Fakultät für Mathematik und Informatik

Ruprecht-Karls-Universität Heidelberg

Verification of probabilistic forecasts for rare and extreme events

Diplomarbeit

von

Sebastian Lerch

Betreuer: Dr. Thordis Linda Thorarinsdottir

Prof. Dr. Tilmann Gneiting

Juni 2012

Abstract

Accurate predictions of extreme events are of critical importance for avoiding human losses and economic damages. Over the last decades, the conviction that forecasts should be probabilistic in nature has gained ground. To assess forecast quality, theoretically justifiable evaluation procedures for the verification of probabilistic forecasts for extreme events thus have to be developed. Despite the large variety of verification methods for general probabilistic forecasts, there is a notable lack of evaluation procedures tailored to extreme events. In many contexts, particularly in the public and media, forecast evaluation takes place by restricting the standard evaluation procedures to subsets of extreme events. However, we demonstrate that conditioning the observation on being an extreme event leads to the use of improper verification procedures that may discredit even the most skillful forecasters. Recently, two novel approaches to the forecast verification for extreme events have been proposed in the economic literature using weighted scoring rules that emphasize specific regions of interest. We develop a general framework for forecast evaluation and analyze these approaches within this framework using a simulation study and a data example. Furthermore, a new approach to forecast evaluation conditional on extreme ensemble predictions and a simple regime-switching forecasting procedure for wind speed are proposed.

Zusammenfassung

Genaue Vorhersagen von Extremereignissen sind aufgrund deren gravierenden Auswirkungen von außerordentlicher Wichtigkeit. In den letzten Jahrzehnten hat sich in vielen wissenschaftlichen Disziplinen die Überzeugung durchgesetzt, dass Vorhersagen probabilistisch sein sollten. Zur Bewertung der Vorhersagequalität sind somit Verifikationsmethoden für probabilistische Vorhersagen von Extremereignissen notwendig. Trotz der Vielfalt entsprechender Methoden zur allgemeinen Bewertung probabilistischer Vorhersagen kann ein deutlicher Mangel an solchen Verifikationsmethoden für Extremereignisse festgestellt werden. Insbesondere in den Medien findet die Bewertung von Vorhersagesystemen oft ausschließlich beschränkt auf Extremereignisse statt. Es kann jedoch gezeigt werden, dass dieses Bedingen auf Extremereignisse zu ungeeigneten Verifikationsmethoden führt, welche selbst die leistungsfähigsten Vorhersagesysteme diskreditieren. In einem allgemeinen mathematischen Rahmen für die Theorie probabilistischer Vorhersagen werden zwei neue Ansätze zur Verifikation probabilistischer Vorhersagen für Extremereignisse analysiert, welche auf der Anwendung gewichteter Bewertungsregeln beruhen. Des Weiteren führen wird eine neue Methode zur Bewertung probabilistischer Vorhersagen bedingt auf extreme Ensemblevorhersagen eingeführt und ein einfaches Vorhersagemodell für Windgeschwindigkeit entwickelt, welches Ergebnisse aus der Extremwerttheorie und aktuelle Vorhersagemodelle für Windgeschwindigkeit kombiniert.

Contents

1. Introduction

Extreme events have a major influence on mankind and present society with significant challenges. Severe weather and climate events are a great threat and repeatedly cause human losses and economic damages. For the year 2005, the Munich Re Foundation estimated economic losses due to weather-related disasters of about US$ 200 billion, see Dlugolecki (2008) and references therein for details. Over the last decades, an increase in weather and climate extremes can be observed (Diaz and Murnane, 2008), for example in storms (Chang and Fu, 2002), flooding (Milly et al., 2002) or heat waves (Schär et al., 2004). These changes in regional extreme weather may result from the global climate change in the 20th century (Beniston et al., 2007).

