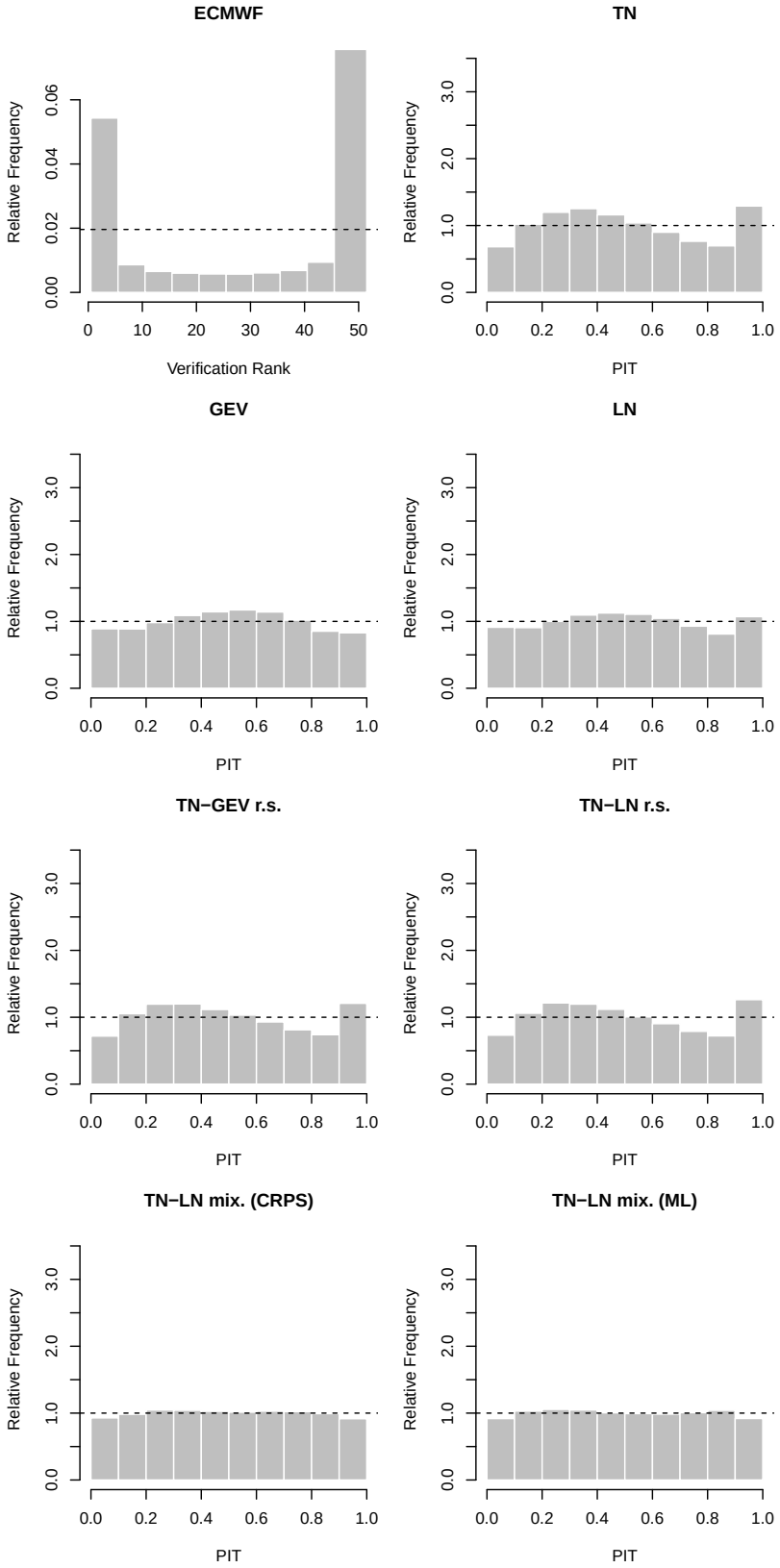


Figure 5.8: VR histogram for the raw ensemble and PIT histograms for the EMOS postprocessed forecasts for the ECMWF data.

Table 5.1: Bootstrap estimates of rejection rates of the α^0_{1234} test of uniformity based on 10 000 random samples of size 1 000 each at the 0.05 level for the different data sets. Lower rejection rates correspond to better calibrated forecasts with the null hypothesis of uniformity being rejected on fewer occasions.

Forecast	ECMWF	ALADIN-HUNEPS	UWME
TN	1	0.76	0.97
GEV	0.95	0.23	0.10
LN	0.98	1	0.99
TN-GEV r.s. comb.	1	0.71	0.56
TN-LN r.s. comb.	1	0.71	0.27
TN-LN mix. (CRPS)	0.28	0.02	0.20
TN-LN mix. (ML)	0.12	0.05	0.13

Table 5.2: Mean CRPS, mean absolute error, average coverage and width of 96.08% prediction intervals of probabilistic one day ahead forecasts of daily maximum wind speed at 228 synoptic stations in Germany from May 1, 2010 to April 30, 2011. For each column, the best value among the postprocessed forecasts is printed in bold.

Forecast	CRPS (m s^{-1})	MAE (m s^{-1})	Coverage (%)	Width (m s^{-1})
Climatology	1.550	2.144	95.84	11.91
Ensemble	1.263	1.441	45.00	1.80
TN	1.045	1.388	92.19	6.39
GEV	1.034	1.388	94.84	8.22
LN	1.037	1.386	93.16	6.91
TN-GEV r.s. comb.	1.033	1.381	92.89	6.60
TN-LN r.s. comb.	1.033	**1.379**	92.49	6.36
TN-LN mix. (CRPS)	**1.030**	1.384	94.34	7.71
TN-LN mix. (ML)	1.034	1.391	**95.81**	8.72

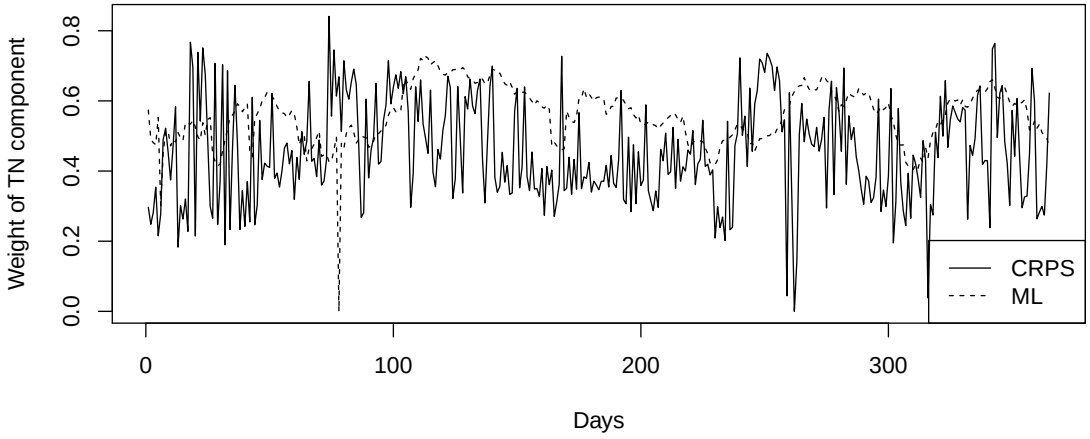


Figure 5.9: Weight of the truncated normal component of the TN-LN mixture models estimated with minimum CRPS and ML estimation for the ECMWF data over the 365 days in the verification period from May 1, 2010 to April 30, 2011.

Table 5.2 shows the mean CRPS, the MAE and average coverage and width of 96.08% prediction intervals for the competing forecasts. The prediction intervals are chosen such that the coverage can be directly compared to the nominal coverage of the ensemble which is 49/51. The point forecast evaluated by the MAE is given by the median of the corresponding predictive distribution.

The ECMWF ensemble predictions outperform the climatological reference forecast and provide sharp prediction intervals at the cost of being uncalibrated. All postprocessing methods outperform the ensemble predictions, with the GEV and LN methods showing small improvements in mean CRPS compared to the TN method. The regime-switching combination of the lighter-tailed TN and heavier-tailed GEV or LN distributions further improves the predictive performance compared to the simple EMOS models based on a single parametric distribution. The predictive performance of the TN-LN mixture models is comparable to the regime-switching combination models, with the mixture model utilizing minimum CRPS estimation showing the overall lowest mean CRPS. Note that due to the heavier tails, models involving a GEV or LN distribution generally result in wider prediction intervals than the TN model. The positive effect of postprocessing on the calibration of the forecast distributions can again be observed from the significantly improved coverage values compared to the underdispersive raw ensemble forecasts.

Two-sided Diebold-Mariano tests indicate that the observed score differences between the models based on a single parametric distribution and the models that combine lighter and heavier-tailed distributions are all significant at the 5% level. However, the only significant score differences among these combination and mixture models are those between the TN-LN mixture model based on minimum CRPS estimation and the remaining models.

Despite the similar predictive performance, the weights w of the truncated

Table 5.3: Mean twCRPS for forecasts of daily maximum wind speed at 228 synoptic stations in Germany from May 1, 2010 to April 30, 2011 using an indicator weight function $w(z) = \mathbb{1}\{z \geq r\}$ for different values of r (in m s^{-1}). For each column, the best value is printed in bold.

	twCRPS		
Forecast	$r = 10$	$r = 12$	$r = 15$
Climatology	0.251	0.128	0.045
Ensemble	0.211	0.113	0.043
TN	0.200	0.110	0.042
GEV	0.195	0.106	0.041
LN	0.198	0.109	0.042
TN-GEV r.s. comb.	**0.191**	**0.103**	**0.039**
TN-LN r.s. comb.	**0.191**	**0.103**	**0.039**
TN-LN mix. (CRPS)	0.194	0.106	0.041
TN-LN mix. (ML)	0.196	0.108	0.041

normal component of the TN-LN mixture models (5.8) for the two parameter estimation variants evolve quite differently as illustrated in Figure 5.9. We observe only a small correlation of 0.063, and the weights determined by ML estimation vary much slower over time. However, the component weights of course only provide a partial image, and the corresponding location and scale/shape parameters of the TN and LN components are much more strongly correlated.

Figure 5.10 compares the station-specific predictive performance of the individual postprocessing models as a function of the site-specific average observed wind speed. In the case of positive values of the shown mean CRPS differences, the proposed competitors outperform the TN model at the specific station corresponding to the point in the plot. All comparisons indicate that the overall improvements of the proposed models involving heavier-tailed distributions over the TN model are mainly due to improvements at stations with high average observed wind speeds.

With a focus on the performance in the upper tail, Table 5.3 reports values of the mean threshold-weighted continuous ranked probability score,

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

for the competing forecasts where we have employed the indicator weight function $w(z) = \mathbb{1}\{z \geq r\}$ for $r = 10, 12$ and 15 m s^{-1}. The threshold values approximately correspond to the 90th, 95th and 98th percentile of the marginal distribution of the wind speed observations. All postprocessing methods improve the ECMWF ensemble predictions, and the new approaches incorporating GEV or LN distributions outperform the TN method. The best results are obtained for regime-switching combination models, potentially due to the bipartite nature of

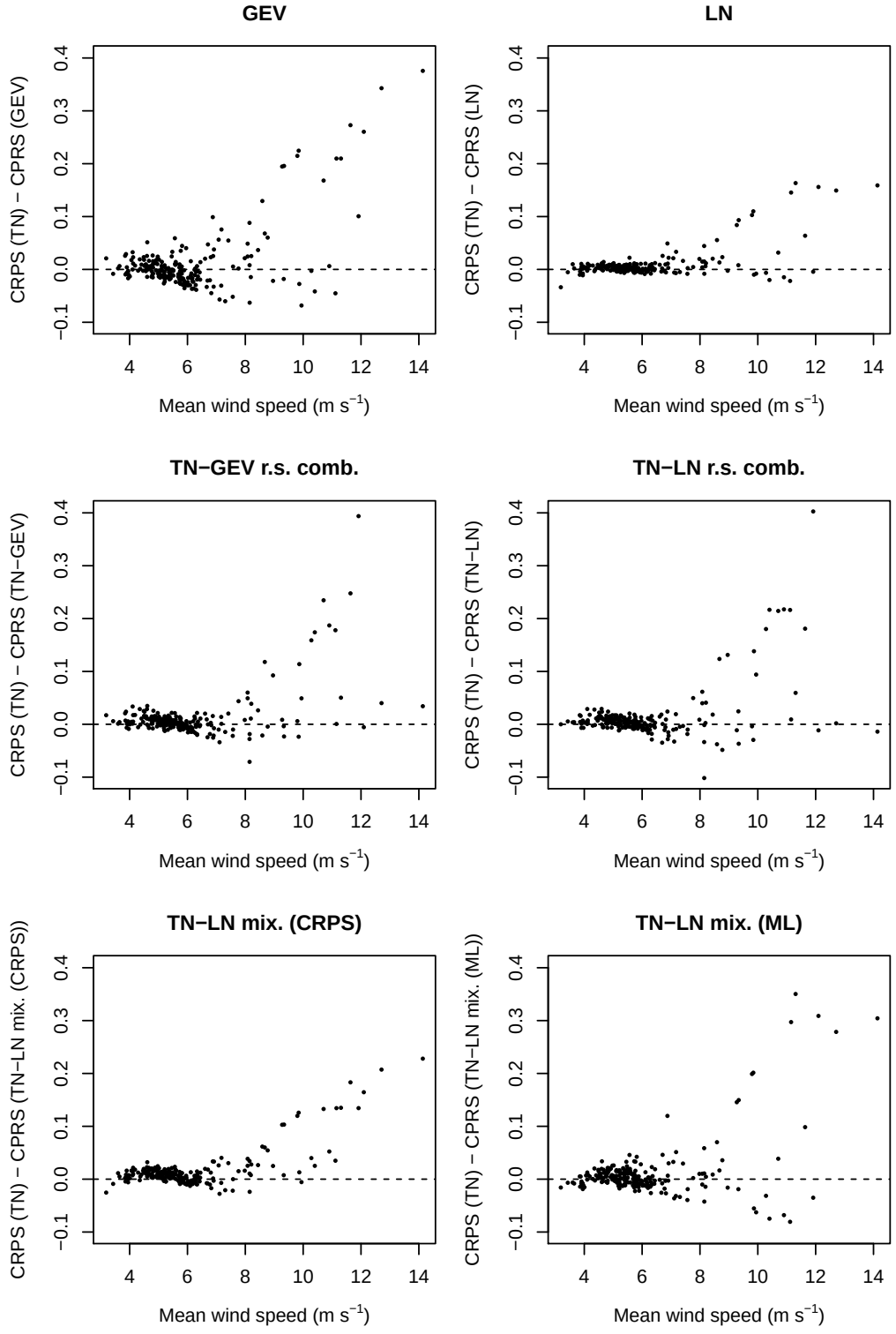


Figure 5.10: Station-specific comparisons of the mean CRPS of the postprocessing methods as a function of the average observed daily maximum wind speed at the station. The plots compare the TN model and all proposed competitors. The horizontal dashed lines indicate equal predictive performance.

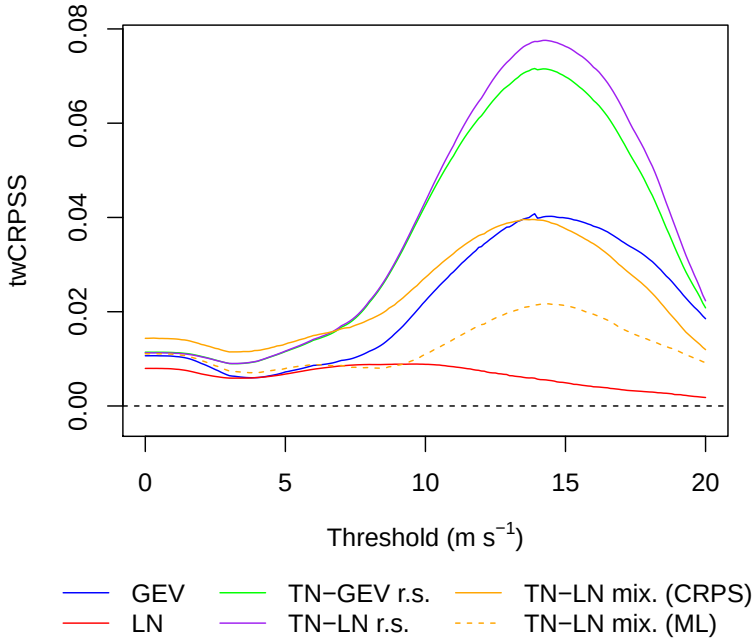


Figure 5.11: Threshold-weighted continuous ranked probability skill score (5.9) of probabilistic forecasts of daily maximum wind speed at 228 synoptic stations in Germany from May 1, 2010 to April 30, 2011 as a function of the threshold r in the indicator weight function $w(z) = \mathbb{1}\{z \geq r\}$, using the forecasts produced by the TN method as reference.

the model estimation. Note that the relative improvement over the TN method for the upper tail is comparatively larger than the improvement under the unweighted CRPS in Table 5.2. Similar rankings hold for any value of r between 10 and 20 m s^{-1}.

To further investigate the predictive performance for high wind speed values we consider the threshold-weighted continuous ranked probability skill score (twCRPSS) given by

$$\text{twCRPSS}(F, y) = 1 - \frac{\text{twCRPS}(F, y)}{\text{twCRPS}(F_{\text{ref}}, y)}, \tag{5.9}$$

where F_{ref} denotes the predictive cumulative distribution function of a reference forecast, in our case the TN method. The twCRPSS is positively oriented and can be interpreted as improvement over the reference forecast.

Figure 5.11 shows the twCRPSS for the proposed methods as a function of the threshold r for the indicator weight function, using the TN method as a reference forecast. For all thresholds and both models, the twCRPSS is strictly positive, indicating improved predictive performance compared to the TN model. Again, it can be observed that the regime-switching combination models result in the greatest improvements. Interestingly, despite the small magnitude of the improvements provided by the LN method, the TN-LN combination model shows the highest twCRPSS values for a wide range of high threshold values. In gen-

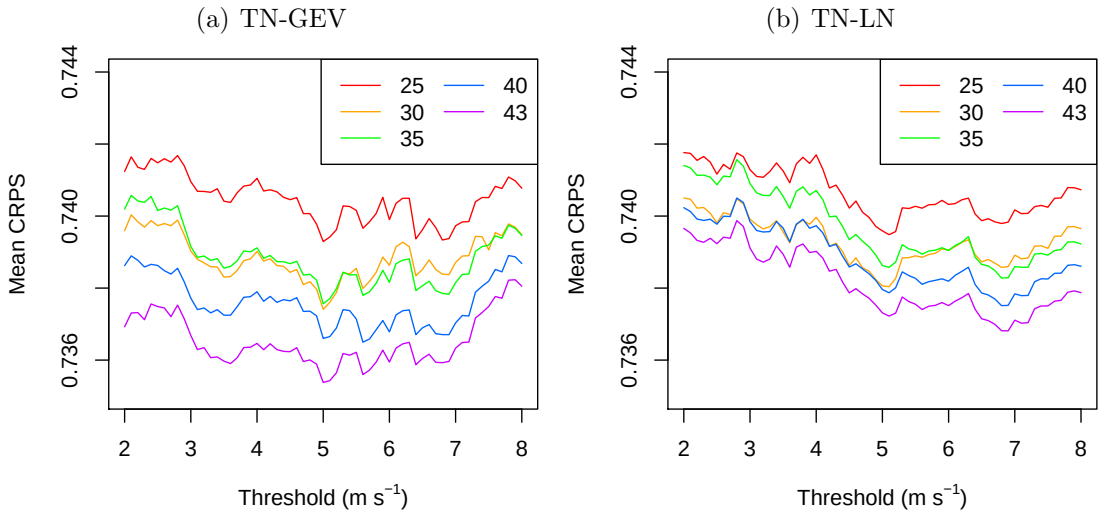


Figure 5.12: Mean CRPS of the regime-switching combination models introduced in Section 5.3.2.2 for the ALADIN-HUNEPS data for various training period lengths (in days) and model selection thresholds.

eral, the score values increase for larger thresholds, with the largest differences being obtained for threshold values around 14 m s^{-1}. The decrease at very large threshold values might constitute an example of the problematic tail dependence of weighted proper scoring rules we observed in Section 3.3.4.

5.4.2 ALADIN-HUNEPS data

As discussed above, the generation of the ALADIN-HUNEPS ensemble leads to two groups of ensemble members given by a control forecast and 10 perturbed members. We therefore use the respective model formulations introduced in Section 5.3.2.1. Note that Baran et al. (2013) consider a different grouping where the odd and even numbered exchangeable ensemble members form two separate groups. This approach was based on deficiencies in the generation of the ensemble forecasts in earlier versions of the NWP model, since only five perturbations were calculated, and then added to and subtracted from the unperturbed initial conditions to obtain the odd and even numbered members, respectively. This does not appear to be the case for the data set considered here, and the two- and three-group models perform very similar. We therefore only report results for the two-group case. The out-of-sample verification period between May 15, 2012 and March 31, 2013 consists of 3150 individual forecast cases.

Following previous studies of Baran et al. (2014) and Baran (2014) that include the TN model, we choose a training period length of 43 days. Due to the lower number of observation stations in the data set leading to smaller training sets, the estimation procedure for the combination models described in Section 5.3.3.2 has to be adapted. Instead of separately estimating the TN and GEV/LN components of the model on the respective subsets of the training period where the median

ensemble forecast was below or above θ, we utilize the whole training sets to estimate the parameters of all involved distributions, and then choose between these distributional models based on the ensemble median.