There is no unique definition of extreme events in the literature, particularly since the words "severe", "rare" and "extreme" are often used interchangeably. Here, we will follow the suggestions of Stephenson (2008) and define extreme events as extreme values of some (meteorological) variables. Extreme events are generally rare, i.e. sparse on a temporal and spatial scale, and often severe, i.e. resulting in large socio-economic losses. In the applications discussed in this thesis, we focus on extreme weather events. However, the results can be easily extended to other areas such as climatology or economics.

Weather and climate extremes are an inherent part of nature (Diaz and Murnane, 2008) and cannot be avoided. Therefore, accurate predictions of these events are of great importance for minimizing damages and human losses. Over the last two decades, the conviction that forecasts should be probabilistic in nature has gained ground (Dawid, 1984; Gneiting and Raftery, 2007). Probabilistic forecasts in the form of predictive densities or cumulative distribution functions are able to provide information about the forecast uncertainty. This is particularly important for predictions of extreme events which are usually associated with high uncertainty. In order to assess the predictive ability and to rank competing forecasting methods, new evaluation procedures for probabilistic forecasts have been developed. Probabilistic forecasts should be evaluated following the paradigm of maximizing the sharpness of the predictive distribution subject to calibration (Gneiting et al., 2007).

In general, the meteorological and the economic literature provide a variety of theoretically justifiable methods to evaluate probabilistic forecasts (Jolliffe and Stephenson, 2003; Gneiting et al., 2007). However, there is a notable lack of such methods for the evaluation of probabilistic forecasts for extreme events. The thesis at hand addresses this problem.

A natural approach for the evaluation of probabilistic forecasts for extreme events is to select extreme events while discarding non-extreme events, and to proceed with the evaluation using standard evaluation procedures. Intuitively, accurate

predictions of extreme events seem to suggest superior predictive performance. However, it can be shown that this approach is bound to discredit even the most skillful forecasters (Gneiting and Ranjan, 2011b). This critical issue will be referred to as the *forecaster's dilemma*.

Note that this conditioning on extreme events is exactly what happens in the public and in the media, where the attention is often only focused on the performance of forecasters in case of extreme events. Therefore, skillful and calibrated forecasters are bound to fail in the public eye. Table 1.1 presents recent examples from renowned German and international newspapers and broadcasting corporations which demonstrate this focus on extreme events in various fields. Striking examples are the international financial crisis of 2007/8 and the L'Aquila earthquake of 2009. After the financial crisis, lots of attention was paid to those economists who correctly predicted the crisis, and a superior predictive ability was attributed to them. In 2011, against the protest of thousands of scientists around the world, a group of Italian seismologists was put on trial for not warning the public of the L'Aquila earthquake of 2009 which caused 309 deaths. For details, see Hall (2011).

With the observed forecaster's dilemma in mind, new verification procedures for probabilistic forecasts for extreme events have to be developed. Recently, two methods have been proposed in the economic literature using weighted scoring rules that emphasize specific regions of interest, such as extreme events in the right tail of the distributions (Gneiting and Ranjan, 2011b; Diks et al., 2011). The purpose of this thesis is to embed the forecaster's dilemma in a broader general framework for the verification of probabilistic forecasts and to analyze these novel approaches to forecast evaluation for extreme events. We will demonstrate that the conditioning on extreme events leads to improper evaluation procedures and we will compare the approaches to forecast evaluation for extreme events by Gneiting and Ranjan (2011b) and Diks et al. (2011) using a simulation study and a data example.

In many meteorological applications, the evaluation of probabilistic forecasts conditional on extreme ensemble predictions might be of interest. We will demonstrate that unlike conditioning on extreme observations, this does not result in the use of improper verification procedures. Furthermore, we will propose a simple regime-switching approach to probabilistic wind speed forecasting dependent on extreme ensemble predictions which is able to significantly improve the predictions of state-of-the-art ensemble postprocessing techniques.