The model selection threshold θ for the regime-switching combination models is determined as for the ECMWF ensemble. Figure 5.12 shows the mean CRPS of the combination models as a function of the threshold for various training period lengths. In contrast to the ECMWF ensemble, the length of the training period here appears to have a stronger influence on the predictive performance compared to the threshold θ. Based on Figure 5.12 we select model thresholds of 5 m s^{-1} for the TN-GEV combination model and 6.9 m s^{-1} for the TN-LN combination model. With these threshold values, GEV and LN distributions are used are 15% and 4% of the forecast cases, respectively.

Figure 5.13 shows verification rank and PIT histograms for the ALADIN-HUNEPS ensemble and the various postprocessing methods. Similar to the ECMWF ensemble, the ALADIN-HUNEPS predictions are underdispersive as indicated by the U-shaped VR histogram. Compared to the ECMWF ensemble, the underdispersive character of the predictions is less pronounced, however, all postprocessing approaches still show substantially better calibration. PIT histograms of the GEV and LN models show a slight overdispersion and over-prediction of low wind speed values, whereas the PIT histograms of the regime-switching combination models closely resemble that of the TN method. The best results are obtained for the TN-LN mixture models with PIT histograms that exhibit no visible systematic deviations from the desired uniform distribution.

These observations are corroborated by the rejection rates of the α^0_{1234} test of uniformity reported in Table 5.1. The null hypothesis of uniformity is rejected in almost none of the randomly selected samples for the TN-LN mixture models, whereas the competing postprocessing methods show substantially higher frequencies of rejections.

Table 5.4 summarizes the predictive performance of the various competing forecasts and shows the mean CRPS, MAE, and average coverage and width of 83.33% prediction intervals. As before, the prediction intervals are chosen to match the nominal coverage of the ensemble which is 10/12. Similar to the results for the ECMWF ensemble, the raw ensemble predictions outperform the climatological forecasts, however, are uncalibrated with too narrow prediction intervals. All postprocessing methods significantly improve the ensemble forecasts. Except for the LN model, the proposed EMOS approaches involving heavier-tailed distributions further outperform the TN model.

The best results are obtained for the regime-switching combination models, however, note that the relative differences among the various postprocessing methods are small. The LN model fails to outperform the TN model, but the regime-switching combination of these two models improves the TN forecasts although an LN distribution is used in only about 4% of the forecast cases. Using two-sided Diebold-Mariano tests of equal predictive performance, we find that most of the observed score differences are not significant at the 5% level, the only exception are comparisons of the LN model and all competitors except for the

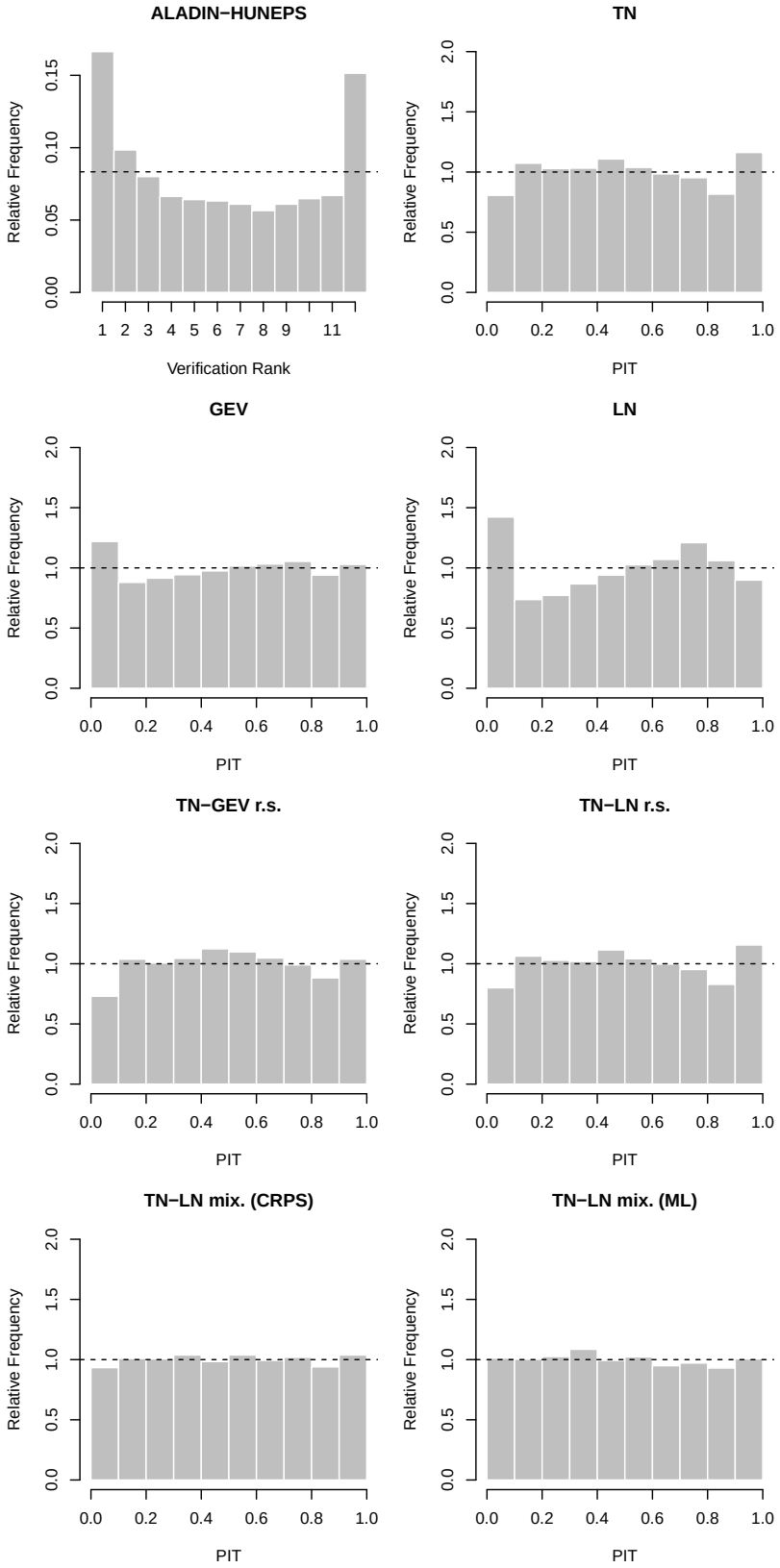


Figure 5.13: VR histogram for the raw ensemble and PIT histograms for the EMOS postprocessed forecasts for the ALADIN-HUNEPS data.

Table 5.4: Mean CRPS, mean absolute error, average coverage and width of 83.33% prediction intervals of probabilistic forecasts of wind speed at 10 major cities in Hungary from May 15, 2012 to March 31, 2013. For each column, the best value among the postprocessed forecasts is printed in bold.

Forecast	CRPS (m s^{-1})	MAE (m s^{-1})	Coverage (%)	Width (m s^{-1})
Climatology	1.046	1.481	82.54	3.43
Ensemble	0.803	1.069	68.22	2.88
TN	0.738	1.037	83.59	3.53
GEV	0.737	1.041	81.21	3.54
LN	0.741	1.038	80.44	3.57
TN-GEV r.s. comb.	**0.735**	1.039	85.59	3.72
TN-LN r.s. comb.	0.737	**1.035**	83.59	3.54
TN-LN mix. (CRPS)	0.736	1.037	83.02	3.62
TN-LN mix. (ML)	0.737	1.040	**83.14**	3.58

TN model.

The two TN-LN mixture models based on minimum CRPS and ML estimation show very similar predictive performance. The weights of the TN component obtained with the two optimization procedures exhibit similar variation and correlation structures as observed for the ECMWF ensemble in Figure 5.9. We omit plots of the corresponding weights and note that again, the parameters of the respective forecast distributions show high correlation.

To investigate the predictive performance at high wind speed values, Table 5.5 shows mean threshold-weighted continuous ranked probability scores where we have employed an indicator weight function. The threshold values of 6, 7 and 9 m s^{-1} approximately correspond to the 90th, 95th and 98th percentile of the marginal distribution of the wind speed observations. As for the ECMWF ensemble, we observe that all postprocessing methods increase the predictive performance at high wind speed values. The best results are obtained for the GEV and TN-GEV combination models, however, note that the relative differences are small, particularly for high threshold values.

Figure 5.14 compares the predictive performance of the competing postprocessing methods for high wind speed values in terms of the mean threshold-weighted continuous ranked probability skill score which is shown as a function of the threshold r employed in the weight function in equation (5.9). The forecasts of the TN model are again used as a reference. The best results are obtained by the two models involving a GEV distribution. The TN-LN mixture models outperform the basic LN and the TN-LN combination model which offer no or only small improvements over the TN method. As for the ECMWF data, the

Table 5.5: Mean twCRPS for forecasts of wind speed at 10 major cities in Hungary from May 15, 2012 to March 31, 2013 using an indicator weight function $w(z) = \mathbb{1}\{z \geq r\}$ for different values of r (in m s^{-1}). For each column, the best value is printed in bold.

	twCRPS		
Forecast	$r = 6$	$r = 7$	$r = 9$
Climatology	0.127	0.064	0.012
Ensemble	0.112	0.059	0.013
TN	0.102	0.054	0.012
GEV	**0.098**	**0.052**	**0.011**
LN	0.102	0.054	**0.011**
TN-GEV r.s. comb.	**0.098**	**0.052**	**0.011**
TN-LN r.s. comb.	0.101	0.054	**0.011**
TN-LN mix. (CRPS)	0.100	0.053	**0.011**
TN-LN mix. (ML)	0.100	0.053	0.012

relative improvements over the TN model generally increase with the threshold value, however, diminish for r exceeding around 6.5 m s^{-1}.

5.4.3 UWME data

The University of Washington Mesoscale Ensemble consists of eight ensemble members which are clearly distinguishable as they are generated with initial conditions from different sources. The number of parameters to be estimated for the model formulations introduced in Section 5.3.2 is therefore larger compared to the ECMWF and ALADIN-HUNEPS data.

We proceed as above, and start by determining the optimal training period length, and the model threshold θ for the TN-GEV and TN-LN combination models. From Figure 5.15 we observe that the temporal extent of the training period only has a small influence on the predictive performance, and the same rankings are obtained for all investigated choices of n. The best results across all three distributions are obtained for training period lengths around $n = 30$ days.

To determine the model threshold for the TN-GEV and TN-LN combination models, Figure 5.16 shows the average CRPS over the out-of-sample verification period from January 1, 2008 to December 31, 2008 for various choices of threshold values and training period lengths. Compared to the corresponding plots for the ALADIN-HUNEPS data in Figure 5.12, the threshold parameter appears to have a bigger effect as the rankings of the forecasts produced with different training period lengths change frequently. The overall best results are obtained with thresholds values of 5.7 m s^{-1} for the TN-GEV combination model and 5.2 m s^{-1} for the TN-LN variant. With these choices, a GEV distribution is used in around 40% of the forecast cases and an LN distribution is used in around 33%.

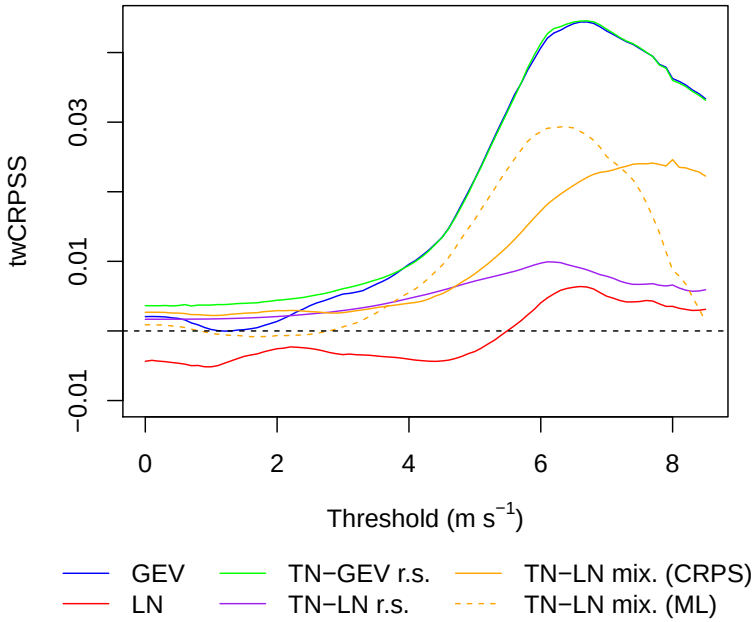


Figure 5.14: Threshold-weighted continuous ranked probability skill score (5.9) of probabilistic forecasts of wind speed at 10 major cities in Hungary from May 15, 2012 to March 31, 2013 as a function of the threshold r in the indicator weight function $w(z) = \mathbb{1}\{z \geq r\}$, using the forecasts produced by the TN method as reference.

To assess calibration, Figure 5.17 shows verification rank and PIT histograms for the UWME ensemble and the various postprocessing methods. Not surprisingly, the raw ensemble forecasts are again underdispersive, whereas all postprocessing approaches significantly improve the calibration. The deviations from uniformity are generally smaller compared to the previously investigated data sets. The TN and LN model both show slight overdispersion, as well as underestimation of high, and overestimation of low wind speed observations, respectively. Again, the TN-GEV and TN-LN combination models are somewhat able to correct for these deficiencies and show smaller deviations from uniformity compared to the TN distribution. The most uniform PIT histograms are obtained for the GEV model and the TN-LN mixture models.

These observations are in line with the results of the formal statistical test of uniformity presented in Table 5.1. On average, all models show lower rejection rates indicating superior calibration compared to the ECMWF and ALADIN-HUNEPS data. The best results are obtained for the GEV model where the null hypothesis of uniformity is rejected in only around 10% of the random samples, followed by the TN-LN mixture model approaches. Over all three investigated data sets, this is only instance where any model exhibits a better calibration than the TN-LN mixture models.

Table 5.6 shows the mean CRPS, MAE, and absolute coverage and width of 77.78% prediction intervals. The prediction intervals are again chosen such that

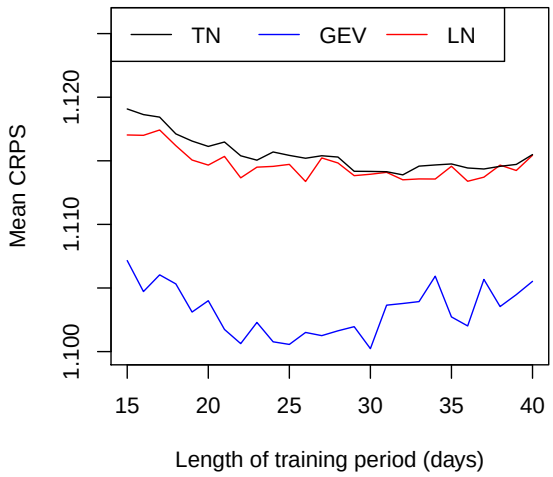


Figure 5.15: Mean CRPS of the EMOS models based on a single parametric family introduced in Section 5.3.2.1 for the UWME data for various training period lengths.

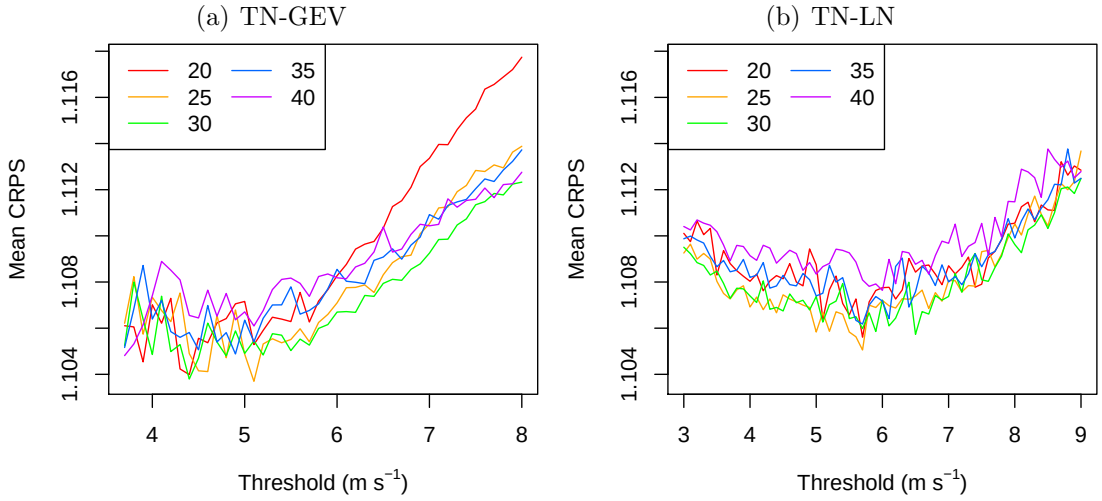


Figure 5.16: Mean CRPS of the regime-switching combination models introduced in Section 5.3.2.2 for the UWME data for various training period lengths (in days) and model selection thresholds.

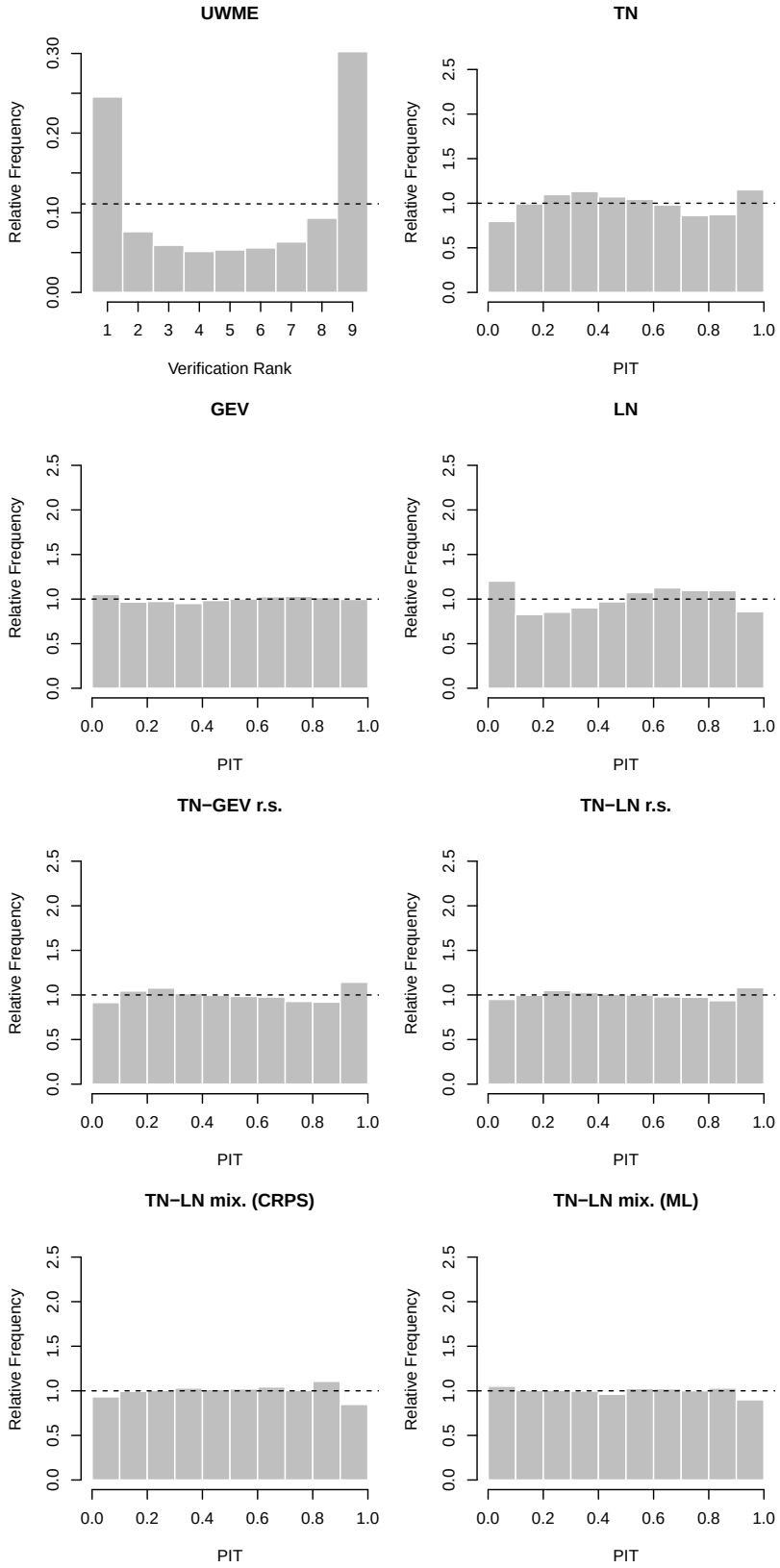


Figure 5.17: VR histogram for the raw ensemble and PIT histograms for the EMOS postprocessed forecasts for the UWME data.

Table 5.6: Mean CRPS, mean absolute error, average coverage and width of 77.78% prediction intervals of probabilistic forecasts of wind speed at 152 observation stations in the Pacific Northwest region of the United States in the calendar year 2008. For each column, the best value among the postprocessed forecasts is printed in bold.

Forecast	CRPS (m s^{-1})	MAE (m s^{-1})	Coverage (%)	Width (m s^{-1})
Climatology	1.412	1.987	81.10	5.90
Ensemble	1.353	1.655	45.24	2.53
TN	1.114	**1.550**	78.65	4.67
GEV	**1.100**	1.554	77.20	4.69
LN	1.114	1.554	77.29	4.69
TN-GEV r.s. comb.	1.105	1.555	77.20	4.60
TN-LN r.s. comb.	1.105	**1.550**	**77.73**	4.64
TN-LN mix. (CRPS)	1.105	**1.550**	79.02	4.77
TN-LN mix. (ML)	1.108	1.560	78.12	4.78

the coverage can be directly compared to the nominal coverage of the ensemble which is 7/9. As for the previously investigated data sets, all postprocessing approaches significantly improve the calibration and predictive performance of the raw ensemble forecasts. All proposed models perform at least as well as the TN method. The best mean CRPS values are obtained for the GEV model, followed by the regime-switching combination models and the TN-LN mixture models. All observed CRPS differences between the basic TN model and the GEV, combination and mixture models are significant at the 5% level in two-sided Diebold-Mariano tests. Comparing the two parameter estimation variants of the TN-LN mixture models we observe that as for the other two data sets, minimum CRPS estimation of the parameters leads to slightly better predictive performance, however, the relative differences are small and the parameters of the component distributions are again strongly correlated.