The remainder of this thesis is organized as follows. Chapter 2 introduces a general framework for the evaluation of probabilistic forecasts and discusses the forecaster's dilemma within this framework. Evaluation procedures for probabilistic forecasts are introduced and extended to verification procedures for probabilistic forecasts of extreme events. The simulation study discussed in Chapter 3 illustrates the theoretical results and is used to closely compare the approaches to forecast verification for extreme events by Gneiting and Ranjan (2011b) and Diks et al. (2011). Chapter 4 analyzes the forecaster's dilemma in a real-world data example of probabilistic wind gust forecasts over the North-American Pacific Northwest. Furthermore, a new approach to forecast evaluation conditional on extreme ensemble predictions and a simple regime-switching forecasting procedure for wind

Table 1.1.: Newspaper articles illustrating the public and media attention focusing on the performance evaluation in case of extreme events. All sources were accessed on May 5, 2012.

Category	Year	Article	Source
Weather	2000	Ratlose Propheten - Ein neues Vorhersagesystem verhinderte die Warnung vor dem Jahrhundertsturm "Lothar".	Der SPIEGEL[1]
Weather	2007	Landratsamt kritisiert Wetterdienst wegen fehlender Warnung	Der SPIEGEL[2]
Weather	2008	Kachelmann: Deutscher Wetterdienst hat versagt	Süddeutsche Zeitung[3]
Weather	2011	Bill Giles accepts blame for BBC Great Storm failures	The Daily Telegraph[4]
Weather	2007	Lessons learned from Great Storm	BBC[5]
Geology	2011	Italian scientists on trial over L'Aquila earthquake	CNN[6]
Geology	2011	Scientists Worry Over "Bizarre" Trial on Earthquake Prediction	Scientific American[7]
Hydrology	2011	Bad Data Failed To Predict Nashville Flood	NBC[8]
Weather, Economics	2011	Un-ending Model Failures in Economics and Weather Forecasting	Asian Tribune[9]
Economics	2009	How Did Economists Get It So Wrong?	The New York Times[10]
Economics	2009	Nouriel Roubini: The economist who predicted worldwide recession	The Guardian[11]
Economics	2009	The Man Nobody Wanted to Hear: Global Banking Economist Warned of Coming Crisis	Der SPIEGEL[12]
Economics	2010	Experts Who Predicted US Economy Crisis See Recovery in 2010	Bloomberg[13]
Economics	2010	An exclusive interview with Med Yones - The expert who predicted the financial crisis	CEOQ Magazine[14]

[1] http://www.spiegel.de/spiegel/print/d-15348803.html
[2] http://www.spiegel.de/panorama/0,1518,495866,00.html
[3] http://www.sueddeutsche.de/panorama/orkan-ueber-deutschland-kachelmann-deutscher-wetterdienst-hat-versagt-1.522051
[4] http://www.telegraph.co.uk/topics/weather/8675175/Bill-Giles-accepts-blame-for-BBC-Great-Storm-failures.html
[5] http://news.bbc.co.uk/2/hi/science/nature/7044050.stm
[6] http://articles.cnn.com/2011-09-20/world/world_europe_italy-quake-trial_1_geophysics-and-vulcanology-l-aquila-seismic-activity?_s=PM:EUROPE
[7] http://www.scientificamerican.com/article.cfm?id=trial-such-as-that-star
[8] http://www.nbc15.com/weather/headlines/January_13_Report_Bad_Data_Failed_To_Predict_Nashville_Flood_113450314.html
[9] http://www.asiantribune.com/news/2011/01/16/un-ending-model-failures-economics-and-weather-forecasting
[10] http://www.nytimes.com/2009/09/06/magazine/06Economic-t.html?_r=1&pagewanted=all
[11] http://www.guardian.co.uk/business/2009/jan/24/nouriel-roubini-credit-crunch
[12] http://www.spiegel.de/international/business/0,1518,635051,00.html
[13] http://www.bloomberg.com/apps/news?pid=conewsstory&refer=conews&tkr=K:US&sid=asziFnEsJSos
[14] http://www.ceoqmagazine.com/whopredictedfinancialcrisis/index.htm

speed are proposed. Connections to existing forecast verification procedures for binary predictions of extreme events are discussed in Chapter 5. We close with a discussion in Chapter 6.

2. Theoretical results

2.1. A mathematical framework for forecasting theory

2.1.1. Ideal forecasts

In order to provide information about their intrinsic uncertainty, forecasts should be probabilistic in nature (Dawid, 1984). Probabilistic forecasts of real-valued quantities can be represented in the form of predictive cumulative distribution functions (CDF) or predictive densities. For complicated models, obtaining the full predictive distributions might be computationally very demanding. However, especially advances in Markov chain Monte Carlo methodology and the advent of ensemble weather prediction systems have led to a growth in the use of predictive distributions in various scientific disciplines (Gneiting and Raftery, 2007; Gneiting, 2008). Based on Gneiting (2010), we develop a mathematical framework for probabilistic forecasting and the evaluation of probabilistic forecasts.

Definition 2.1. We consider a probability space

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

which will be referred to as the *global space*. The observation, a random variable Y which takes values on Ω_Y, is a measurable mapping

$$Y : (\Omega, \mathcal{A}, \mathcal{Q}) \longrightarrow (\Omega_Y, \mathcal{A}_Y)$$

for a σ-algebra $\mathcal{A}_Y$ on Ω_Y. The measurable space $(\Omega_Y, \mathcal{A}_Y)$ is called the *observation space*.

We think of probabilistic forecasts $P_1, \ldots, P_k$ as random quantities. Consider a σ-algebra $\mathcal{A}_\mathcal{P}$ on the class of probability measures $\mathcal{P}$ on the observation space and sub-σ-algebras $\mathcal{A}_1, \ldots, \mathcal{A}_k$ of $\mathcal{A}$. The sub-σ-algebra $\mathcal{A}_i$ is called the *information basis* of forecaster i who issues the probabilistic forecast

$$P_i : (\Omega, \mathcal{A}_i, \mathcal{Q}) \longrightarrow (\mathcal{P}, \mathcal{A}_\mathcal{P}).$$

The information basis $\mathcal{A}_i$ can be interpreted as data, expertise, theories and assumptions at hand (Gneiting and Ranjan, 2011a).

For a real-valued quantity Y, the observation space is given by $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ denotes the Borel-σ-algebra on $\mathbb{R}$. In this case, $\mathcal{P}$ is a subset of the class of Borel measures and a probabilistic forecast P can be identified with the corresponding CDF F.

A probabilistic forecast P_i with CDF F_i is called *ideal* or $\mathcal{Q}$-*ideal* relative to the sub-σ-algebra $\mathcal{A}_i$ if

$$P_i = \mathcal{L}(Y|\mathcal{A}_i) \quad \mathcal{Q} - \text{almost surely.}$$

In practical applications, the global space $(\Omega, \mathcal{A}, \mathcal{Q})$ and the information bases $\mathcal{A}_1, \ldots, \mathcal{A}_k$ will always remain hypothetical. However, in theoretical examples and simulation studies it is possible to define them explicitly. Henceforth, we will only consider real-valued quantities Y and always identify the probabilistic forecast P with the corresponding predictive CDF F or the corresponding predictive density f. A simple example is given below.

Example 2.2. Let

$$\begin{aligned} \Omega &= \mathbb{R} \times \mathbb{R}, \\ \mathcal{A} &= \mathcal{B} \otimes \mathcal{B}, \\ \mathcal{Q} &= \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\right) \end{aligned}$$

be the global space.