With an average coverage of 77.73%, the TN-LN combination model shows the smallest deviations from the nominal coverage of 77.78% among the investigated models. However, the PIT histograms in Figure 5.17 and the results of the α^0_{1234} test in Table 5.1 indicate that the forecasts of the GEV and TN-LN mixture models are much better calibrated. This observation illustrates that studying the coverage of prediction intervals alone might result in misleading conclusions.

Turning to the predictive performance for high wind speed observations, Table 5.7 shows mean values of the threshold-weighted CRPS for the various forecasts. The threshold values in the indicator weight function approximately correspond to the 90th, 95th and 98th percentile of the empirical distributions of all wind speeds observations. The GEV and TN-GEV combination model show the best

Table 5.7: Mean twCRPS for forecasts of wind speed at 152 observation stations in the Pacific Northwest region of the United States in the calendar year 2008 using an indicator weight function $w(z) = \mathbb{1}\{z \geq r\}$ for different values of r (in m s^{-1}). For the first two columns, the best value is printed in bold.

	twCRPS		
Forecast	$r = 9$	$r = 10.5$	$r = 14$
Climatology	0.173	0.081	0.010
Ensemble	0.175	0.081	0.011
TN	0.150	0.074	0.010
GEV	**0.145**	**0.072**	0.010
LN	0.149	0.073	0.010
TN-GEV r.s. comb.	**0.145**	**0.072**	0.010
TN-LN r.s. comb.	0.149	0.073	0.010
TN-LN mix. (CRPS)	0.147	0.073	0.010
TN-LN mix. (ML)	0.147	0.073	0.010

predictive performance at high wind speed values, however, the relative differences are again small.

Similar conclusions can be drawn from Figure 5.18 where the twCRPSS is shown as a function of the threshold r in the indicator weight function with the TN model as reference forecast. Again, the relative improvements are smaller compared to the ECMWF data. The twCRPSS decreases for threshold values exceeding around 10 m s^{-1}, and is negative at higher threshold values. However, as discussed above, this might be a consequence of the lighter tails of the TN distribution. Interestingly, the regime-switching combination of TN and GEV distributions is unable to outperform the GEV model.

5.5 Discussion

We propose several extensions of the TN approach to ensemble postprocessing for wind speed of Thorarinsdottir and Gneiting (2010), and employ GEV and LN predictive distributions with heavy right tails. Combination and mixture model approaches aim to combine advantages of lighter and heavier-tailed distributions. In three case studies with different ensemble prediction systems and observed wind quantities all postprocessing methods significantly improve the calibration as well as the overall skill of the raw ensemble. The novel EMOS models show consistent improvements over the standard TN approach, particularly for high wind speed observations which the light tails of the TN distribution fail to resolve correctly. The overall best results in terms of verification scores are generally obtained by the GEV model and the TN-GEV combination model, whereas the

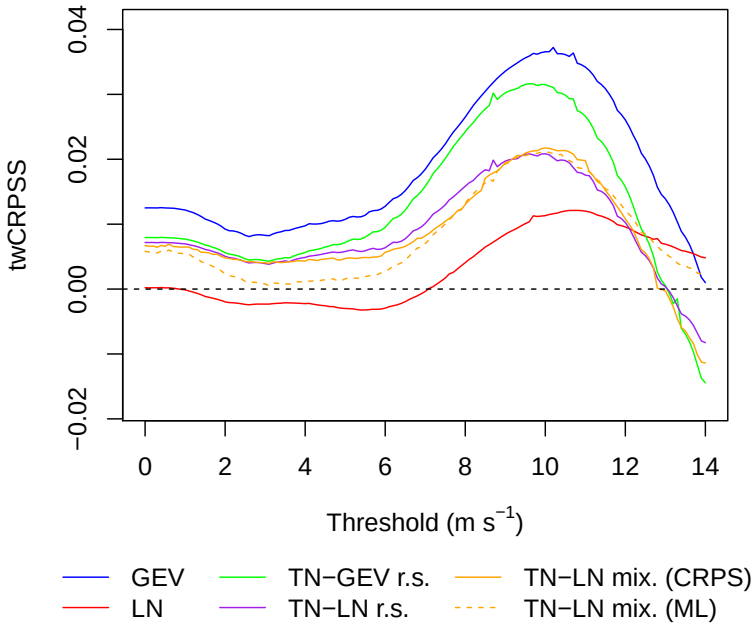


Figure 5.18: Threshold-weighted continuous ranked probability skill score (5.9) of probabilistic forecasts of wind speed at 152 observation stations in the Pacific Northwest region of the United States in the calendar year 2008 as a function of the threshold r in the indicator weight function $w(z) = \mathbb{1}\{z \geq r\}$, using the forecasts produced by the TN method as reference.

TN-LN mixture models show superior calibration.

A comparison of the TN, GEV and LN models suggests that the choice of a suitable parametric EMOS model for postprocessing ensemble forecasts is a non-trivial exercise and depends on the employed performance measure as well as the data. The GEV model generally results in slightly better predictions, however, has the disadvantage that the support of the GEV distribution depends on the parameter values and might include negative wind speeds. In our GEV approach, we have not accounted for the possibility of the method predicting negative wind speeds, however this rarely happens in our case studies. The mean probability masses assigned to negative wind speed values by the GEV and the TN-GEV combination models are approximately 0.01% and 10^{-7}% for the ECMWF data, 0.33% and 10^{-3}% for the ALADIN-HUNEPS data, and 0.04% and 10^{-4}% for the UWME data. In an application to quantitative precipitation, Scheuerer (2014) considers the GEV distribution to be left-censored at zero assigning all mass below zero to exactly zero. This approach seems very appropriate for precipitation where there is often high probability of zero precipitation. However, it seems less appropriate for wind variables. Instead, one might consider a truncation of the GEV distribution similar to the truncated normal distribution in (5.4).

The regime-switching combination of the TN and the GEV or LN model allows to combine advantages of lighter and heavier-tailed distributions and generally re-

sults in improvements of the predictive performance. However, these combination models require a suitable covariate to formulate a model selection criterion. Here, we used a TN distribution if the ensemble median was below a threshold θ, and a heavier-tailed alternative otherwise. The median of the ensemble predictions works well for the case studies investigated here, however, results might change for other ensembles or observation data. A large variety of alternative choices comes to mind, for example, the model selection could be based on station-specific information or other weather variables. Thereby, the necessary step of finding a suitable covariate limits the flexibility of the combination models in practice as the adequacy of covariates might depend on the data at hand.

By contrast, the TN-LN mixture models combine lighter-tailed TN and heavier-tailed LN distributions as a weighted mixture and thus do not require the exclusive choice of one of multiple parametric families. In this light, they are more flexible than models based on single parametric distributions or regime-switching combination models. From a practical perspective, they show significantly better calibration compared to the aforementioned alternatives, however, their estimation requires substantially more computational resources due to the larger number of parameters and the lack of a closed-form analytical expression of the CRPS. Therefore, we have investigated maximum likelihood estimation of the parameters as a more parsimonious, and thus practically viable alternative. In the three case studies ML estimation yields slightly better calibration, whereas minimum CRPS estimation results in slightly better verification scores.

Regarding computational aspects, the EMOS models based on single parametric distributions clearly require the least computational resources. On average, the LN model shows slightly lower computation times compared to the TN and GEV model, see Baran and Lerch (2015) for detailed results. By the bipartite nature of the model estimation for the combination models, the computation times are longer, but typically still do not exceed a few seconds for all forecast cases on one day in the verification period. Similarly, ML estimation of the TN-LN mixture model parameters requires more computational resources due to the higher number of parameters to be estimated. On the other hand, minimum CRPS estimation of the TN-LN mixture model results in the highest computational costs by far, with mean computation times exceeding those of the alternative models by a factor of more than 500. In light of the similar predictive performance, we therefore prefer ML estimation of the TN-LN mixture model parameters over minimum CRPS estimation. In general, however, the differences in the computation times are negligible from an operational point of view when compared to the amount of time and resources needed to generate the forecast ensemble.

Alternative postprocessing models have been proposed in the literature. Recently, Scheuerer and Möller (2015) use an EMOS model for wind speed based on gamma distributions. Bayesian implementations of BMA models for temperature (see, e.g., Bishop and Shanley, 2008; Di Narzo and Cocchi, 2010) allow for incorporating uncertainty information in both observations and parameters, and might help to overcome the effects of missing data. However, they are computationally quite demanding compared to the frequentist approach taken here, and might

thus be infeasible for a large number of stations and observations. On the other hand, there exist non-parametric approaches to statistical postprocessing that avoid the choice of a parametric family (Hamill and Whitaker, 2006; Flowerdew, 2014), but suffer from other limitations, see Gneiting (2014).

As illustrated in Section 5.3.3, estimation of the model parameters requires choices regarding the spatial composition of the training sets. For all models discussed above, we have used the regional (or global) approach where data from all available stations are composited to form a single training set for all stations. While this approach works well for the investigated data sets, the results are likely to change for larger ensemble domains with substantial differences in the climatological properties of observations and forecast errors. A local parameter estimation approach where only data from the single station of interest are used often results in better predictive performance, but can lead to unstable optimization algorithms and requires long training periods. In the subsequent chapter, we propose alternative approaches that combine the advantages of regional and local parameter estimation.

A general issue with all of the above considerations is the inherently univariate nature of the presented approaches. Accounting for temporal and spatial dependencies of weather trajectories is, however, critically important for many applications such as renewable energy forecasting. Therefore, multivariate approaches to statistical postprocessing are an important focus of recent interest and compelling topic for future research.

Multivariate postprocessing techniques can be separated into parametric and non-parametric methods. Parametric approaches model spatial, temporal and inter-variable dependencies in specific, typically low-dimensional settings, for example by suitably adapted EMOS and BMA variants (Berrocal et al., 2007; Möller et al., 2013; Baran and Möller, 2015; Feldmann et al., 2015). On the other hand, non-parametric approaches combine univariate postprocessing and reordering methods by imposing dependence structures from ensemble forecasts or historical observations (Schefzik et al., 2013; Wilks, 2015). See also Schefzik (2015) for a recent overview and further examples.

Despite the univariate nature of the results presented here, they are still of interest in a multivariate context. For example. non-parametric multivariate postprocessing techniques require suitable univariate models for weather variables, and parametric techniques might benefit from similarity-based semi-local approaches to parameter estimation that will be investigated in the subsequent chapter due to the generally large number of parameters that have to be estimated.

6 Similarity-based semi-local estimation of EMOS models

> It is far better to foresee even without certainty than not to foresee at all.[1]
>
> Henri Poincaré, 1913

In this chapter, we propose semi-local methods for estimating coefficients of the EMOS models introduced in Chapter 5. The training data for a specific observation station are augmented with corresponding forecast cases from stations with similar characteristics. Similarities between stations are determined using either distance functions or clustering based on various features of the climatology, forecast errors, ensemble predictions and locations of the observation stations. The present chapter is based on Lerch and Baran (2016).

6.1 Introduction

In Section 5.1, we have highlighted the importance of ensemble forecasts for the transition from deterministic to probabilistic weather prediction. However, as illustrated by the examples in Chapter 5, ensemble prediction systems often fail to correctly represent the uncertainty in the forecasts and require statistical postprocessing.

Recent developments in ensemble forecasting include multi-model ensemble prediction systems such as the THORPEX Interactive Grand Global Ensemble (TIGGE, Swinbank et al., 2016) where several single-model ensembles each based on multiple runs of individual NWP models are combined, see, e.g., Johnson and Swinbank (2009); Hagedorn et al. (2012). Another example is the Grand Limited Area Model Ensemble Prediction System (GLAMEPS, Iversen et al., 2011) considered here which is described in more detail in Section 6.2.

In the chapter at hand, we apply the truncated normal EMOS model of Thorarinsdottir and Gneiting (2010) for statistical postprocessing of wind speed forecasts of the GLAMEPS ensemble. The GLAMEPS ensemble covers a large domain across Europe and Northern Africa, however, only a short period of data is available

This disparity between the spatial and temporal extent of the data set causes challenges in the numerical estimation of the parameters. A regional approach as pursued in Section 5.3 where data from all observation stations are composited

[1] Poincaré (2015, p. 129), first edition published 1913.

to form a single training set for all stations is undesirable due to the potentially significant differences in the climatological properties of the observation stations and forecast errors over the large ensemble domain. On the other hand, a local approach where only forecast cases from the single observation station of interest are considered for the parameter estimation proves to be problematic due to the limited amount of available training data and leads to numerical stability issues in the optimization algorithms.

We propose two similarity-based semi-local approaches to parameter estimation which combine advantages of local and regional estimation in order to account for these challenges. A distance-based approach uses data from stations with similar characteristics to augment the training data for a given station. Our novel clustering-based approach employs k-means clustering to obtain groups of similar observation stations with respect to various features which then form shared training sets for parameter estimation.

As we will demonstrate below, the proposed similarity-based semi-local models show significant improvement in predictive performance compared to the standard regional and local estimation methods. They further allow for estimating complex models without numerical stability issues and are computationally more efficient than local parameter estimation.

The remainder of this chapter is organized as follows. In Section 6.2, we introduce the GLAMEPS ensemble and the observation data. In Section 6.3, we generalize the formulation of the TN model to account for the multi-model structure of the GLAMEPS predictions, and propose similarity-based semi-local approaches to parameter estimation based on distance functions and clustering. In Section 6.4, we report the results of the case study based on the GLAMEPS data. We conclude with a discussion in Section 6.5.

6.2 GLAMEPS data

The GLAMEPS ensemble is a short-range multi-model EPS launched in 2006 as a part of the cooperation between the ALADIN and HIRLAM (High Resolution Limited Area Modelling) consortia. It operates on a large domain covering Europe, North Africa and the Northern Atlantic. The currently running version is a combination of the subensembles from two versions of the ALADIN and AROME combined model (ALARO model with ISBA and SURFEX schemes, see, e.g., Noilhan and Planton (1989) and Hamdi et al. (2014)), and two version of the HIRLAM model (with Kain-Fritsch and STRACO schemes, see, e.g., Kain and Fritsch (1990) and Sass (2002)). Each subensemble consists of 12 perturbed members and a control forecast, and half of the perturbed members are lagged by 6 hours (Deckmyn, 2014).

Our data set contains 52 ensemble members of 18 hour ahead forecasts of 10-meter wind speed for 1738 observation sites together with the corresponding validating observations for October 2 – November 25, 2013, and February 2 – May 18, 2014. The locations of the observation stations are shown in Figure 6.1.

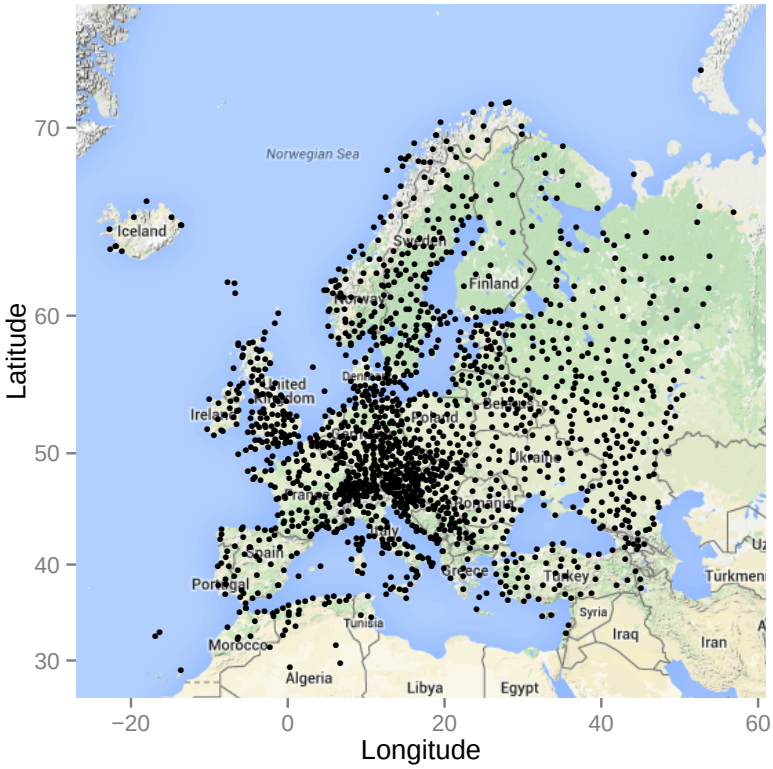


Figure 6.1: Map of Europe and Northern Africa showing the locations of the 1738 observation stations over which GLAMEPS forecasts are evaluated.

All maps in this chapter were produced using the `ggmap` package for `R` (Kahle and Wickham, 2013).

We divide the available data into two equally large periods from October 2013 to February 2014 and from March 2014 to May 2014 in order to allow for rolling training periods of sufficient length. The forecasts are evaluated over the second period. This out-of-sample verification period contains a total number of 137 302 individual forecast cases. Data from the first period are used to obtain training periods of equal lengths for all days, and to determine the similarities between the stations, see Section 6.3 for details.

While Iversen et al. (2011) apply BMA to calibrate GLAMEPS temperature forecasts, the case study at hand is the first application of postprocessing techniques to the corresponding wind speed forecasts. The data were provided by Maurice Schmeits, Jan Barkmeijer and John Bjørnar Bremnes.

6.3 Similarity-based semi-local models

Here, we propose new similarity-based semi-local approaches to estimating the model coefficients of the link functions connecting the parameters of the TN model to the ensemble predictions. Beforehand, we provide generalized specifications of the link functions to account for the specific structure of the GLAMEPS forecasts

which are based on multiple subensembles from distinct NWP models.

6.3.1 Generalized formulation of the TN model

We use the truncated normal EMOS model of Thorarinsdottir and Gneiting (2010) introduced in Section 5.3.2.1 to postprocess the GLAMEPS predictions of wind speed. Recall that wind speed y is modeled as

$$y|x_1, \ldots, x_M \sim \mathcal{N}_{[0,\infty)}(\mu, \sigma^2),$$

where the parameters μ and σ are suitably adapted functions of the ensemble forecasts $x_1, \ldots, x_M$.

We have also investigated the alternative wind speed models incorporating GEV or log-normal distributions that have been proposed in Section 5.3. However, as there are no substantial differences in the qualitative results and our focus here lies on the effect of different approaches to parameter estimation, we restrict our attention to the TN model.

The multi-model structure of the GLAMEPS ensemble suggests the existence of groups of exchangeable ensemble members, and thus requires a generalization of the model formulations introduced in Section 5.3.2. In what follows, if we have M ensemble members divided into m groups where the kth group contains $M_k \geq 1$ exchangeable members ($\sum_{k=1}^{m} M_k = M$), notation $x_{k,\ell}$ is used for the ℓth member of the kth group. Ensemble members within a given group are exchangeable and should thus share the same coefficient of the location parameter resulting in the predictive distribution

$$\mathcal{N}_{[0,\infty)}\left(a_0 + a_1 \sum_{\ell_1=1}^{M_1} x_{1,\ell_1} + \cdots + a_m \sum_{\ell_m=1}^{M_m} x_{m,\ell_m}, b_0 + b_1 S^2\right), \tag{6.1}$$

where S^2 denotes the ensemble variance. As before, the EMOS model coefficients $a_0, a_1, \ldots, a_m, b_0, b_1$ have to be estimated from training data. Model formulations that take into account the existence of groups in modeling the variance, for example by considering affine functions of the within-group variances, have also been investigated, but result in slightly worse predictive performance, potentially due to the larger number of parameters that have to be estimated. In Section 6.4, we will further investigate different specifications of the model groups for the GLAMEPS predictions.

In the light of the general model formulation (6.1), the ensemble prediction systems in Section 5.3.1 were introduced in ascending order of the number groups of members, m. The ECMWF ensemble only contains 1 group of 50 exchangeable members, the ALADIN-HUNEPS predictions consist of 2 groups of $M_1 = 1$ control and $M_2 = 10$ perturbed members, and the UWME predictions consist of $m = 8$ groups of size 1 each.

6.3.2 Similarity-based semi-local parameter estimation

As described above in Section 5.2.2, the coefficients of EMOS models are generally estimated by minimizing the mean value of a proper scoring rule, typically the CRPS, of the predictive distributions over suitably chosen rolling training periods consisting of the preceding n days.

Two basic approaches for selecting the spatial composition of the training data are given by regional and local methods. The regional (or global) approach composites ensemble forecasts and validating observations from all available stations during the rolling training period. Therefore, one obtains a single universal set of parameters across the entire ensemble domain, which is then used to produce forecasts at all observation sites. In the case of the GLAMEPS ensemble this means that a single set of coefficients is used for the wide-ranging domain, and the geographical and climatological variability might thus not be sufficiently taken into account. While the regional approach to parameter estimation can be implemented without numerical stability issues and offers slight gains in predictive performance compared to the raw ensemble (see Section 6.4), there is room for further improvement for large and heterogeneous domains.