The observation Y is given by

$$Y = Y(\omega_1, \omega_2) = \omega_1 + \omega_2$$

and the corresponding observation space

$$\begin{aligned} \Omega_Y &= \mathbb{R}, \\ \mathcal{A}_Y &= \mathcal{B}. \end{aligned}$$

Suppose there exists a forecaster with information basis $\mathcal{A}_1 = \mathcal{B} \otimes \{\emptyset, \mathbb{R}\}$. This forecaster knows the value of ω_1, but does not know the value of ω_2. The ideal prediction relative to $\mathcal{A}_1$ is given by

$$\begin{aligned} F_1 &= \mathcal{L}(Y|\mathcal{A}_1) \\ &= \mathcal{L}(\omega_1 + \omega_2|\mathcal{A}_1) \\ &= \omega_1 + \mathcal{N}(0, 2) \\ &= \mathcal{N}(\omega_1, 2). \end{aligned}$$

Example 2.3. In the situation of Example 2.2, let

$$Y \sim \mathcal{N}(\omega_1 + \omega_2, 1).$$

The information basis of the perfect forecaster who has access to all available information and therefore knows the value of ω_1 and ω_2 is given by

$$\mathcal{A}_p = \mathcal{A} = \mathcal{B} \otimes \mathcal{B}.$$

The perfect forecaster issues the forecast

$$\begin{aligned} F_p &= \mathcal{L}(Y|\mathcal{A}_p) \\ &= \mathcal{L}(Y|\sigma(\omega_1, \omega_2)) \\ &= \mathcal{N}(\omega_1 + \omega_2, 1) \\ &= F_Y, \end{aligned}$$

which is the true distribution of the observation Y. The perfect forecaster is ideal relative to $\mathcal{A}$.

2.1.2. Calibration and Sharpness

The general goal of probabilistic forecasting should be to maximize the sharpness of the predictive distribution subject to calibration (Murphy and Winkler, 1987; Gneiting et al., 2007; Gneiting and Ranjan, 2011a). While "calibration refers to the statistical consistency between the predictive distributions and the associated observations, and is a joint property of the predictions and the values that materialize", "sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only" (Gneiting et al., 2007, page 264). Calibration requires that the observation is indistinguishable from a random draw of the predictive distribution. Sharper probabilistic forecasts correspond to less uncertainty, thus the sharper the better, subject to calibration (Gneiting et al., 2008).

Definition 2.4. Let $V \sim \mathcal{U}([0,1])$ be independent of the prediction F, the information bases $\mathcal{A}_1, \ldots, \mathcal{A}_k$ and the observation Y.

The random variable

$$Z_F = \lim_{y \uparrow Y} F(y) + V(F(Y) - \lim_{y \uparrow Y} F(y))$$

is the *probability integral transform* of the predictive distribution F.

For an almost surely continuous predictive distribution F, it holds that $Z_F(\omega) = [F(\omega)](Y(\omega))$ almost surely. The probability integral transform dates back to Rosenblatt (1952), see also Diebold et al. (1998). Since $F(X) \sim \mathcal{U}([0,1])$ for any real-valued random variable X with cumulative distribution function F, the probability integral transform can be used to define calibration.

Definition 2.5. In the situation of Definition 2.1 with global space $(\Omega, \mathcal{A}, \mathcal{Q})$, a probabilistic forecast F is

(a) *probabilistically calibrated* if its probability integral transform Z_F is uniformly distributed on $[0,1]$ and

(b) *marginally calibrated* if $\mathbb{E}_{\mathcal{Q}}[F(y)] = \mathcal{Q}(Y \leq y)$ for all $y \in \mathbb{R}$.

Theorem 2.6 (Gneiting and Ranjan, 2011a)**.** *A forecast F that is ideal relative to a σ-algebra $\mathcal{A}_0$ is both probabilistically and marginally calibrated.*

Proof. The mathematical framework of Gneiting and Ranjan (2011a) slightly differs from the framework developed in 2.1.1. However, the arguments of the proof of Theorem 2.9 of Gneiting and Ranjan (2011a) still hold in our framework: Suppose that $F = \mathcal{L}(Y|\mathcal{A}_0)$ is ideal relative to the σ-algebra $\mathcal{A}_0$, so that by Definition 2.1 $F(y) = \mathcal{Q}(Y \leq y|\mathcal{A}_0)$ almost surely for all $y \in \mathbb{R}$. Marginal calibration follows from

$$\mathbb{E}_{\mathcal{Q}}[F(y)] = \mathbb{E}_{\mathcal{Q}}[\mathcal{Q}(Y \leq y|\mathcal{A}_0)] = \mathbb{E}_{\mathcal{Q}}\mathbb{E}_{\mathcal{Q}}[\mathbb{1}(Y \leq y)|\mathcal{A}_0] = \mathcal{Q}(Y \leq y),$$

where $\mathbb{1}$ denotes an indicator function.

To prove probabilistic calibration, let $\mathcal{Q}_0$ denote the marginal distribution of Y under $\mathcal{Q}$ so that $Z_F = \mathcal{Q}_0((-\infty, Y)|\mathcal{A}_0) + V\mathcal{Q}_0(\{Y\}|\mathcal{A}_0)$ and

$$\mathcal{Q}(Z_F \leq z) = \mathbb{E}_{\mathcal{Q}}\mathbb{E}_{\mathcal{Q}}[\mathbb{1}(Z_F \leq z)|\mathcal{A}_0]$$

for $z \in (0, 1)$, where the final equality uses the key results of Brockwell (2007). □

Example 2.7. Both F_1 from Example 2.2 and F_p from Example 2.3 are ideal relative to a sub-σ-algebra of $\mathcal{A}$ and therefore probabilistically and marginally calibrated which follows directly from Theorem 2.6.

The sharpness of predictive distributions can be empirically assessed by computing the average width of associated prediction intervals. Gneiting et al. (2007) propose methods for the empirical evaluation of probabilistic and marginal calibration in practical applications. Here, the true data-generating distribution $\mathcal{L}(Y)$ remains hypothetical and therefore has to be replaced by the empirical distribution function. For a set of forecast-observation pairs $(F_t, y_t), t = 1, \ldots, T$, a set of PIT values $p_t = F_t(y_t)$ is obtained.

Tests for the uniformity of the PIT values can be applied in order to test for probabilistic calibration by employing formal tests as proposed by Diebold et al. (1998) or graphical methods such as probability plots or PIT histograms. For small sample sizes and notable deviations from uniformity, this can be done by plotting the empirical CDF of the PIT values against the CDF of a uniform distribution on $[0, 1]$. For large sample sizes and small departures from uniformity, histograms of the PIT values with 10 or 20 bins can be checked for uniformity.

Inspection of the shape of the PIT histograms can furthermore suggest reasons for the deficiencies of the probabilistic forecast at hand. While hump-shaped histograms indicate overdispersion and too wide prediction intervals, U-shaped histograms are caused by underdispersion and prediction intervals that are too narrow. Biased predictive distributions result in triangle-shaped histograms (Gneiting et al., 2007).

Examples of Hamill (2001) and Gneiting et al. (2007) show that there are situations in which evaluation solely based on PIT histograms fails to detect a bias in the single probabilistic forecasts or is unable to distinguish between the ideal forecaster and competitors. Therefore, Gneiting et al. (2007) propose the paradigm of maximizing the sharpness of the predictive distributions subject to calibration.

Marginal calibration can be empirically assessed by comparing the average pre-

dictive CDF,

$$\bar{F}_T(x) = \frac{1}{T}\sum_{t=1}^{T} F_t(x), \quad x \in \mathbb{R},$$

with the empirical CDF of the observations,

$$\hat{G}_T(x) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}(x_t \leq x), \quad x \in \mathbb{R}.$$

Gneiting et al. (2007) propose *marginal calibration diagrams* which are plots of the difference

$$\bar{F}_T(x) - \hat{G}_T(x), \quad x \in \mathbb{R},$$

as a function of the threshold value x. Larger fluctuations around 0 contradict the hypothesis of marginal calibration. To compare predictive distributions whose marginal calibration diagrams exhibit fluctuations around 0 that are of different magnitudes, plots of $\bar{F}_T(x)$ and $\hat{G}_T(x)$ against x can be used instead.