By contrast, the local approach produces distinct parameter estimates for different stations by using only past forecast-observation pairs of the given station. Local models typically result in better predictive performance compared to regional models (see, e.g., Thorarinsdottir and Gneiting, 2010; Schuhen et al., 2012), however, these training sets contain only one observation per day and the estimation of local EMOS models thus requires significantly longer training periods to avoid numerical stability issues. For example, in case of the GLAMEPS data, model (6.1) has 15 parameters to be estimated which makes the use of local EMOS problematic. In a recent case study on EMOS models for the ECMWF ensemble, Hemri et al. (2014) find that training period lengths between 365 and 1816 days give the best results for local parameter estimation. For the GLAMEPS data, choosing such long training periods is impossible as the whole data set consists of only 161 days.

We propose two alternative similarity-based semi-local approaches which avoid the problems that make both regional and local estimation of the EMOS coefficients undesirable for the GLAMEPS data. The basic idea of the semi-local methods is to combine the advantages of regional and local estimation by augmenting the training data for a given station with data from stations with similar characteristics. The choice of similar stations is either based on suitably defined distance functions, or on clustering.

6.3.2.1 Distance-based semi-local model

Following Hamill et al. (2008), the training sets of a given station are increased by including training data from other stations with similar features. The similarity between stations is determined based on suitably defined distance functions. We use the term *distance function* in a general sense with only one of the proposed similarity measures depending on the actual geographical locations

of the observation stations. From a mathematical point of view, all considered distance functions are semimetrics, i.e., non-negative and symmetric functions $d : \{1, \ldots, 1738\} \times \{1, \ldots, 1738\} \to \mathbb{R}$ with $d(i, i) = 0$. Distance functions can thus be seen as negatively oriented similarity measures with smaller values indicating more similar characteristics of the stations of interest.

Note that compared to Hamill et al. (2008), we consider alternative choices of distance functions, and our forecasts are evaluated over a set of observation stations whereas the forecasts and analysis data used by Hamill et al. (2008) are given on a grid where different conclusions may apply.

Generally, the distance between two stations i and j denoted by $d(i, j)$ with $i, j \in \{1, \ldots, 1738\}$ is determined using the first period of available data from October 2013 to February 2014 which is distinct from the verification period. In the semi-local estimation of the EMOS model for a given station i_0, we then add the corresponding forecast cases in the rolling training period from the L most similar stations, i.e., the L stations with the smallest distances $d(i_0, j),\ j \in \{1, \ldots, 1738\}$.

Alternatively, one could also iteratively determine the similarities anew in every rolling training period. However, this approach requires lots of computational resources as the $\frac{1737 \cdot 1738}{2} \approx 1.5 \cdot 10^6$ pairwise distances between stations have to be re-computed for every training period, and is thus infeasible due to the large number of observation stations. In particular, note that already the non-iterative distance-based model estimation with a fixed set of similarities is computationally more demanding compared to local parameter estimation which arises as special case for $L = 1$. Furthermore, initial tests did not indicate substantial improvements in the predictive performance for the GLAMEPS data, we thus limit our discussion to the use of a fixed period of data for determining the similarities.

We investigate the following five distance functions.

Distance 1: Geographical locations. The distance between stations i and j is given by the Euclidean distance of the locations $(\mathcal{X}_i, \mathcal{Y}_i)$ and $(\mathcal{X}_j, \mathcal{Y}_j)$ of the two stations, i.e.,

$$d^{(1)}(i, j) = \sqrt{(\mathcal{X}_i - \mathcal{X}_j)^2 + (\mathcal{Y}_i - \mathcal{Y}_j)^2}.$$

The Euclidean distance is employed here since the station locations in the data set are given on the linearly transformed model estimation grid. In general, the great-circle distance is a more appropriate distance measure for actual geographical locations on the globe.

Distance 2: Station climatology. Let $\hat{F}_i$ denote the empirical CDF of wind speed observations at station i over the first period of data. Similar to the distance function proposed by Hamill et al. (2008), the distance to station j is given by the normalized sum over the absolute differences of the respective empirical CDFs $\hat{F}_i$ and $\hat{F}_j$ evaluated at a set of fixed values S, i.e.,

$$d^{(2)}(i, j) = \frac{1}{|S|} \sum_{s \in S} \left| \hat{F}_i(s) - \hat{F}_j(s) \right|,$$

where $|S|$ denotes the cardinality of S. Here, we choose $S = \{0, 0.5, 1, \ldots, 14.5, 15\}$ (equidistant evaluation points between the minimum observation of 0 m s^{-1} and the 99th percentile of all observations at 15 m s^{-1}) and note that the obtained sets of similar stations are somewhat robust to minor changes in the definition of the set of evaluation points, e.g., setting $S = \{0, 1, \ldots, 20\}$ results in very similar sets of close stations.

Distance 3: Ensemble forecast errors. Denote the ensemble mean for station i and date t, by $\bar{x}_{i,t}$ and the corresponding verifying observation by $y_{i,t}$, then the forecast error $e_{i,t}$ of the ensemble mean is given by

$$e_{i,t} = \bar{x}_{i,t} - y_{i,t}.$$

The third distance function is based on the distribution of these forecast errors. To that end, we define the empirical CDF of the forecast errors at station i as

$$\hat{G}_i^e(z) = \frac{1}{|T|} \sum_{t \in T} \mathbb{1}\{\bar{x}_{i,t} - y_{i,t} \leq z\}, \tag{6.2}$$

where T denotes the set of dates in the first period of data. The distance between two stations i and j is then given by

$$d^{(3)}(i,j) = \frac{1}{|S'|} \sum_{s \in S'} \left| \hat{G}_i^e(s) - \hat{G}_j^e(s) \right|,$$

where $S' = \{-10, -9.5, -9, -8.5, \ldots, 0, \ldots, 8.5, 9, 9.5, 10\}$ denotes the set of fixed values at which the empirical CDFs of the forecast errors are evaluated.

Distance 4: Combination of distances 2 and 3. We add up the values of distances 2 and 3 to define a distance function which depends on both the climatology of the observations as well as the distribution of the forecast errors of the ensemble, i.e., with the above notation,

$$\begin{aligned} d^{(4)}(i,j) &= d^{(2)}(i,j) + d^{(3)}(i,j) \\ &= \frac{1}{|S|} \sum_{s \in S} \left| \hat{F}_i(s) - \hat{F}_j(s) \right| + \frac{1}{|S'|} \sum_{s \in S'} \left| \hat{G}_i^e(s) - \hat{G}_j^e(s) \right|. \end{aligned}$$

Distance 5: Ensemble characteristics. Schefzik (2016) proposes a similarity-based implementation of the Schaake shuffle using a distance function that depends on summary statistics of the ensemble. With $\bar{x}_{i,t}$ and $S_{i,t}$ denoting the mean and standard deviation of the ensemble member forecasts at station i and date t, the distance between station i and j is given by

$$d^{(5)}(i,j) := \sum_{t \in T} \sqrt{(\bar{x}_{i,t} - \bar{x}_{j,t})^2 + (S_{i,t} - S_{j,t})^2},$$

where T again denotes the set of dates during the first period of data.

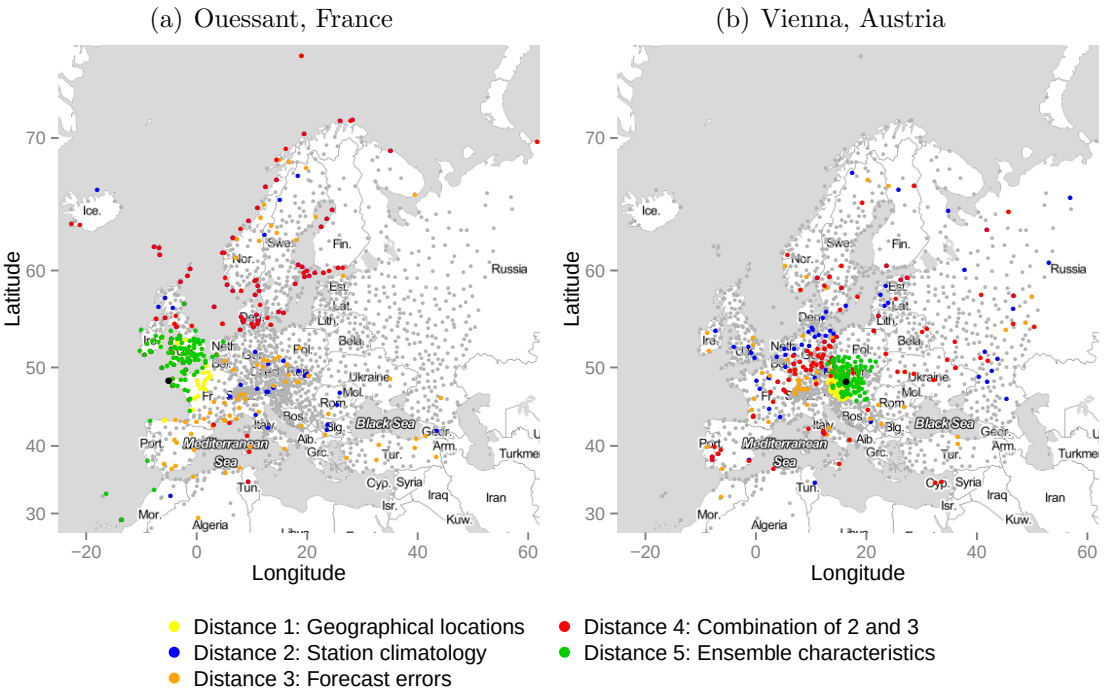


Figure 6.2: Illustration of the 100 most similar stations measured by the five distance functions for two reference stations at Ouessant, France (a) and Vienna, Austria (b). The reference stations are indicated by black dots. Note that several points are part of the set of similar stations in more than one similarity measure. In this case, the color of the last mentioned distance is assigned to them. See Figures 6.8 and 6.9 in Appendix 6.A for individual plots.

Figure 6.2 illustrates the five distance functions for two of the observation stations by displaying the 100 most similar stations in a specific color each. For both stations, a portion of the sets of most similar stations measured by two or more distance functions overlaps. See Figures 6.8 and 6.9 in Appendix 6.A for individual plots for the five distance functions and the two stations.

For the station at Ouessant (Figure 6.2a) located on the North-Western coast of France, it can be observed that the 100 most similar stations measured by the distance functions depending on the distribution of the observations and forecast errors (distances 2–4) are mostly located at coastal regions and islands in Northern Europe, in particular if these characteristics are combined (distance 4). By contrast, the most similar stations to the observation site at Vienna (Figure 6.2b) are distributed over continental central Europe and mostly located in France, Germany and Poland.

As implied by the definition, the most similar stations measured by distance 1 (and due to the large overlap also by distance 5) are located in close geographical proximity around the two observation sites. Due to the differences in the density of the observation station network, the stations in close geographical proximity to

the reference station at Ouessant are spread out over larger geographical distances compared to the respective stations around Vienna. Therefore, data from stations with different climatological properties might be added to the training sets for parameter estimation which indicates a potential drawback of the location-based distance 1.

6.3.2.2 Clustering-based semi-local model

Further, as an alternative to the distance-based approach we propose a novel semi-local approach based on cluster analysis. Here, the observation sites are grouped into clusters, and parameter estimation is performed for each cluster individually using only ensemble forecasts and validating observations at stations within the given cluster. To determine the clusters of observation stations we apply k-means clustering (see, e.g., Hastie et al., 2009) to various choices of feature sets which are based on climatological characteristics of the observation stations and the distribution of forecast errors, and are described in more detail below.

In comparison with the distance-based method, the clustering-based semi-local approach is computationally much more efficient as the parameter estimation is only performed for k distinct training sets for each given day, whereas the distance-based approach requires individual estimation of the coefficients at each of the 1738 stations with partially overlapping training sets. Further, the similarities between the observation stations are obtained in a more efficient way as clustering is computationally less demanding compared to the computation of pairwise distances between all observation stations (up to symmetry). In particular, clustering-based semi-local estimation is also computationally more efficient than local parameter estimation which arises as a special case with $k = 1738$ clusters of size 1 each. In this light, clustering-based semi-local models offer a compromise between adaptivity and parsimony of the numerical estimation.

The above discussion does not account for the computational costs of the actual clustering. However, there exist efficient algorithms for k-means clustering such as the Hartigan-Wong algorithm (Hartigan and Wong, 1979), which converge rapidly for the data at hand. The costs of the actual clustering are thus negligible compared to the computational costs of the numerical parameter estimation. In contrast to the distance-based approach, this allows for iteratively determining the clusters anew in every training period without a significant increase in the overall computational costs. This adaptive approach will be pursued for all clustering-based semi-local models discussed below.

We denote the number of features used in the k-means clustering procedure by N and consider the following feature sets.

Feature set 1: Station climatology. Let $\hat{F}_{i,n}$ denote the empirical CDF of the wind speed observations at station i over the rolling training period consisting of the preceding n forecast cases at this station. The feature set for station i is given by the set of equidistant quantiles of $\hat{F}_{i,n}$ at levels $\frac{1}{N+1}, \frac{2}{N+1}, \ldots, \frac{N}{N+1}$.

Feature set 2: Forecast errors. Denote the empirical CDF (6.2) of forecast errors $e_{i,t}$ by $\hat{G}^e_{i,n}(z)$. With a slight abuse of the above notation, the set T in

the expression $t \in T$ denotes the preceding n dates as the clusters are iteratively determined anew in every rolling training period. The feature set for station i is then given by the set of equidistant quantiles of $\hat{G}^e_{i,n}$ at levels $\frac{1}{N+1}, \frac{2}{N+1}, \ldots, \frac{N}{N+1}$.

Feature set 3: Combination of feature sets 1 and 2. To define a feature set that depends on both the station climatology and the distribution of forecast errors, we combine equidistant quantiles of $\hat{F}_{i,n}$ at levels $\frac{1}{N_1+1}, \ldots, \frac{N_1}{N_1+1}$ and equidistant quantiles of $\hat{G}^e_{i,n}$ at levels $\frac{1}{N_2+1}, \ldots, \frac{N_2}{N_2+1}$ into one single set of size $N = N_1 + N_2$, where N_1 and N_2 are defined as follows. If N is an even number, let $N_1 = N_2 = \frac{N}{2}$, otherwise let $N_1 = \lceil \frac{N}{2} \rceil$ and $N_2 = N - N_1$.

Alternative choices of feature sets where the geographical location of the observation stations is included in the definition have also been investigated, but result in a reduction of the predictive performance and are thus omitted in the following discussion.

Figure 6.3 illustrates the obtained clusters of observation stations for the different feature sets with a fixed number of $k = 5$ clusters. For the feature set defined in terms of the distribution of the observations (feature set 1, Figure 6.3a), one can observe two larger clusters distributed over central Europe, where one cluster mainly contains stations in Germany and France, while the other one contains most of the stations in the Alps and continental Eastern Europe. The remaining clusters are predominantly centered around the United Kingdom and coastal regions of France and Northern Europe. If the clusters are determined based on forecast errors (feature set 2, Figure 6.3b), the stations are mainly grouped into three almost equally large clusters, where the most notable difference compared to the fist feature set is the predominant presence of the third cluster in North-Eastern Europe. Further, the stations in the United Kingdom and coastal regions of Europe now mostly belong to the two biggest clusters rather than forming separate sets. Clustering based on a combination of the distribution of the observations and forecast errors (feature set 3, Figure 6.3c) results in a pattern of cluster memberships in between the other two choices. In particular, the alpine regions, continental Europe and the coastal regions around the United Kingdom show the most clear-cut separation compared to the other feature sets.

6.4 Case study

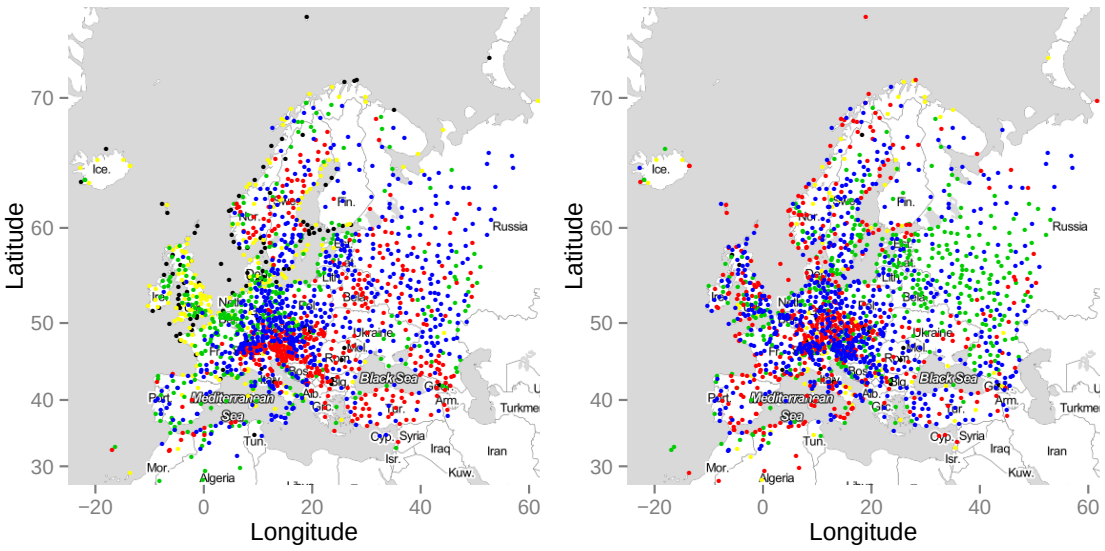
Here, we present the results of the case study based on the GLAMEPS data. In particular, we investigate various model formulations and the effect of tuning parameters of the semi-local approaches.

6.4.1 Model formulations

As discussed in Section 6.3, the link functions connecting the parameters of the predictive distribution of the EMOS models and the ensemble forecasts depend on the stochastic properties of the ensemble. To account for the multi-model

(a) Climatology

(b) Forecast errors



(c) Climatology + forecast errors

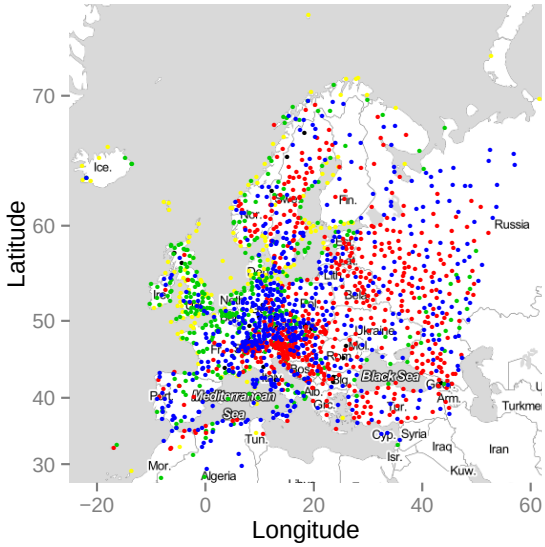


Figure 6.3: Illustration of cluster memberships of the observation stations based on feature sets 1 (a), 2 (b) and 3 (c) obtained with a fixed number of 5 clusters and 24 features. Colors are assigned to the clusters by size (in descending order: blue, red, green, yellow, black).

structure of the GLAMEPS predictions, we have introduced a generalized model formulation for the TN model in (6.1).

The GLAMEPS ensemble consists of four subensembles which differ in the choice of numerical model and parametrization scheme. Each subensemble contains a control forecast and $6 + 6$ (non-lagged and lagged) perturbed members. This induces a natural grouping into twelve groups:

$x_{\mathrm{AI},1}, \ldots, x_{\mathrm{AI},6}$	ALARO model with ISBA parameterization scheme
$x_{\mathrm{AS},1}, \ldots, x_{\mathrm{AS},6}$	ALARO model with SURFEX parameterization
$x_{\mathrm{HK},1}, \ldots, x_{\mathrm{HK},6}$	HIRLAM model with Kain-Fritsch parameterization
$x_{\mathrm{HS},1}, \ldots, x_{\mathrm{HS},6}$	HIRLAM model with STRACO parameterization
$x_{\bullet\mathrm{L},1}, \ldots, x_{\bullet\mathrm{L},6}$	lagged versions of above groups, 4 individual groups of size 6, where $\bullet \in \{\mathrm{AI}, \mathrm{AS}, \mathrm{HK}, \mathrm{HS}\}$
$x_{\mathrm{AI},c}, x_{\mathrm{AS},c}, x_{\mathrm{HK},c}, x_{\mathrm{HS},c}$	control forecasts, 4 individual groups of size 1.

The members within each individual group are exchangeable and should share a common set of EMOS coefficients, resulting in a predictive TN distribution with location

$$\begin{aligned} a_0 &+ a_{\mathrm{AI},c} x_{\mathrm{AI},c} + \sum_{\ell_1=1}^{6} \left(a_{\mathrm{AI}} x_{\mathrm{AI},\ell_1} + a_{\mathrm{AIL}} x_{\mathrm{AIL},\ell_1} \right) \\ &+ a_{\mathrm{AS},c} x_{\mathrm{AS},c} + \sum_{\ell_2=1}^{6} \left(a_{\mathrm{AS}} x_{\mathrm{AS},\ell_2} + a_{\mathrm{ASL}} x_{\mathrm{ASL},\ell_2} \right) \\ &+ a_{\mathrm{HK},c} x_{\mathrm{HK},c} + \sum_{\ell_3=1}^{6} \left(a_{\mathrm{HK}} x_{\mathrm{HK},\ell_3} + a_{\mathrm{HKL}} x_{\mathrm{HKL},\ell_3} \right) \\ &+ a_{\mathrm{HS},c} x_{\mathrm{HS},c} + \sum_{\ell_4=1}^{6} \left(a_{\mathrm{HS}} x_{\mathrm{HS},\ell_4} + a_{\mathrm{HSL}} x_{\mathrm{HSL},\ell_4} \right) \end{aligned} \tag{6.3}$$

and scale $b_0 + b_1 S^2$, which is a special case of model (6.1). This model has a total number of 15 parameters to be estimated and will be referred to as *full model*.

From a theoretical point of view, model (6.3) is the most appropriate specification. However, as local estimation of the full model proves difficult, we further investigate more parsimonious alternatives. A natural simplification is to ignore the existence of lag in the NWP model runs by assigning the same parameter values to the lagged and non-lagged members of a subensemble. That is, we set

$$a_{\mathrm{AI}} = a_{\mathrm{AIL}}, a_{\mathrm{AS}} = a_{\mathrm{ASL}}, a_{\mathrm{HK}} = a_{\mathrm{HKL}}, a_{\mathrm{HS}} = a_{\mathrm{HSL}}$$

in (6.3) which results in a reduced model with with location

$$\begin{aligned} a_0 + a_{\mathrm{AI},c} x_{\mathrm{AI},c} + \sum_{\ell_1=1}^{6} a_{\mathrm{AI}} \left(x_{\mathrm{AI},\ell_1} + x_{\mathrm{AIL},\ell_1} \right) \\ + a_{\mathrm{AS},c} x_{\mathrm{AS},c} + \sum_{\ell_2=1}^{6} a_{\mathrm{AS}} \left(x_{\mathrm{AS},\ell_2} + x_{\mathrm{ASL},\ell_2} \right) \\ + a_{\mathrm{HK},c} x_{\mathrm{HK},c} + \sum_{\ell_3=1}^{6} a_{\mathrm{HK}} \left(x_{\mathrm{HK},\ell_3} + x_{\mathrm{HKL},\ell_3} \right) \\ + a_{\mathrm{HS},c} x_{\mathrm{HS},c} + \sum_{\ell_4=1}^{6} a_{\mathrm{HS}} \left(x_{\mathrm{HS},\ell_4} + x_{\mathrm{HSL},\ell_4} \right) \end{aligned} \tag{6.4}$$

and 11 parameters to be estimated. This model will be referred to as *lag-ignoring model*.

Finally, we also investigate the situation where the existence of the aforementioned groups is ignored, and all ensemble members are assumed to form a single exchangeable group. In this case the predictive distribution is given by

$$\mathcal{N}_{[0,\infty)} \left(a_0 + a_1 \bar{x}, b_0 + b_1 S^2 \right), \tag{6.5}$$

where again, $\bar{x}$ denotes the ensemble mean. We refer to this model as *simplified model*.

6.4.2 Selection of tuning parameters for semi-local parameter estimation methods

Both semi-local parameter estimation techniques require the choice of various tuning parameters given by the length of the rolling training period, the number of similar stations to be taken into account, the number of features and the number of clusters. We now discuss the effect of these tuning parameters on the predictive performance of the forecast models. To that end, the full, lag-ignoring and simplified model were estimated using the distance-based and clustering-based semi-local parameter estimation techniques described in Section 6.3. Conclusions are drawn based on the mean CRPS over the evaluation period. For comparison, note that the average CRPS values of the GLAMEPS ensemble and the best regional TN model are 1.058 and 0.955, respectively.

Due to numerical stability issues in the parameter estimation, a comparison with local models is mostly impossible. An estimate of the mean CRPS of the locally estimated simplified TN model (6.5) can be obtained if problematic parameter estimates (around 0.1% of the forecast cases) are replaced by corresponding estimates from preceding forecast cases. The mean CRPS of the local simplified model with such subsequent modifications equals 0.790, see Section 6.4.3 for details on the numerical problems and required modifications.

6.4.2.1 Distance-based approach

In the distance-based semi-local approach to parameter estimation, the size of the training set for a given station i is increased by including corresponding training data from the L most similar stations, i.e., the L stations with the smallest distances $d(i, j)$, $j \in \{1, \ldots, 1738\}$. Note that for the distance functions defined in Section 6.3.2, $d(i, i) = 0$, a value of, e.g, $L = 5$ thus means that the training set for station i consists of data from this station, and of data from the 4 stations with the smallest distances to station i. Figure 6.4 illustrates the effect of the number of close stations on the predictive performance measured as mean CRPS of the three proposed models for selected lengths of the training period. Due to the large overlap of close stations determined by distance functions 1 and 5 (see, e.g., Figure 6.2) we omit the corresponding plots for distance 5 which closely resemble the plots for distance 1 and remark that similar conclusions apply, in particular for small values of L.

For distance 1 the predictive performance decreases with the number of similar stations added to the training sets, except for the more complex lag-ignoring and full models and shorter training periods, where the best mean CRPS values are attained for values around $L = 20$. Clearly, the inclusion of similar stations then allows for unproblematic parameter estimation, but generally, if the similarities are determined based on geographical locations as few stations as possible should be used in order to achieve results as close as possible to the favorable (but even for long training periods impossible) local parameter estimation corresponding to $L = 1$. Similar conclusions apply for the climatology-based distance 2, however, the predictive performance of these models is notably better.

A different pattern emerges for distances 3 and 4 based on forecast errors and combinations with climatology shown in the second row of Figure 6.4. In contrast to distances 1 and 2, augmenting the training sets with data from similar stations here generally improves the forecasts. The best predictive performances are achieved with choices of L between 10 and 30 depending on the similarity measure and the length of the training periods, whereas smaller values of L result in worse predictions. The mean CRPS increases for values of L exceeding around 30, however, note that these semi-local models still perform better than the local model for a wide range of tuning parameter values.

The effect of the length of the rolling training periods consisting of the preceding n days can also be seen from Figure 6.4 where each individual plot contains three different choices of n. Together with further investigations of plots of the average CRPS against the employed training period lengths which are omitted in the interest of brevity, one can observe that n only has a small effect on the predictive performance of the models. For all considered distance functions, the predictive performance increases slightly with longer training periods, in particular for the more complex models and smaller values of L. This is to be expected from the smaller size of the training sets as parameter estimation becomes problematic for short training periods and few additional forecast cases from similar stations taken into account.

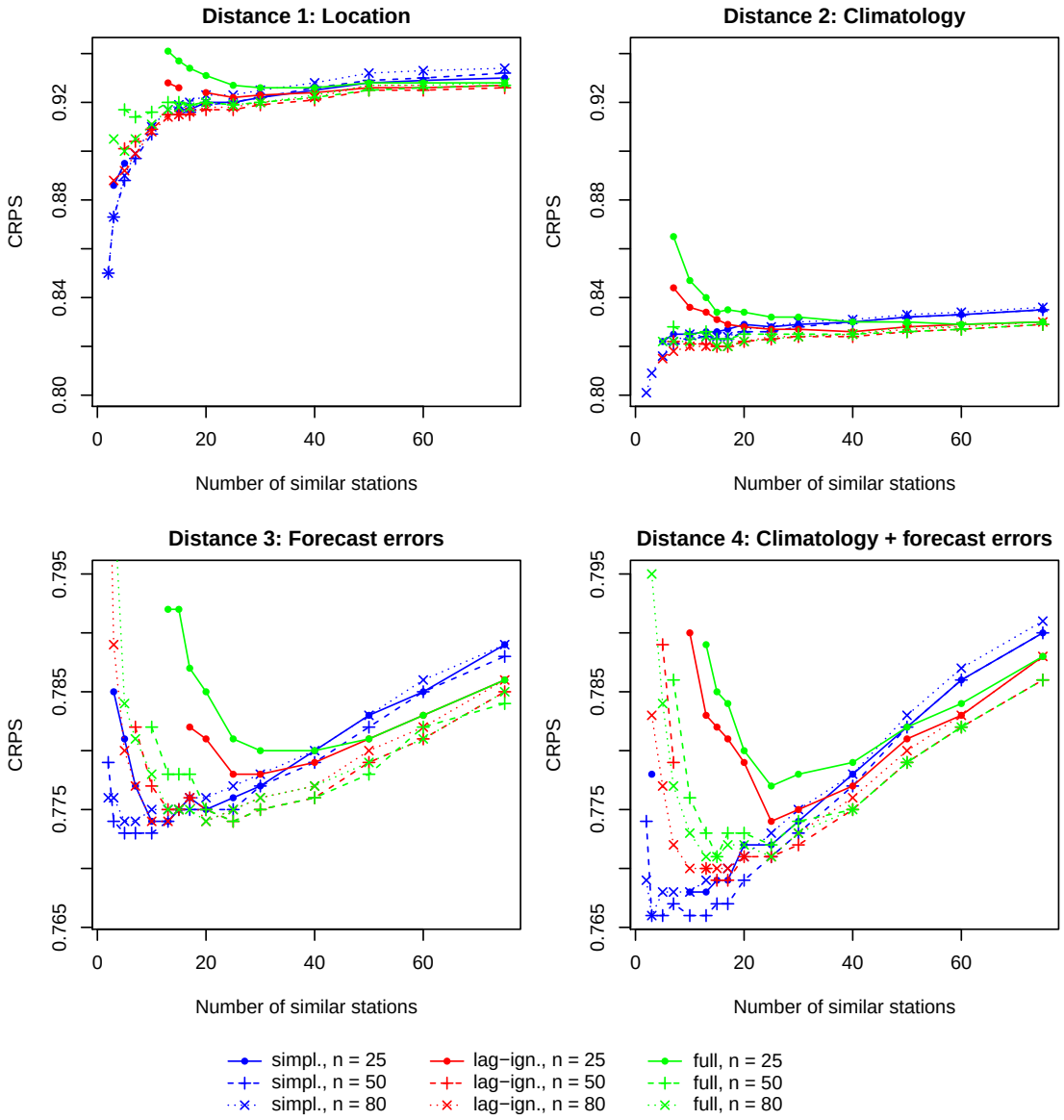


Figure 6.4: Effect of the number of similar stations L on the predictive performance of the distance-based semi-local models for three choices of training period lengths n (in days). Missing line segments indicate unsuccessful parameter estimation for these choices of tuning parameters. Note the different scales of the plots in the first and second row caused by the varying predictive performances of the respective models.

The simplified models show a slight decrease in predictive performance for training periods longer than 40–50 days, however, the differences are negligible compared to those between models based on varying choices of distance functions or varying numbers of similar stations taken into account. The overall best predictive performances across the three considered model formulations are achieved with training period lengths of 80 days.

6.4.2.2 Clustering-based approach

In the clustering-based semi-local approach k-means clustering based on the different feature sets is employed to group the observation stations into clusters. The lower computational costs of this approach allow for iterative computation of the clusters in every training period. This adaptive application of k-means clustering leads to improvements in mean CRPS of around 1–5% compared to a non-iterative implementation.

Figure 6.5 illustrates the effect of the number of clusters k on the predictive performance. Choosing $k = 1$ obviously corresponds to regional parameter estimation. For all three feature sets considered here, the predictive performance increases for larger values of k up to around 100 clusters except for shorter training periods. Clearly, a larger number of clusters allows for a more refined grouping into sets of observation stations with similar characteristics. The predictive performance generally decreases if much more than $k = 100$ clusters are used. This behavior is not surprising as the clusters become smaller and parameter estimation eventually becomes numerically unstable, particularly for the lag-ignoring and full models. Note that depending on training period length and feature set, only small improvements can be observed for k exceeding values of around 40 to 70 clusters.

As observed for the distance-based models, the clustering-based semi-local models defined in terms of the distribution of forecast errors and the station climatology (feature sets 2 and 3) are able to outperform the local model over a wide range of tuning parameter choices except for short training periods. The worse predictive performance for shorter training periods is to be expected as the smaller amount of forecasts cases used to determine the clusters might result in a less accurate partitioning of the observation stations. Compared to the distance-based approach it can be observed that for some k, training period lengths below 80 days are optimal. However, in comparison to the effect of different choices of feature sets the effect of the length of the training period is negligible.

Thus far, all clustering-based semi-local models shown in Figure 6.5 were estimated for a fixed feature set size of $N = 24$. To illustrate the effect of N on the predictive performance, Figure 6.6 shows the average CRPS of the clustering-based models as functions of the number of features N considered in k-means clustering for three choices of k. Given that sufficiently many features (around 5-10 depending on the other tuning parameters) are used, the feature set size has only a small effect on the predictive performance compared to different choices of k or n. Reasons for this behavior clearly include the aforementioned robustness

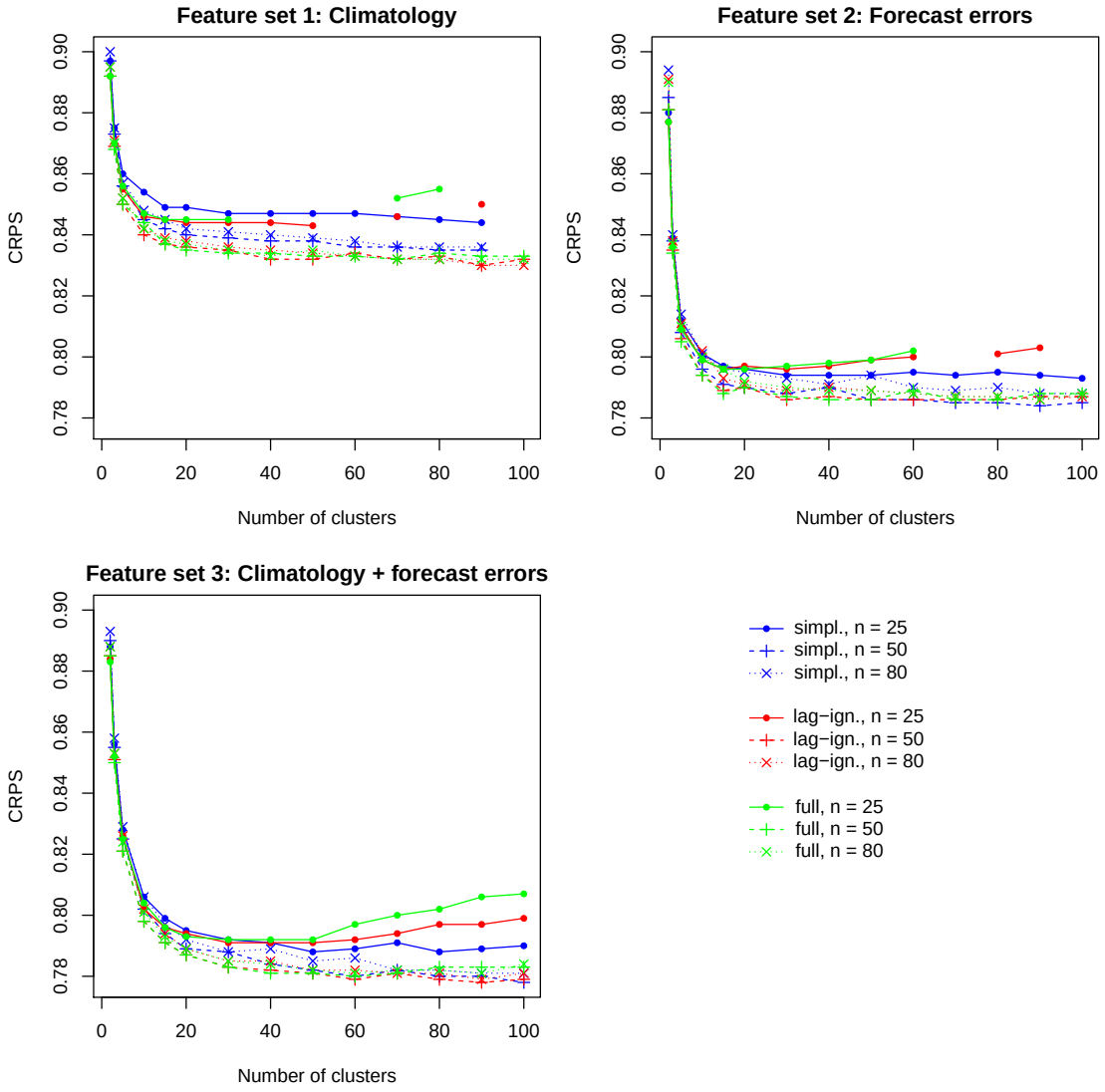


Figure 6.5: Effect of the number of clusters k on the predictive performance of clustering-based semi-local models for three choices of training period lengths n (in days). All models are estimated with feature sets of size $N = 24$. Missing line segments indicate unsuccessful parameter estimation for these choices of tuning parameters.

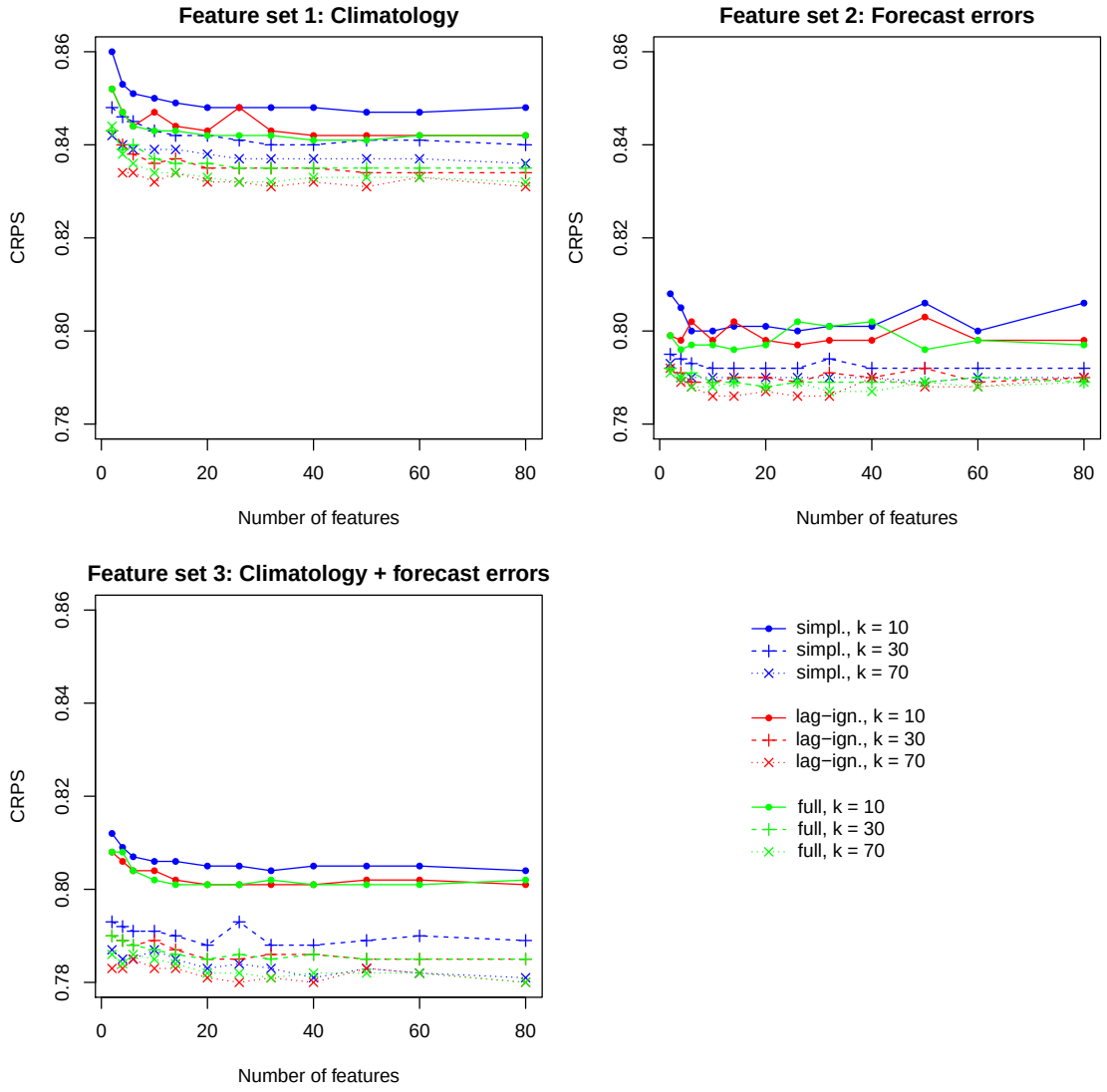


Figure 6.6: Effect of the size of the feature set N on the predictive performance of clustering-based semi-local models for three choices of numbers of clusters k. All models are estimated over a training period of 80 days. Missing line segments indicate unsuccessful parameter estimation for these choices of tuning parameters.

of the obtained cluster memberships with regards to N. The best results across all considered tuning parameter combinations are generally obtained for feature set sizes between 20 and 40 thus justifying our previous choice of $N = 24$.

6.4.3 Forecast performance

The predictive performance of the semi-local models is evaluated over the verification period March 1 – May 18, 2014 which contains 137 302 individual forecast cases. We use the local climatological forecasts given by the observations at the corresponding station during the rolling training periods, the raw GLAMEPS ensemble predictions, and probabilistic forecast by the regional TN model as benchmark models. While locally estimated models are desirable, the estimation of these models is highly problematic for the GLAMEPS data due to the issues discussed earlier. Even for the simplified model (6.5) with a maximum training period length of 80 days, numerical issues occur in the local parameter estimation, e.g., some shape parameters are estimated to be 0. In this case the problematic parameter estimates are replaced by the preceding ones. Note that such subsequent adjustments are not necessary for the semi-local or regional models. Further, neither the lag-ignoring nor the full local model can be successfully estimated as the employed numerical optimization algorithms fail to converge or produce numerical errors.

In the interest of brevity, we limit our discussion to the simplified and the lag-ignoring models. It can be seen from Figures 6.4–6.6 that the full semi-local models generally result in slightly worse predictive performance compared to the lag-ignoring models, therefore, the additional computational costs of taking into account the lagging in the subensembles are not justified. Note that different conclusions may apply for other ensemble prediction systems with lagged members.

With regard to the tuning parameters for the semi-local approaches, we employ a fixed training period length of 80 days, and use a fixed number of $N = 24$ features for k-means clustering to ensure comparability across the different models. For the individual distance-based and clustering-based semi-local models we then choose suitable values for the number of most similar stations L and the number of clusters k from Figures 6.4–6.6. While the chosen tuning parameter combinations might not be the overall optimal values for the individual models, the results hold for a wide range of tuning parameter choices as indicated by the sensitivity considerations in Section 6.4.2.