2.1.3. Ideal forecasts and extreme events

Probabilistically and marginally calibrated probabilistic forecasts may loose their desirable properties when extreme events are investigated. Conditioning Y on being an extreme event, for example Y being larger than the 95th percentile q of its marginal distribution, changes the underlying global space and constitutes a mapping

$$(\Omega, \mathcal{A}, \mathcal{Q}) \longrightarrow (\Omega^\star, \mathcal{A}^\star, \mathcal{Q}^\star),$$

where $\mathcal{A}^\star$ is a σ-algebra on $\Omega^\star$ and $\mathcal{Q}^\star$ is a probability measure on $\Omega^\star$. In general, $\Omega^\star \subset \Omega$, but $\Omega^\star$ may also be equal to Ω, while $\mathcal{Q}^\star = \mathcal{Q}|Y > q \neq \mathcal{Q}$, otherwise every outcome would have to be considered an extreme event which is inconsistent with our definition of extreme events. This can be illustrated using the examples from above.

Example 2.8. In the situation of Example 2.2, suppose that $Y^\star = Y|Y > q$, where q is the 95th percentile of the marginal distribution of Y. Then $\Omega^\star = \{(\omega_1, \omega_2) \in \mathbb{R}^2 | \omega_1 + \omega_2 > q\} \neq \Omega = \mathbb{R} \times \mathbb{R}$ and $\mathcal{A}^\star$ is the corresponding σ-algebra on $\Omega^\star$. Here, $\mathcal{Q}^\star$ is a two-dimensional truncated normal distribution given by

$$\mathcal{Q}^\star(a, b) = \mathcal{Q}|_{\mathcal{C}}(a, b),$$

where $\mathcal{C} = \{(a, b) \in \mathbb{R}^2 | a + b > q\}$.

Suppose that a forecaster knows the value ω_1 but not the value ω_2 and furthermore knows that the outcome is an extreme event, then the corresponding information basis becomes

$$\begin{aligned}\mathcal{A}_1^\star &= (\mathcal{B} \otimes \{\emptyset, \mathbb{R}\}) \cap \sigma(\{Y > q\}) \\ &= (\mathcal{B} \otimes \{\emptyset, \mathbb{R}\}) \cap \sigma(\{\omega_1 + \omega_2 > q\}).\end{aligned}$$

The ideal forecast relative to this information basis is

$$\begin{aligned} F_1^\star &= \mathcal{L}(Y|\mathcal{A}_1^\star) \\ &= \omega_1 + \mathcal{N}_{[q-\omega_1,\infty)}(0,2) \\ &= \mathcal{N}_{[q,\infty)}(\omega_1, 2), \end{aligned}$$

where $\mathcal{N}_{[q,\infty)}(a,b)$ denotes a truncated normal distribution with mean a and variance b which is restricted to the interval $[q,\infty)$.

Because there exist sets A such that $A \in \mathcal{A}$ and $A \in \mathcal{A}^\star$ for which $\mathcal{Q}(A) \neq \mathcal{Q}^\star(A)$, F_1 will not be $\mathcal{Q}^\star$-ideal relative to $\mathcal{A}_1$. Since $F_1 \neq F_1^\star$, F_1 will also not be $\mathcal{Q}^\star$-ideal relative to $\mathcal{A}_1^\star$.