To determine the optimal tuning parameter values for a new data set we suggest to follow common practice from the extant literature on ensemble postprocessing, and to test various combinations of parameter values, perhaps on a shorter initial test set, similar to the approach we have taken in Section 5.3. For the GLAMEPS ensemble, our analysis indicates that the most influential tuning parameters for the semi-local model estimation are the number of similar stations L, and the number of clusters k, respectively, see Section 6.4.2 for details.

Table 6.1 shows the average CRPS, MAE of median values, and coverage and

Table 6.1: Mean CRPS, MAE, coverage and width of 96.2% prediction intervals of probabilistic 18 hour ahead forecasts of wind speed evaluated over the verification period from March to May 2014. A training period length of 80 days is used for all models. For the clustering-based model estimation, a fixed number of $N = 24$ features is applied.

Forecast		CRPS (m s^{-1})	MAE (m s^{-1})	Coverage (%)	Width (m s^{-1})
Local climatology		1.127	1.580	96.6	7.96
GLAMEPS ensemble		1.058	1.376	67.1	3.50
Regional TN models					
simpl.		0.957	1.324	90.3	6.36
lag-ign.		0.955	1.320	90.3	6.33
Local TN models (with subsequent modifications)					
simpl.		0.790	1.100	88.7	5.12
Distance-based semi-local TN models					
D1 simpl.	$L = 3$	0.873	1.218	90.2	5.99
D1 lag-ign.	$L = 3$	0.887	1.236	89.2	5.71
D2 simpl.	$L = 5$	0.816	1.136	90.0	5.61
D2 lag-ign	$L = 5$	0.815	1.136	89.6	5.42
D3 simpl.	$L = 5$	0.774	1.083	90.3	5.25
D3 lag-ign.	$L = 10$	0.774	1.083	90.2	5.21
D4 simpl.	$L = 3$	0.766	1.069	89.9	5.16
D4 lag-ign.	$L = 10$	0.770	1.075	90.0	5.18
D5 simpl.	$L = 3$	0.874	1.220	90.2	5.95
D5 lag-ign.	$L = 5$	0.895	1.248	89.8	5.91
Clustering-based semi-local TN models					
C1 simpl.	$k = 70$	0.836	1.162	89.8	5.68
C1 lag-ign.	$k = 70$	0.832	1.156	89.6	5.55
C2 simpl.	$k = 70$	0.789	1.103	89.9	5.25
C2 lag-ign.	$k = 70$	0.787	1.099	89.8	5.22
C3 simpl.	$k = 70$	0.782	1.091	89.7	5.19
C3 lag-ign.	$k = 70$	0.781	1.090	89.7	5.17

average width of 96.2% prediction intervals for the considered models. The raw GLAMEPS ensemble predictions outperform the climatological forecasts and provide sharp prediction intervals, however, at the cost of being uncalibrated. Regional TN models are able to improve the calibration of the ensemble, and result in around 10% better mean CRPS values, however, the semi-local approaches significantly outperform the regional approaches for all considered models and tuning parameter choices, see also Figures 6.4 and 6.5.

Among the distance-based semi-local models, the best predictive performances are obtained by distance functions 3 and 4 which utilize the distribution of forecast errors and combinations with the station climatology to determine similarities between stations. Note that these semi-local models are also able to outperform the local TN model for a wide range of tuning parameter choices without requiring subsequent corrections and while further allowing for a successful estimation of the more complex lag-ignoring and full semi-local models. The semi-local models based on distance functions 1 and 5 exhibit similar predictive performances which are slightly worse compared to the other distances, but are still able to outperform the regional model. The similarity is clearly caused by the large overlap of selected similar stations, see Figure 6.2. Except for distance 2, the simplified model performs slightly better than the lag-ignoring model, however, the differences are negligible compared to the differences between the different model estimation approaches.

We obtain similar results for the clustering-based semi-local models which perform slightly worse compared to the corresponding distance-based models, however, still outperform the regional models and the local model if the clusters are determined based on forecast errors and station climatology. Here, the lag-ignoring models show better predictive performances compared to the simplified models, but again, the differences are small compared to the influence of the choice of feature sets.

Figure 6.7 shows the VR histogram for the GLAMEPS predictions and PIT histograms of the regional, the local, and the semi-local models with the best average CRPS values. Compared to the raw ensemble forecasts, all postprocessing models exhibit substantially improved calibration with PIT histograms showing much smaller deviations from the desired uniform distribution. The hump-shaped PIT histogram of the regional TN model indicates a slight under-prediction of lower wind speed values. The local and semi-local models are able to correct for this deficiency and show slightly better calibration, in particular for the semi-local models. However, all models consistently show a slight under-dispersion that can also be seen from the coverage values reported in Table 6.1. This deficiency appears to be a general drawback of models based on the TN distribution, see also the results of the previous case studies reported in Section 5.4. Alternative distributional choices proposed above might lead to further improvements in calibration. The novel semi-local estimation approaches might be of particular interest for the TN-LN mixture models where the large number of parameters impedes local estimation.

To conclude, we note that the overall best predictive performance is achieved by

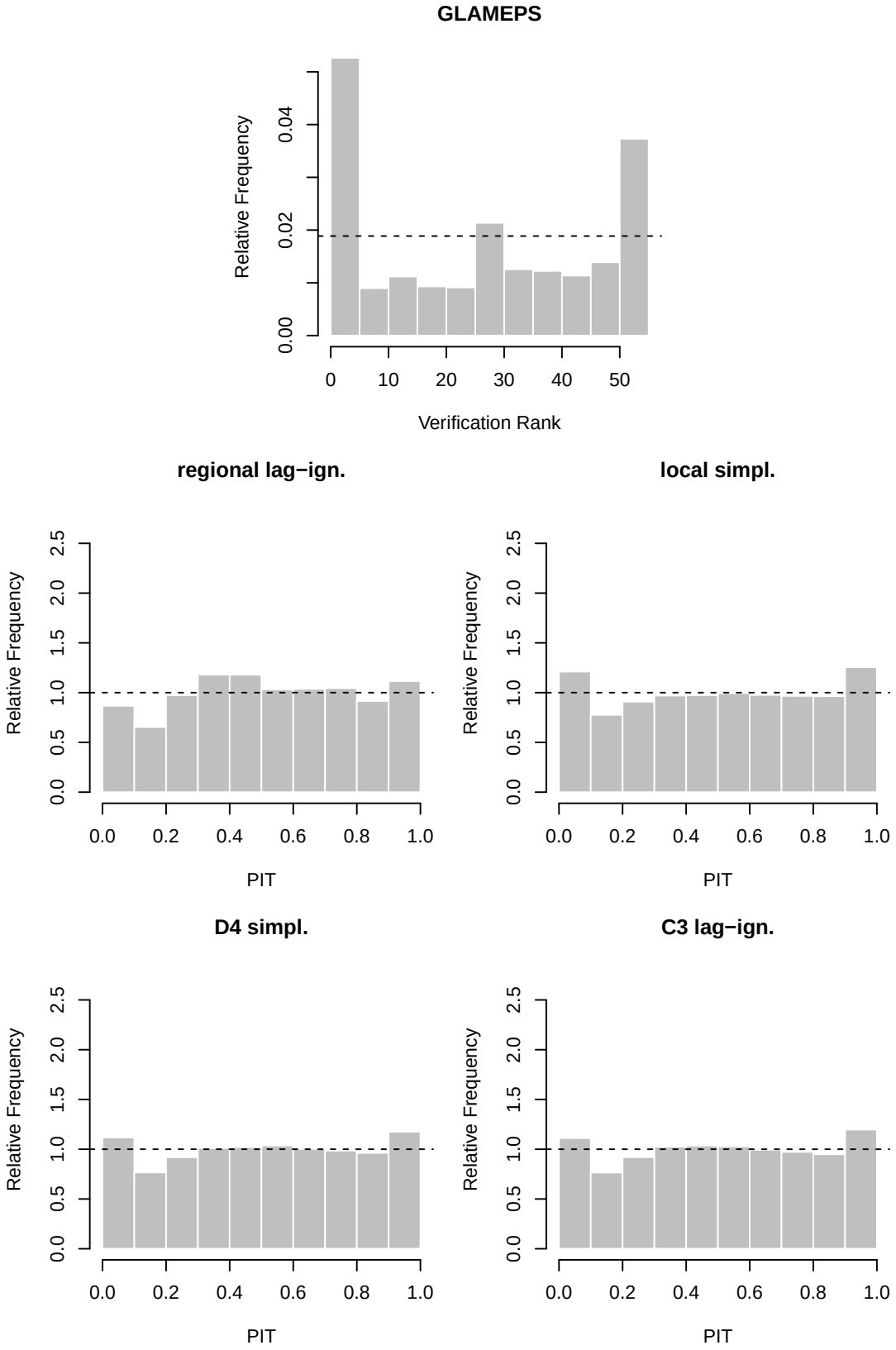


Figure 6.7: VR histogram for the raw ensemble and PIT histograms for selected EMOS postprocessed forecasts for the GLAMEPS data.

semi-local models where the similarities between stations are determined based on combinations of the distributions of observations and forecast errors at the given stations. While all semi-local models show significantly better predictive performance than the regional models, these best models are also able to outperform the locally estimated model. The semi-local parameter estimation methods further allow for estimating more complex models without numerical issues, whereas local estimation is only possible for simplified model formulations with a reduced number of parameters and still requires subsequent modifications. Figures 6.4 and 6.5 indicate that these conclusions hold for a wide range of tuning parameter choices.

6.5 Discussion

We have proposed two semi-local approaches to parameter estimation for ensemble postprocessing where the training data for a given observation station are augmented with data from stations with similar characteristics. The distance-based approach roughly follows the ideas of Hamill et al. (2008) and uses distance functions to determine the similarities between observations stations, whereas the novel clustering-based approach employs k-means clustering to obtain groups of similar stations.

The semi-local models outperform regional and local models and offer several advantages over these standard approaches to parameter estimation while being straightforward to implement. The clustering-based semi-local model estimation is further computationally much more efficient than local estimation. While distance-based semi-local models show slightly better predictive performance compared to the clustering-based models, the estimation requires substantially more computational resources. In particular, an iterative computation of the similarities in every training period is not feasible for the distance-based models.

Compared to the work of Hamill et al. (2008), we propose several alternative distance functions and use the distance-based approach for observations at specific stations instead of gridded data. It would be interesting to apply the novel similarity measures as well as the clustering-based approach to grid-based forecast and analysis data and assess potential differences. In particular, similarity measures incorporating the distribution of forecast errors (distances 3 and 4) might also offer improvements over the climatology-based distance function used by Hamill et al. (2008) when applied to gridded data. In connected works, Kleiber et al. (2011), Scheuerer and Büermann (2014), and Scheuerer and Möller (2015) consider alternative approaches incorporating techniques from geostatistics and novel model formulations that entail local adaptivity of the parameters, and allow for extrapolating the forecasts to locations or grid points without observations. These schemes are particularly important for interpolating local forecasts obtained at observation stations to the model grid.

The distance functions considered here are defined in terms of station loca-

tions, observations, forecast errors of the ensemble and mean and variance of the ensemble forecast. It might appear somewhat surprising that models based on similarities defined by characteristics of the ensemble (mean and variance) as measured by distance 5 do not result in improvements compared to simple location-based similarities (distance 1). However, this might be due to the fact that these characteristics of the ensemble are substantially influenced by the locations of the stations, and the training sets thus largely overlap with those of the location-based distance 1. These results might change for other ensemble prediction systems. Potential improvements might be obtained by including different summary statistics of the ensemble, e.g., by adding information about the within-group variances of the subensembles, or quantiles of the distribution of ensemble forecasts. Alternative choices of similarity measures proposed in related works may further improve the predictive performance. For example, Kleiber et al. (2011) include covariates such as elevation and land use information. Pursuing similar approaches was not possible for the GLAMEPS data as such covariate information was not available for the data at hand.

The group memberships of the observation stations in the clustering-based semi-local models are determined by k-means clustering. Alternative clustering methods exist and might potentially lead to improvements (for reviews and comparisons see, e.g., Fraley and Raftery, 1998; Kaufman and Rousseeuw, 2009; Wilks, 2011). We did not incorporate informations on the geographical locations of the stations or characteristics of the ensemble into the selected feature sets as initial tests indicated a worse predictive performance. For different ensemble prediction systems, alternative choices of feature sets may lead to further improvements.

Junk et al. (2015) propose analog-based local EMOS models where the training set for a given station is chosen by selecting forecast cases with similar ensemble forecasts for that station. This analog-based approach thus utilizes information for a given station in an optimal way by selecting subsets of the local training sets, whereas our semi-local models combine informations from multiple observation stations based on similarities. While the analog-based modification of the local parameter estimation method shows good predictive performance in a case study on hub height wind speed, it requires sufficiently long training periods for locally selecting similar forecast cases. The implementation of this analog-based approach is thus infeasible for the GLAMEPS data, however, comparisons and combinations with the similarity-based semi-local approaches proposed here are of interest and might result in further improvement in predictive performance.

Both the analog-based approach of Junk et al. (2015) and the similarity-based semi-local models proposed here can be viewed within the broader framework of customized training in transductive learning (Powers et al., 2015). Mathematical results from the respective machine learning literature for classification problems can potentially provide theoretical justification for the presented empirical evidence, see Bottou and Vapnik (1992).

Appendix 6.A Illustration of similarity measures

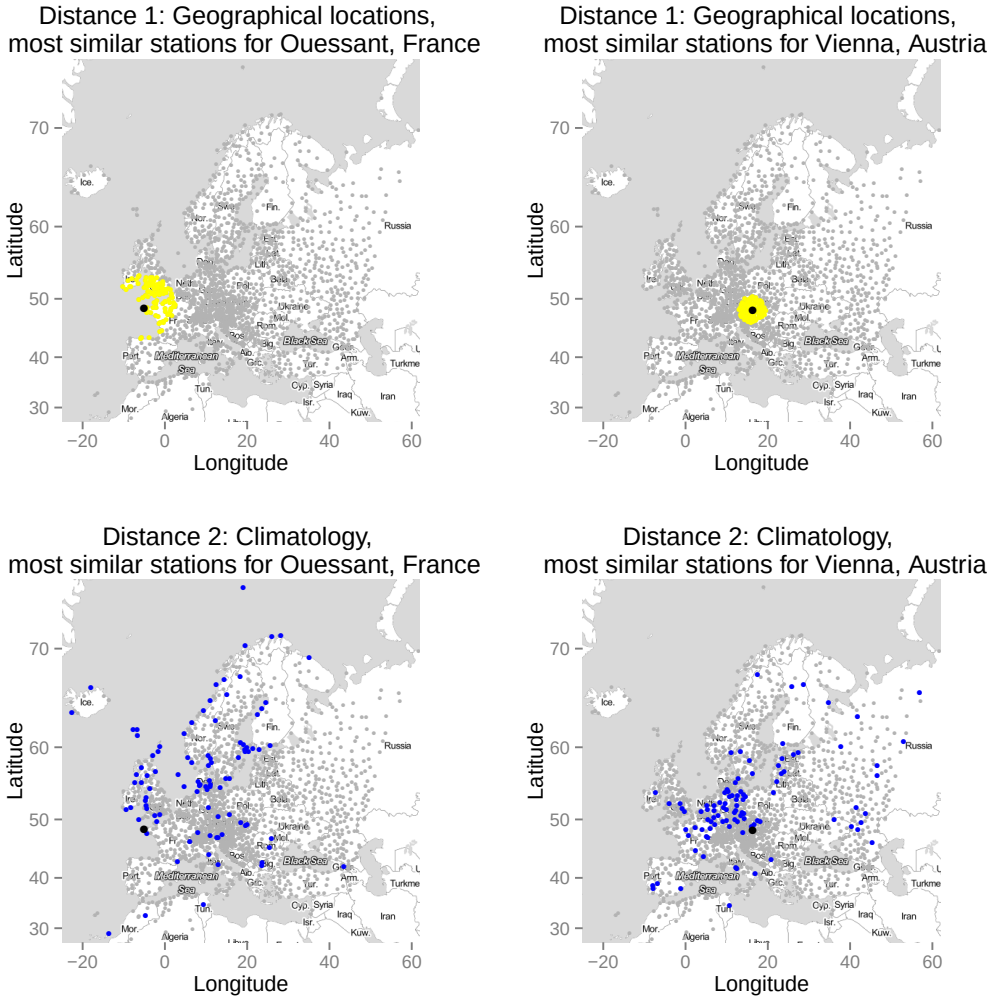


Figure 6.8: Illustration of the 100 most similar stations measured by distance functions 1 and 2 for two reference stations at Ouessant, France (left column) and Vienna, Austria (right column).

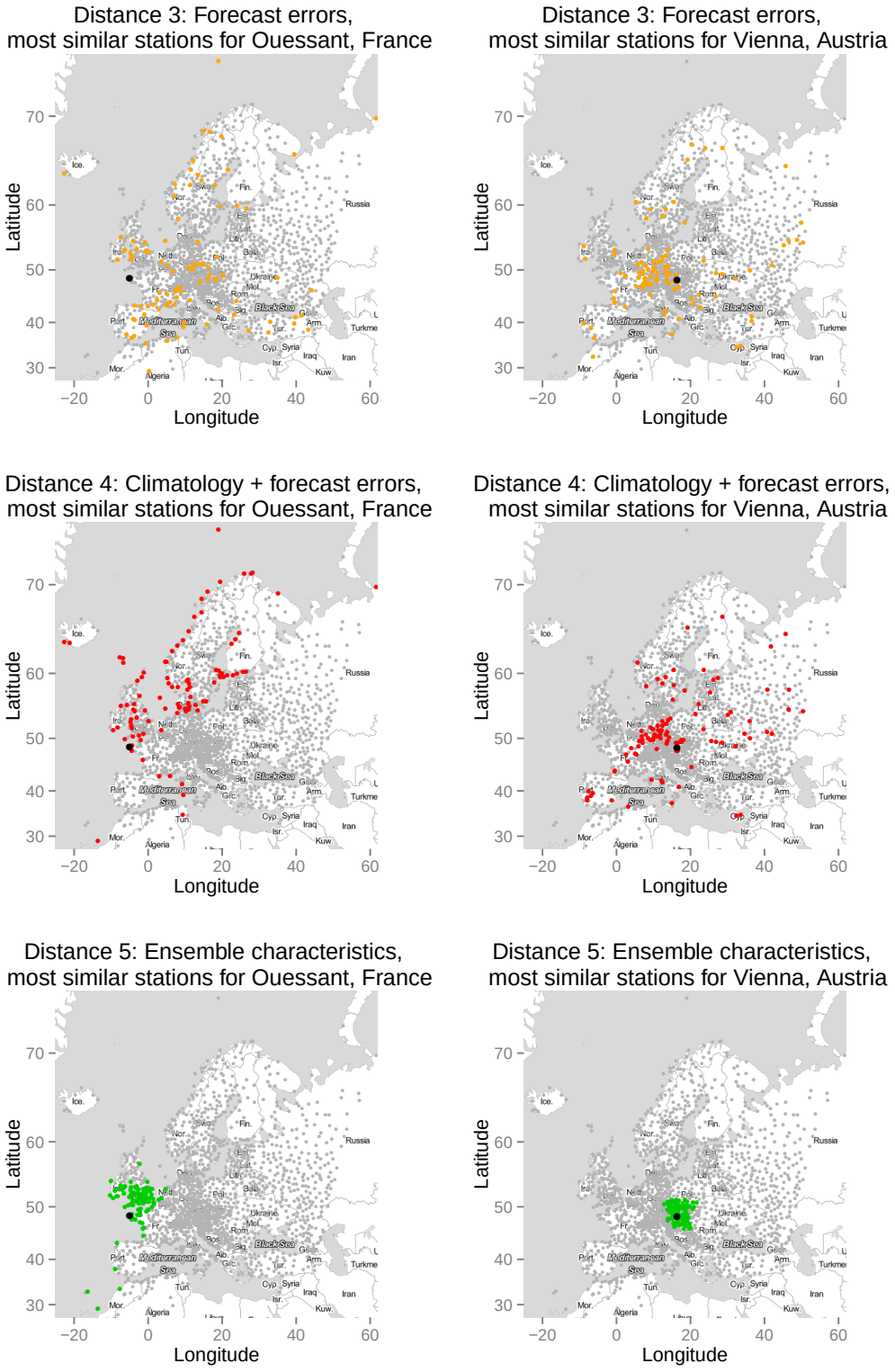


Figure 6.9: Illustration of the 100 most similar stations measured by distance functions 3–5 for two reference stations at Ouessant, France (left column) and Vienna, Austria (right column).

7 | Conclusion

It's difficult to make predictions, especially about the future.[1]

Danish proverb

Throughout this thesis, we have demonstrated that not only making, but also evaluating forecasts can be a challenging exercise. We have seen that these two tasks are closely intertwined and should be addressed jointly based on a sound theoretical framework. The concluding remarks in this chapter aim to summarize and discuss the main results, and to give an overview of directions for future research.

As indicated above, the decision theoretical framework of proper scoring rules constitutes a key tool for probabilistic forecasting and comparative model evaluation. In Chapter 3, we investigated how to evaluate probabilistic forecasts with an emphasis on extreme events. Suitably adapted weighted scoring rules provide measures of forecast quality that can be flexibly adapted to the situation at hand. In particular, they allow for proper forecast verification for extreme events. They can thus be seen as a remedy of the forecaster's dilemma frequently observed in public discussions of forecast quality where the evaluation is often restricted to subsets of selected extreme events and skillful forecasts are thereby potentially discredited. However, as demonstrated in simulation experiments, the practical benefits of using proper weighted scoring rules in terms of statistical power for model comparisons might be limited by the disadvantageous dependence on the tail properties of the forecast distributions in finite samples.

In Chapter 4, we moved the focus to the close connections of making and evaluating forecasts and investigated how to estimate the unknown underlying forecast distribution based on simulation output in Bayesian forecasting methods. The theoretical framework of proper scoring rules provided a foundation for our proposed notion of consistency that allows to assess the adequacy of approximation methods from a theoretical perspective. Conditions under which choices from the literature are consistent relative to popular scoring rules were derived with the help of classical asymptotic results. These considerations were illustrated in simulation experiments and a case study. Consistency and efficiency of approximation methods strongly depend on the scoring rule of interest. The empirical CDF and a mixture-of-parameters estimator that exploits the structure of the Bayesian model are consistent under weak regularity assumptions and work well in practical applications. By contrast, nonparametric kernel density estimation

[1] Frequently attributed to Niels Bohr (e.g., Kac, 1975, p. 5).

proved to be problematic, particularly when used in conjunction with the logarithmic score.

Chapters 5 and 6 focused on applications in numerical weather prediction where proper scoring rules are used in the statistical estimation of model parameters by minimizing the average score over suitably chosen training sets of past forecasts and observations. We addressed two topics in probabilistic wind speed forecasting based on statistical postprocessing of ensemble forecasts. In Chapter 5, we investigated the choice of a suitable parametric model for non-homogeneous regression approaches to postprocessing forecast ensembles. Models based on distributions with heavy right tails and various combination approaches are able to improve the predictive performance of the standard truncated normal model, particularly for high wind speed observations. In Chapter 6, we investigated similarity-based approaches for selecting the training sets for estimating the model parameters. Augmenting training data for a specific observation station with data from similar sites improves the predictive performance and allows for efficiently estimating complex models without numerical stability issues.

As indicated in Section 5.5, a common shortcoming and relevant starting point for future research is the inherently univariate nature of the presented approaches to making and evaluating probabilistic forecasts. Extensions of the theoretical framework of forecast evaluation based on an assessment of calibration and sharpness, and proper scoring rules to higher dimensional spaces are straightforward, however, the practical application of suitable measures of calibration and scoring rules is more involved compared to the univariate case.

Based on work of Gneiting et al. (2008), various methods for multivariate calibration assessment have been developed. While the multivariate rank histogram proposed by Gneiting et al. (2008) works well in low-dimensional settings (Schuhen et al., 2012; Schefzik et al., 2013), the multivariate ordering loses power in higher-dimensions (Thorarinsdottir et al., 2016). The band depth rank histogram approach proposed by Thorarinsdottir et al. (2016) aims to overcome this shortcoming and scales efficiently to higher dimensional settings.

Several multivariate extensions of various univariate proper scoring rules have been proposed over the last years. A multivariate generalization of the CRPS to $\mathbb{R}^d$ is given by the *energy score* (ES; Gneiting and Raftery, 2007; Gneiting et al., 2008),

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

where $\| \cdot \|$ denotes the Euclidean norm on $\mathbb{R}^d$, $y \in \mathbb{R}^d$ is the observation vector, and X, X' are independent d-dimensional random vectors with distribution F such that $\mathbb{E}_F \|X\|$ is finite. The corresponding representation of the univariate CRPS has been introduced in equation (2.4). The energy score has appealing theoretical properties and can be readily computed for multivariate ensemble forecasts via an empirical variant similar to equation (4.8), however, it is often not sufficiently sensitive to detect misspecifications of the correlation structure (Pinson and Girard, 2012; Pinson and Tastu, 2013; Scheuerer and Hamill, 2015b).

An alternative is given by a multivariate extension of the Dawid-Sebastiani score,

$$\mathrm{DSS}_d(F, y) = \log \det \Sigma_F + (y - \mu_F)' \, \Sigma_F^{-1} \, (y - \mu_F) \,,$$

where μ_F and Σ_F denote the mean vector and covariance matrix of the forecast distribution F, respectively. DSS_d is strictly proper relative to any class of distributions characterized by the first and second moment, however, applications to ensemble forecasts are problematic as Σ_F (and Σ_F^{-1}) have to be estimated from the ensemble (Scheuerer and Hamill, 2015b). To overcome the shortcomings of ES and DSS_d, Scheuerer and Hamill (2015b) propose a multivariate proper scoring rule based on the concept of variograms from geostatistics. The *variogram score* (VS) of order $p > 0$ is given by

$$\mathrm{VS}^p(F, y) = \sum_{i,j=1}^{d} w_{ij} \left(|y_i - y_j|^p - \mathbb{E}_F |X_i - X_j|^p \right)^2 ,$$

where $w_{ij} > 0$ are weights, and X_i and X_j are the ith and jth component of a d dimensional random vector X with distribution F. Suggested typical choices for p include 0.5 and 1.

In light of the results presented in Chapter 3, it might be of interest to investigate how to evaluate multivariate probabilistic forecasts with an emphasis on extreme events. Adaptations of multivariate proper scoring rules towards suitable weighted variants are likely not straightforward, but are potentially important for various applications considering the multivariate nature of many relevant problems in extreme value theory (see, e.g., Coles and Tawn, 1991). The multivariate proper scoring rules introduced above are typically computed by empirical approximations based on samples from the forecast distribution F. Therefore, multivariate extensions of the concepts and results presented in Chapter 4 might be of interest for making and evaluating multivariate probabilistic forecasts based on Bayesian methods. For example, Bayesian VAR models such as those used in Section 3.4 are popular in the econometric literature as they allow for jointly modeling multiple variables of interest, however, the evaluation is typically restricted to the use of univariate scoring rules. Multivariate approaches to numerical weather prediction via statistical postprocessing of ensemble forecasts have been discussed in Section 5.5. The availability of suitable multivariate proper scoring rules is of critical importance to assess the quality and benefits of such approaches.

Bibliography

Aaronson, J., Burton, R., Dehling, H., Gilat, D., Hill, T. and Weiss, B. (1996). Strong laws for L- and U-statistics. *Transactions of the American Mathematical Society*, 348, 2845–2866.

Abernethy, J. D. and Frongillo, R. M. (2012). A characterization of scoring rules for linear properties. Journal of Machine Learning, Workshop and Conference Proceedings, 25th Annual Conference on Learning Theory. Available at `http://jmlr.org/proceedings/papers/v23/abernethy12/abernethy12.pdf`.

Adams, D. (1996). *The Ultimate Hitchhiker's Guide*. Wings Books, New York.

Adcock, S. A. and McCammon, J. A. (2006). Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chemical Reviews*, 106, 1589–1615.

Adolfson, M., Linde, J. and Villani, M. (2007). Forecasting performance of an open economy DSGE model. *Econometric Reviews*, 26, 289–328.

Albeverio, S., Jentsch, V. and Kantz, H. (eds.) (2006). *Extreme Events in Nature and Society*. Springer.

Amisano, G. and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business and Economic Statistics*, 25, 177–190.

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

Araújo, M. B. and New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, 22, 42–47.

Banerjee, A., Merugu, S., Dhillon, I. S. and Ghosh, J. (2005). Clustering with Bregman divergences. *The Journal of Machine Learning Research*, 6, 1705–1749.

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

Baran, S. (2014). Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Computational Statistics & Data Analysis*, 75, 227–238.

Baran, S., Horányi, A. and Nemoda, D. (2013). Statistical post-processing of probabilistic wind speed forecasting in Hungary. *Meteorologische Zeitschrift*, 22, 273–282.

Baran, S., Horányi, A. and Nemoda, D. (2014). Comparison of BMA and EMOS statistical calibration methods for temperature and wind speed ensemble weather prediction. *Időjárás*, 118, 217–241.

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

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

Bauer, P., Thorpe, A. and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525, 47–55.

Bauschke, H. H., Borwein, J. M. and Combettes, P. L. (2001). Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces. *Communications in Contemporary Mathematics*, 3, 615–647.

Bauwens, L., Koop, G., Korobilis, D. and Rombouts, J. V. K. (2014). The contribution of structural break models to forecasting macroeconomic series. *Journal of Applied Econometrics*, 30, 596–620.

Beirlant, J., Goegebeur, Y., Teugels, J. and Segers, J. (2004). *Statistics of Extremes*. John Wiley & Sons, Chichester.

Bentzien, S. and Friederichs, P. (2012). Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Weather and Forecasting*, 27, 988–1002.

Berg, T. O. and Henzel, S. R. (2015). Point and density forecasts for the Euro area using Bayesian VARs. *International Journal of Forecasting*, 31, 1067–1095.

Bernoulli, J. (1713). *Ars Conjectandi*. Impensis Thurnisiorum, Basileae. Reproduction of original from Sterling Memorial Library, Yale University. Online edition of Gale Digital Collections: The Making of the Modern World: Part I: The Goldsmiths'-Kress Collection, 1450-1850. Available at `http://nbn-resolving.de/urn%3Anbn%3Ade%3Agbv%3A3%3A1-146753`.

Bernoulli, J. (2006). *The Art of Conjecturing, together with Letter to a Friend on Sets in Court Tennis, translated and with an introduction and notes by Edith Dudley Sylla*. John Hopkins Univ. Press, Baltimore.

Berrocal, V. J., Gelfand, A. E. and Holland, D. M. (2014). Assessing exceedance of ozone standards: A space-time downscaler for fourth highest ozone concentrations. *Environmetrics*, 25, 279–291.

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

Bishop, C. H. and Shanley, K. T. (2008). Bayesian model averaging's problematic treatment of extreme weather and a paradigm shift that fixes it. *Monthly Weather Review*, 136, 4641–4652.

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

Bottou, L. and Vapnik, V. (1992). Local learning algorithms. *Neural Computation*, 4, 888–900.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2, 107–144.

Brandt, P. T., Freeman, J. R. and Schrodt, P. A. (2014). Evaluating forecasts of political conflict dynamics. *International Journal of Forecasting*, 30, 944–962.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7, 200–217.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.

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

Bröcker, J. and Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22, 382–388.

Bröcker, J. and Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. *Tellus A*, 60, 663–678.

Bücher, A. and Segers, J. (2016). On the maximum likelihood estimator for the Generalized Extreme-Value distribution. Preprint, available at `http://arxiv.org/abs/1601.05702`.

Buizza, R. (2006). The ECMWF ensemble prediction system. In *Predictability of Weather and Climate* (T. N. Palmer and R. Hagedorn, eds.), chap. 17. Cambridge University Press, 459–488.

Buizza, R., Miller, M. and Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125, 2887–2908.

Buizza, R. and Palmer, T. N. (1995). The singular-vector structure of the atmospheric global circulation. *Journal of the Atmospheric Sciences*, 52, 1434–1456.

Buizza, R., Tribbia, J., Molteni, F. and Palmer, T. N. (1993). Computation of optimal unstable structures for a numerical weather prediction model. *Tellus A*, 45, 388–407.

Buja, A., Stuetzle, W. and Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. Available at `http://www.stat.washington.edu/people/wxs/Learning-papers/paper-proper-scoring.pdf`.

Byrne, S. (2016). A note on the use of empirical AUC for evaluating probabilistic forecasts. *Electronic Journal of Statistics*, 10, 380–393.

Carriero, A., Clark, T. E. and Marcellino, M. (2015a). Bayesian VARs: Specification choices and forecast accuracy. *Journal of Applied Econometrics*, 30, 46–73.

Carriero, A., Clark, T. E. and Marcellino, M. (2015b). Common drifting volatility in large Bayesian VARs. *Journal of Business and Economic Statistics*, 33, in press.

Carriero, A., Clark, T. E. and Marcellino, M. (2015c). Realtime nowcasting with a Bayesian mixed frequency model with stochastic volatility. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 178, 837–862.

Carriero, A., Mumtaz, H. and Theophilopoulou, A. (2015d). Macroeconomic information, structural change, and the prediction of fiscal aggregates. *International Journal of Forecasting*, 31, 325–348.

Casati, B., Ross, G. and Stephenson, D. (2004). A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorological Applications*, 11, 141–154.

Celik, A. N. (2004). A statistical analysis of wind power density based on the Weibull and Rayleigh models at the southern region of Turkey. *Renewable Energy*, 29, 593–604.

Chen, M.-H. and Shao, Q.-M. (1997). Performance study of marginal posterior density estimation via Kullback-Leibler divergence. *Test*, 6, 321–350.

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

Christensen, H. M., Moroz, I. M. and Palmer, T. N. (2015). Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, 141, 538–549.

Clark, T. E. (2011). Real-time density forecasts from BVARs with stochastic volatility. *Journal of Business and Economic Statistics*, 29, 327–341.

Clark, T. E. and Ravazzolo, F. (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *Journal of Applied Econometrics*, 30, 551–575.

Cloke, H. L. and Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, 375, 613–626.

Cogley, T. S. M. and Sargent, T. J. (2005). Drifts and volatilities: Monetary policies and outcomes in the post-World War II U.S. *Review of Economic Dynamics*, 8, 262–302.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.

Coles, S. G. and Tawn, J. A. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 53, 377–392.

Cooke, W. E. (1906). Forecasts and verifications in Western Australia. *Monthly Weather Review*, 34, 23–24.

Cooley, D., Davis, R. A. and Naveau, P. (2012). Approximating the conditional density given large observed values via a multivariate extremes framework, with application to environmental data. *The Annals of Applied Statistics*, 6, 1406–1429.

Corradi, V. and Swanson, N. R. (2006). Bootstrap conditional distribution tests in the presence of dynamic misspecification. *Journal of Econometrics*, 133, 779–806.

Courtney, J. F., Lynch, P. and Sweeney, C. (2013). High resolution forecasting for wind energy applications using Bayesian model averaging. *Tellus A*, 65, 19669.

Craiu, R. V. and Rosenthal, J. S. (2014). Bayesian computation via Markov chain Monte Carlo. *Annual Review of Statistics and Its Application*, 1, 179–201.

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

Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59, 77–93.

Dawid, A. P., Musio, M. and Ventura, L. (2016). Minimum scoring rule inference. *Scandinavian Journal of Statistics*, 43, 123–138.

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

De la Cruz, R. and Branco, M. D. (2009). Bayesian analysis for nonlinear regression model under skewed errors, with application in growth curves. *Biometrical Journal*, 51, 588–609.

Deckmyn, A. (2014). Introducing GLAMEPSv2. ALADIN Forecasters Meeting, Ankara, Turkey, September 10–11, 2014. Available at: `http://www.cnrm.meteo.fr/aladin/meshtml/FM2014/presentation/AladinFm_AD_be.pdf`.

Dehling, H. and Philipp, W. (2002). Empirical process techniques for dependent data. In *Empirical Process Techniques for Dependent Data* (H. Dehling, T. Mikosch and M. Sørensen, eds.). Birkhäuser Boston, 3–113.

Delatola, E.-I. and Griffin, J. E. (2011). Bayesian nonparametric modelling of the return distribution with stochastic volatility. *Bayesian Analysis*, 6, 901–926.

Denrell, J. and Fang, C. (2010). Predicting the next big thing: Success as a signal of poor judgment. *Management Science*, 56, 1653–1667.

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P. and Cébron, P. (2014). PEARP, the Météo-France short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 141, 1671–1685.

Di Narzo, A. F. and Cocchi, D. (2010). A Bayesian hierarchical approach to ensemble weather forecasting. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 59, 405–422.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business and Economic Statistics*, 33, 1–9.

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

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253–263.

Diks, C., Panchenko, V. and van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163, 215–230.

Doswell, C. A., Davies-Jones, R. and Keller, D. L. (1990). On summary measures of skill in rare event forecasting based on contingency tables. *Weather and Forecasting*, 5, 576–585.

Easterling, D. R., Meehl, G. A., Parmesan, C., Changnon, S. A., Karl, T. R. and Mearns, L. O. (2000). Climate extremes: Observations, modeling, and impacts. *Science*, 289, 2068–2074.

Eckel, F. A. and Mass, C. F. (2005). Aspects of effective mesoscale, short-range ensemble forecasting. *Weather and Forecasting*, 20, 328–350.

ECMWF Directorate (2012). Describing ECMWF's forecasts and forecasting system. ECMWF Newsletter No. 133 – Autumn 2012. Available at `http://www.ecmwf.int/sites/default/files/elibrary/2012/14576-newsletter-no133-autumn-2012.pdf`.

Eguchi, S. and Copas, J. (2006). Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma. *Journal of Multivariate Analysis*, 97, 2034–2040.

Ehm, W. and Gneiting, T. (2012). Local proper scoring rules of order two. *The Annals of Statistics*, 40, 609–637.

Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations, and forecast rankings. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 78, 505–562.

Ehrman, C. M. and Shugan, S. M. (1995). The forecaster's dilemma. *Marketing Science*, 14, 123–147.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin.

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

European Wind Energy Association (2012). Wind in power: 2012 European statistics. Available at: `http://www.ewea.org/fileadmin/files/library/publications/statistics/Wind_in_power_annual_statistics_2012.pdf`.

Faust, J. and Wright, J. H. (2009). Comparing Greenbook and reduced form forecasts using a large realtime dataset. *Journal of Business and Economic Statistics*, 27, 468–479.

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

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140, 1917–1923.

Ferro, C. A. T., Richardson, D. S. and Weigel, A. P. (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, 15, 19–24.

Ferro, C. A. T. and Stephenson, D. B. (2011). Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, 26, 699–713.

Feuerverger, A. and Rahman, S. (1992). Some aspects of probability forecasting. *Communications in Statistics – Theory and Methods*, 21, 1615–1632.

Fissler, T., Ziegel, J. F. and Gneiting, T. (2016). Expected shortfall is jointly elicitable with value-at-risk: Implications for backtesting. *Risk*, January, in press.

Flowerdew, J. (2014). Calibrating ensemble reliability whilst preserving spatial structure. *Tellus A*, 66, 22662.

Fox, E. B. and West, M. (2011). Autoregressive models for variance matrices: Stationary inverse Wishart processes. Preprint, available at `http://arxiv.org/abs/1107.5239`.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41, 578–588.

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

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

Fraley, C., Raftery, A. E., Sloughter, J. M. and Gneiting, T. (2015). *ensembleBMA: Probabilistic Forecasting using Ensembles and Bayesian Model Averaging*. R package version 5.1.1, URL `http://CRAN.R-project.org/package=ensembleBMA`.

Fricker, T. E., Ferro, C. A. T. and Stephenson, D. B. (2013). Three recommendations for evaluating climate predictions. *Meteorological Applications*, 20, 246–255.

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23, 579–594.

Frigessi, A., Haug, O. and Rue, H. (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5, 219–235.

Frigyik, B. A., Srivastava, S. and Gupta, M. R. (2008). Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54, 5130–5139.

Garcia, A., Torres, J. L., Prieto, E. and De Francisco, A. (1998). Fitting wind speed distributions: A case study. *Solar Energy*, 62, 139–144.

Gebhardt, C., Theis, S. E., Paulat, M. and Ben Bouallègue, Z. (2011). Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmospheric Research*, 100, 168–177.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2014a). *Bayesian Data Analysis*. 3rd ed. Chapman & Hall/CRC, Boca Raton.

Gelman, A., Hwang, J. and Vehtari, A. (2014b). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24, 997–1016.

Genton, M. and Hering, A. (2007). Blowing in the wind. *Significance*, 4, 11–14.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. John Wiley & Sons, Hoboken.

Geweke, J. and Amisano, G. (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting*, 26, 216–230.

Geweke, J. and Amisano, G. (2011). Hierarchical Markov normal mixture models with applications to financial asset returns. *Journal of Applied Econometrics*, 26, 1–29.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7, 473–483.

Geyer, C. J. (1995). Conditioning in Markov Chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 4, 148–154.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74, 1545–1578.

Giannone, D., Lenza, M. and Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, 97, 436–451.

Gneiting, T. (1997). Normal scale mixtures and dual probability densities. *Journal of Statistical Computation and Simulation*, 59, 375–384.

Gneiting, T. (2008). Editorial: Probabilistic forecasting. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 171, 319–321.

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

Gneiting, T. (2014). Calibration of medium-range weather forecasts. ECMWF Technical Memorandum 719. Available at `http://www.ecmwf.int/sites/default/files/elibrary/2014/9607-calibration-medium-range-weather-forecasts.pdf`.

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

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

Gneiting, T., Larson, K., Westrick, K., Genton, M. G. and Aldrich, E. (2006). Calibrated probabilistic forecasting at the Stateline Wind Energy Center: The regime-switching space–time method. *Journal of the American Statistical Association*, 101, 968–979.

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

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

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

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

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

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

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 14, 107–114.

Greenberg, E. (2013). *Introduction to Bayesian Econometrics*. 2nd ed. Cambridge University Press, New York.

Grell, G. A., Dudhia, J. and Stauffer, D. R. (1995). A description of the fifth-generation Penn state/NCAR mesoscale model (MM5). Technical Note NCAR/TN-398+STR. National Center for Atmospheric Research. Available at: `http://nldr.library.ucar.edu/repository/assets/technotes/TECH-NOTE-000-000-000-214.pdf`.

Grimit, E. P., Gneiting, T., Berrocal, V. J. and Johnson, N. A. (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.

Groen, J. J. J., Paap, R. and Ravazzolo, F. (2013). Real-time inflation forecasting in a changing world. *Journal of Business and Economic Statistics*, 31, 29–44.

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics*, 32, 1367–1433.

Gschlößl, S. and Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, 2007, 202–225.

Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia University Press, New York.

Gutmann, M. and Hirayama, J. (2011). Bregman divergence as general framework to estimate unnormalized statistical models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*. Corvallis, Oregon, 283–290.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series*. Springer, Berlin.

Hagedorn, R. (2010). On the relative benefits of TIGGE multi-model forecasts and reforecast-calibrated EPS forecasts. ECMWF Newsletter No. 124 – Summer 2010. Available at `http://www.ecmwf.int/sites/default/files/elibrary/2010/14601-newsletter-no124-summer-2010.pdf`.

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

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

Hagel, E. (2010). The quasi-operational LAMEPS system of the Hungarian Meteorological Service. *Időjárás*, 114, 121–133.

Haiden, T., Magnusson, L. and Richardson, D. (2014). Statistical evaluation of ECMWF extreme wind forecasts. ECMWF Newsletter No. 139 – Spring 2014. Available at `http://www.ecmwf.int/sites/default/files/elibrary/2014/14582-newsletter-no139-spring-2014.pdf`.

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *The Annals of Statistics* 1491–1519.

Hall, P., Lahiri, S. N. and Truong, Y. K. (1995). On bandwidth choice for density estimation with dependent data. *The Annals of Statistics*, 23, 2241–2263.

Hall, S. S. (2011). Scientists on trial: At fault? *Nature*, 477, 264–269.

Hamdi, R., Degrauwe, D., Duerinckx, A., Cedilnik, J., Costa, V., Dalkilic, T., Essaouini, K., Jerczynki, M., Kocaman, F., Kullmann, L., Mahfouf, J.-F., Meier, F., Sassi, M., Schneider, S., Váňa, F. and Termonia, P. (2014). Evaluating the performance of SURFEXv5 as a new land surface scheme for the ALADINcy36 and ALARO-0 models. *Geoscientific Model Development*, 7, 23–39.

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

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

Hamill, T. M., Hagedorn, R. and Whitaker, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Monthly Weather Review*, 136, 2620–2632.

Hamill, T. M. and Whitaker, J. S. (2006). Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review*, 134, 3209–3229.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57, 357–384.

Hart, J. D. and Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *The Annals of Statistics*, 18, 873–890.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 28, 100–108.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer, Berlin.

Heffernan, J. E., Stephenson, A. G. and Gilleland, E. (2014). *ismev: An Introduction to Statistical Modeling of Extreme Values*. R package version 1.40, URL `http://CRAN.R-project.org/package=ismev`.

Held, L., Rufibach, K. and Balabdaoui, F. (2010). A score regression approach to assess calibration of continuous probabilistic predictions. *Biometrics*, 66, 1295–1305.

Hemri, S., Haiden, T. and Pappenberger, F. (2016). Discrete post-processing of total cloud cover ensemble forecasts. *Monthly Weather Review*, 144, in press.

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

Hendrickson, A. D. and Buehler, R. J. (1971). Proper scores for probability forecasters. *The Annals of Mathematical Statistics* 1916–1921.

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

Hilden, J. and Gerds, T. A. (2014). A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Statistics in Medicine*, 33, 3405–3414.

Hogan, R. J. and Mason, S. (2012). Deterministic forecasts of binary events. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (I. T. Jolliffe and D. B. Stephenson, eds.), 2nd ed., chap. 3. John Wiley & Sons, 31–59.

Holland, M. and Ikeda, K. (2016). Minimum proper loss estimators for parametric models. *IEEE Transactions on Signal Processing*, 64, 704–713.

Holzmann, H. and Eulert, M. (2014). The role of the information set for forecasting – with applications to risk management. *Annals of Applied Statistics*, 8, 595–621.

Hooten, M. B. and Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85, 3–28.

Horányi, A., Kertész, S., Kullmann, L. and Radnóti, G. (2006). The ARPEGE/ALADIN mesoscale numerical modeling system and its application at the Hungarian Meteorological Service. *Időjárás*, 110, 203–227.

Hosking, J. R. M. (1985). Algorithm AS 215: Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 34, 301–310.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35, 73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1. 221–233.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6, 695–709.

Ikeda, S. (1960). A remark on the convergence of Kullback-Leibler's mean information. *Annals of the Institute of Statistical Mathematics*, 12, 81–88.

Iversen, T., Deckmyn, A., Santos, C., Sattler, K., Bremnes, J. B., Feddersen, H. and Frogner, I.-L. (2011). Evaluation of 'GLAMEPS' – a proposed multimodel EPS for short range forecasting. *Tellus A*, 63, 513–530.

Jeon, J. and Taylor, J. W. (2012). Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association*, 107, 66–79.

Johnson, C. and Swinbank, R. (2009). Medium-range multimodel ensemble combination and calibration. *Quarterly Journal of the Royal Meteorological Society*, 135, 777–794.

Jordan, A. (2015). Closed form expressions for the continuous ranked probability score. Available at `https://github.com/FK83/scoringRules/blob/master/crps.pdf`.

Jordan, A., Krüger, F. and Lerch, S. (2016). *The scoringRules package*. URL `https://github.com/FK83/scoringRules`.

Junk, C., Delle Monache, L. and Alessandrini, S. (2015). Analog-based ensemble model output statistics. *Monthly Weather Review*, 143, 2909–2917.

Justus, C. G., Hargraves, W. R., Mikhail, A. and Graber, D. (1978). Methods for estimating wind speed frequency distributions. *Journal of Applied Meteorology*, 17, 350–353.

Juutilainen, I., Tamminen, S. and Röning, J. (2012). Exceedance probability score: A novel measure for comparing probabilistic predictions. *Journal of Statistical Theory and Practice*, 6, 452–467.

Kac, M. (1975). Some reflections of a mathematician on the nature and the role of statistics. *Advances in Applied Probability* 5–11.

Kahle, D. and Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *The R Journal*, 5, 144–161.

Kain, J. S. and Fritsch, J. M. (1990). A one-dimensional entraining/detraining plume model and its application in convective parameterization. *Journal of the Atmospheric Sciences*, 47, 2784–2802.

Kallache, M., Maksimovich, E., Michelangeli, P.-A. and Naveau, P. (2010). Multimodel combination by a Bayesian hierarchical model: Assessment of ice accumulation over the oceanic Arctic region. *Journal of Climate*, 23, 5421–5436.

Kann, A., Wittmann, C., Wang, Y. and Ma, X. (2009). Calibrating 2-m temperature of limited-area ensemble forecasts using high-resolution analysis. *Monthly Weather Review*, 137, 3373–3387.

Katz, R. W., Parlange, M. B. and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, 25, 1287–1304.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Hoboken.

Kleiber, W., Raftery, A. E., Baars, J., Gneiting, T., Mass, C. F. and Grimit, E. (2011). Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Monthly Weather Review*, 139, 2630–2649.

Knüppel, M. (2015). Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business and Economic Statistics*, 33, 270–281.

Koop, G. M. (2013). Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics*, 28, 177–203.

Krüger, F. (2014). Combining density forecasts under various scoring rules: An analysis of UK inflation. Preprint, available at `https://sites.google.com/site/fk83research/papers`.

Krüger, F., Clark, T. E. and Ravazzolo, F. (2015). Using entropic tilting to combine BVAR forecasts with external nowcasts. *Journal of Business and Economic Statistics*, 33, in press.

Krüger, F., Lerch, S., Thorarinsdottir, T. L. and Gneiting, T. (2016). Probabilistic forecasting and comparative model assessment based on MCMC output. Working paper.

Krüger, F. and Nolte, I. (2015). Disagreement versus uncertainty: Evidence from distribution forecasts. *Journal of Banking & Finance*, 65, in press.

Krzysztofowicz, R. (2001). The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, 249, 2–9.

Kullback, S. (1959). *Information Theory and Statistics*. Dover, New York.

Lehmann, E. L. and Romano, J. B. (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer, New York.

Lei, M., Shiyan, L., Chuanwen, J., Hongling, L. and Yan, Z. (2009). A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews*, 13, 915–920.

Leininger, T. J., Gelfand, A. E., Allen, J. M. and Silander Jr, J. A. (2013). Spatial regression modeling for compositional data with many zeros. *Journal of Agricultural, Biological, and Environmental Statistics*, 18, 314–334.

Leith, C. E. (1974). Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, 102, 409–418.

Lerch, S. (2012). *Verification of probabilistic forecasts for rare and extreme events.* Diplom thesis, Heidelberg University.

Lerch, S. and Baran, S. (2016). Similarity-based semi-local estimation of EMOS models. Accepted for publication at the *Journal of the Royal Statistical Society Series C (Applied Statistics)*. Preprint available at `http://arxiv.org/abs/1509.03521`.

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

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. and Gneiting, T. (2016). Forecaster's dilemma: Extreme events and forecast evaluation. Preprint, available at `http://arxiv.org/abs/1512.09244`.

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

Li, F., Villani, M. and Kohn, R. (2010). Flexible modeling of conditional distributions using smooth mixtures of asymmetric Student t densities. *Journal of Statistical Planning and Inference*, 140, 3638–3654.

Link, W. A. and Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, 3, 112–115.

Little, R. J. (2006). Calibrated Bayes: a Bayes/frequentist roadmap. *The American Statistician*, 60, 213–223.

Lopes, H. F., Salazar, E. and Gamerman, D. (2008). Spatial dynamic factor analysis. *Bayesian Analysis*, 3, 759–792.

Lozano, R., Wang, H., Foreman, K. J., Rajaratnam, J. K., Naghavi, M., Marcus, J. R., Dwyer-Lindgren, L., Lofgren, K. T., Phillips, D., Atkinson, C. et al. (2011). Progress towards Millennium Development Goals 4 and 5 on maternal and child mortality: An updated systematic analysis. *The Lancet*, 378, 1139–1165.

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

Lynch, P. (2008). The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, 227, 3431–3444.

MacEachern, S. N. and Berliner, L. M. (1994). Subsampling the Gibbs sampler. *The American Statistician*, 48, 188–190.

Magnusson, L., Haiden, T. and Richardson, D. (2014). Verification of extreme weather events: Discrete predictands. ECMWF Technical Memorandum 731. Available at `http://www.ecmwf.int/sites/default/files/elibrary/2014/10909-verification-extreme-weather-events-discrete-predictands.pdf`.

Maneesoonthorn, W., Martin, G. M., Forbes, C. S. and Grose, S. D. (2012). Probabilistic forecasts of volatility and its risk premia. *Journal of Econometrics*, 171, 217–236.

Manzan, S. and Zerom, D. (2013). Are macroeconomic variables useful for forecasting the distribution of US inflation? *International Journal of Forecasting*, 29, 469–478.

Marzban, C. (1998). Scalar measures of performance in rare-event situations. *Weather and Forecasting*, 13, 753–763.

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

McNeil, A. J., Frey, R. and Embrechts, P. (2015). *Quantitative Risk Management*. Revised ed. Princeton University Press, Princeton and Oxford.

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

Messner, J. W., Mayr, G. J., Zeileis, A. and Wilks, D. S. (2014b). Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Monthly Weather Review*, 142, 448–456.

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

Möller, A., Thorarinsdottir, T. L., Lenkoski, A. and Gneiting, T. (2015). Spatially adaptive, Bayesian estimation for probabilistic temperature forecasts. Preprint, available at `http://arxiv.org/abs/1507.05066`.

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

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

Nadaraya, E. A. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, 10, 186–190.

National Center for Atmospheric Research (2015). *verification: Weather Forecast Verification Utilities*. R package version 1.42, URL `http://CRAN.R-project.org/package=verification`.

National Weather Service (1998). Automated Surface Observing System (ASOS) User's Guide. Available at: `http://www.weather.gov/asos/aum-toc.pdf`.

Nau, R. F. (1985). Should scoring rules be 'effective'? *Management Science*, 31, 527–535.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transations of the Royal Society Series A*, 231, 289–337.

Noilhan, J. and Planton, S. (1989). A simple parameterization of land surface processes for meteorological models. *Monthly Weather Review*, 117, 536–549.

Ovcharov, E. Y. (2015a). Existence and uniqueness of proper scoring rules. *Journal of Machine Learning Research*, 16, 2207–2230.

Ovcharov, E. Y. (2015b). Proper scoring rules and Bregman divergences. Preprint, available at `http://arxiv.org/abs/1502.01178`.

Owen, J. (1607). *Epigrammatum, Book IV*. Hypertext critical edition by Dana F. Sutton, The University of California, Irvine (1999), available at `http://www.philological.bham.ac.uk/owen/`.

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

Palutikof, J. P., Brabson, B. B., Lister, D. H. and Adcock, S. T. (1999). A review of methods to calculate extreme wind speeds. *Meteorological Applications*, 6, 119–132.

Panagiotelis, A. and Smith, M. (2008). Bayesian density forecasting of intraday electricity prices using multivariate skew t distributions. *International Journal of Forecasting*, 24, 710–727.

Parry, M., Dawid, A. P. and Lauritzen, S. (2012). Proper local scoring rules. *The Annals of Statistics*, 40, 561–592.

Pelenis, J. (2014). Weighted scoring rules for comparison of density forecasts on subsets of interest. Preprint, available at `http://elaine.ihs.ac.at/~pelenis/JPelenis_wsr.pdf`. Accessed January 28, 2016.

Peralta, C., Ben Bouallègue, Z., Theis, S. E., Gebhardt, C. and Buchhold, M. (2012). Accounting for initial condition uncertainties in COSMO-DE-EPS. *Journal of Geophysical Research: Atmospheres*, 117, D07108.

Pinson, P. (2012). Adaptive calibration of (u, v)-wind ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138, 1273–1284.

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

Pinson, P., Chevallier, C. and Kariniotakis, G. N. (2007). Trading wind generation from short-term probabilistic forecasts of wind power. *IEEE Transactions on Power Systems*, 22, 1148–1156.

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

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

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

Pocernich, M. (2012). Verification software. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (I. T. Jolliffe and D. B. Stephenson, eds.), 2nd ed. John Wiley & Sons, 221–240.

Poincaré, H. (2015). *The Foundations of Science*. Cambridge University Press, Cambridge.

Powers, S., Hastie, T. and Tibshirani, R. (2015). Customized training with an application to mass spectrometric imaging of cancer tissue. *The Annals of Applied Statistics*, 9, 1709–1725.

Prescott, P. and Walden, A. T. (1980). Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika*, 67, 723–724.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

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

Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126, 649–668.

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

Riebler, A., Held, L. and Rue, H. (2012). Estimation and extrapolation of time trends in registry data – borrowing strength from related populations. *The Annals of Applied Statistics*, 6, 304–333.

Risser, M. D. and Calder, C. A. (2015). Regression-based covariance functions for nonstationary spatial modeling. *Environmetrics*, 26, 284–297.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. 2nd ed. Springer, New York.

Rodrigues, L. R. L., García-Serrano, J. and Doblas-Reyes, F. (2014). Seasonal forecast quality of the West African monsoon rainfall regimes by multiple forecast systems. *Journal of Geophysical Research: Atmospheres*, 119, 7908–7930.

Romer, C. D. and Romer, D. H. (2000). Federal Reserve information and the behavior of interest rates. *American Economic Review*, 90, 429–457.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27, 832–837.

Rosenblatt, M. (1970). Density estimates and Markov sequences. In *Nonparametric Techniques in Statistical Inference* (M. L. Puri, ed.). Cambridge University Press, 199–211.

Rosenblatt, M. (1971). *Markov Processes. Structure and Asymptotic Behavior*. Springer, Berlin.

Roulin, E. and Vannitsem, S. (2012). Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Monthly Weather Review*, 140, 874–888.

Roulston, M. S. and Smith, L. A. (2003). Combining dynamical and statistical ensembles. *Tellus A*, 55, 16–30.

Roussas, G. G. (1969). Nonparametric estimation in Markov processes. *Annals of the Institute of Statistical Mathematics*, 21, 73–87.

Roussas, G. G. (1988). Nonparametric estimation in mixing sequences of random variables. *Journal of Statistical Planning and Inference*, 18, 135–149.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12, 1151–1172.

Ruiz, J. J. and Saulo, C. (2012). How sensitive are probabilistic precipitation forecasts to the choice of calibration algorithms and the ensemble generation method? Part I: Sensitivity to calibration methods. *Meteorological Applications*, 19, 302–313.

Russell, B. (2001). *The Scientific Outlook*. 3rd ed. Routledge, London and New York.

Sahu, S. K., Bakar, K. S. and Awang, N. (2015). Bayesian forecasting using spatio-temporal models with applications to ozone levels in the eastern United States. In *Geometry Driven Statistics* (I. L. Dryden and J. T. Kent, eds.), chap. 13. John Wiley & Sons, 260–281.

Salazar, E., Sansó, B., Finley, A. O., Hammerling, D., Steinsland, I., Wang, X. and Delamater, P. (2011). Comparing and blending regional climate model predictions for the American Southwest. *Journal of Agricultural, Biological, and Environmental Statistics*, 16, 586–605.

Sass, B. H. (2002). A research version of the STRACO cloud scheme. Danish Meteorological Institute Technical Report 02-10. Available at `http://www.dmi.dk/dmi/index/viden/dmi-publikationer/tekniskerapporter.htm`.

Schefzik, R. (2015). Combining low-dimensional ensemble postprocessing with reordering methods. Preprint, available at `http://arxiv.org/abs/1512.05566`.

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

Schefzik, R., Thorarinsdottir, T. L. and Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28, 616–640.

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140, 1086–1096.

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

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

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

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

Schmeits, M. J. and Kok, K. J. (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.

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

Seguro, J. and Lambert, T. (2000). Modern estimation of the parameters of the Weibull wind speed distribution for wind energy analysis. *Journal of Wind Engineering and Industrial Aerodynamics*, 85, 75–84.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 53, 683–690.

Siegert, S. (2015). *SpecsVerification: Forecast Verification Routines for the SPECS FP7 Project*. R package version 0.4-1, URL `http://CRAN.R-project.org/package=SpecsVerification`.

Sigrist, F., Künsch, H. R. and Stahel, W. A. (2012). A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. *The Annals of Applied Statistics*, 6, 1452–1477.

Sigrist, F., Künsch, H. R. and Stahel, W. A. (2015). Stochastic partial differential equation based modelling of large space–time data sets. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 77, 3–33.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Sköld, M. and Roberts, G. O. (2003). Density estimation for the Metropolis-Hastings algorithm. *Scandinavian Journal of Statistics*, 30, 699–718.

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

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

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

Smith, M. S. and Vahey, S. (2015). Asymmetric density forecasting of U.S. macroeconomic variables. *Journal of Business and Economic Statistics*, 33, in press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 64, 583–639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 76, 485–493.

Stephenson, D. B., Casati, B., Ferro, C. A. T. and Wilson, C. A. (2008). The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteorological Applications*, 15, 41–50.

Strähl, C. and Ziegel, J. F. (2015). Cross-calibration of probabilistic forecasts. Preprint, available at `http://arxiv.org/abs/1505.05314`.

Stummer, W. and Vajda, I. (2012). On Bregman distances and divergences of probability measures. *IEEE Transactions on Information Theory*, 58, 1277–1288.

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

Székely, G. J. and Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, 93, 58–80.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143, 1249–1272.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528–540.

Tay, A. S. and Wallis, K. F. (2000). Density forecasting: A survey. *Journal of Forecasting*, 19, 124–143.

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, Princeton.

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

Thorarinsdottir, T. L., Gneiting, T. and Gissibl, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, 1, 522–534.

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

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

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22, 1701–1728.

Timmermann, A. (2000). Density forecasting in economics and finance. *Journal of Forecasting*, 19, 231–234.

Tödter, J. and Ahrens, B. (2012). Generalization of the ignorance score: Continuous ranked version and its decomposition. *Monthly Weather Review*, 140, 2005–2017.

Toth, Z. and Kalnay, E. (1997). Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review*, 125, 3297–3319.

Tran, L. T. (1989). The L_1 convergence of kernel density estimates under dependence. *The Canadian Journal of Statistics*, 17, 197–208.

Tran, M.-N., Pitt, M. K. and Kohn, R. (2016). Adaptive Metropolis–Hastings sampling using reversible dependent mixture proposals. *Statistics and Computing*, 26, 361–381.

Trombe, P.-J., Pinson, P. and Madsen, H. (2012). A general probabilistic forecasting framework for offshore wind power fluctuations. *Energies*, 5, 621–657.

Tsyplakov, A. (2013). Evaluation of Probabilistic Forecasts: Proper Scoring Rules and Moments. Preprint, available at `http://ssrn.com/abstract=2236605`.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Wang, X. and Bishop, C. H. (2005). Improvement of ensemble reliability with a new dressing kernel. *Quarterly Journal of the Royal Meteorological Society*, 131, 965–986.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer, New York.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, 11, 3571–3594.

Weigel, A. P. (2012). Verification of ensemble forecasts. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (I. T. Jolliffe and D. B. Stephenson, eds.), 2nd ed., chap. 8. John Wiley & Sons, 141–166.

Wied, D. and Weißbach, R. (2012). Consistency of the kernel density estimator: A survey. *Statistical Papers*, 53, 1–21.

Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications*, 16, 361–368.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier Academic Press.

Wilks, D. S. (2015). Multivariate ensemble model output statistics using empirical copulas. *Quarterly Journal of the Royal Meteorological Society*, 141, 945–952.

Wilks, D. S. and Hamill, T. M. (2007). Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, 135, 2379–2390.

Williams, R. M., Ferro, C. A. T. and Kwasniok, F. (2014). A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140, 1112–1120.

Yu, B. (1993). Density estimation in the L^∞ norm for dependent data with applications to the Gibbs sampler. *The Annals of Statistics*, 21, 711–735.

Yuen, R. A., Gneiting, T., Thorarinsdottir, T. L. and Fraley, C. (2013). *ensembleMOS: Ensemble Model Output Statistics*. R package version 0.7, URL `http://CRAN.R-project.org/package=ensembleMOS`.

Zhou, Z., Matteson, D. S., Woodard, D. B., Henderson, S. G. and Micheas, A. C. (2015). A spatio-temporal point process model for ambulance demand. *Journal of the American Statistical Association*, 110, 6–15.