A simulation study suggests that F_1 is neither probabilistically nor marginally calibrated in case of extreme events. Figure 2.1 shows PIT histograms and plots of the predictive and empirical cumulative distribution functions for 100 000 random samples of Y. While the PIT histogram of F_1 is uniform, the triangle-shaped PIT histogram of $F_1^\star$ indicates that this probabilistic forecast is biased and not probabilistically calibrated. The comparative plot of the empirical and the predictive cumulative distribution function exhibits no notable deviations from the empirical distribution function of the simulated values for F_1 while there are significant deviations for $F_1^\star$. Therefore, F_1 appears to be probabilistically and marginally calibrated and $F_1^\star$ is neither probabilistically nor marginally calibrated when all events are used for the evaluation.

However, if only observations larger than the 95th percentile of the marginal distribution of Y are considered, F_1 is no longer probabilistically or marginally calibrated which can be seen from Figure 2.2. For these extreme events, the PIT histogram of F_1 basically consists of only one bar while the PIT histogram of $F_1^\star$ is essentially uniform. The comparison of empirical and predictive CDF shows that unlike $F_1^\star$, F_1 is not marginally calibrated.

From Theorem 2.6 it follows that there exists no sub-σ-algebra $\mathcal{A}_0 \subset \mathcal{A}^\star$ such that F_1 is $\mathcal{Q}^\star$-ideal relative to $\mathcal{A}_0$ and the calibrated probabilistic forecast F_1 is no longer probabilistically or marginally calibrated if extreme events are investigated while $F_1^\star$ is probabilistically and marginally calibrated for these events.

Example 2.9. In the situation of Example 2.3, Ω and $\mathcal{A}$ remain unchanged after conditioning on Y being an extreme event since no choice of ω_1, ω_2 would restrict Y to non-extreme events. However, $\mathcal{Q}^\star$ will differ from $\mathcal{Q}$ by assigning larger probabilities to values of ω_1, ω_2 which are more likely to produce extreme outcomes.

If we condition on $Y > q$, the predictive CDF of the perfect forecaster becomes

$$F_p^\star = \mathcal{L}(Y|\sigma(\omega_1,\omega_2) \cap \sigma(\{\omega_1 + \omega_2 > q\})) = \mathcal{N}_{[q,\infty)}(\omega_1 + \omega_2, 1) = F_{Y^\star}.$$

Obviously, $F_p \neq F_p^\star$, and another simulation study suggests that F_p is neither probabilistically nor marginally calibrated for the subset of extreme events. For 100 000 random samples from Y, PIT histograms and plots of the predictive and empirical cumulative distribution functions for the subset of observations larger than the 95th percentile of the marginal distribution of Y are displayed in Fig-

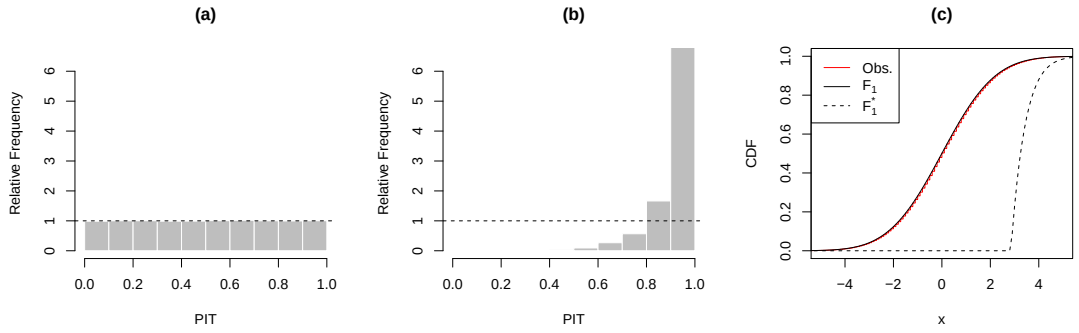

Figure 2.1.: PIT histograms for (a) F_1, (b) $F_1^\star$ and (c) plots of the empirical (red) and predictive (black) cumulative distribution functions for random samples of Y.

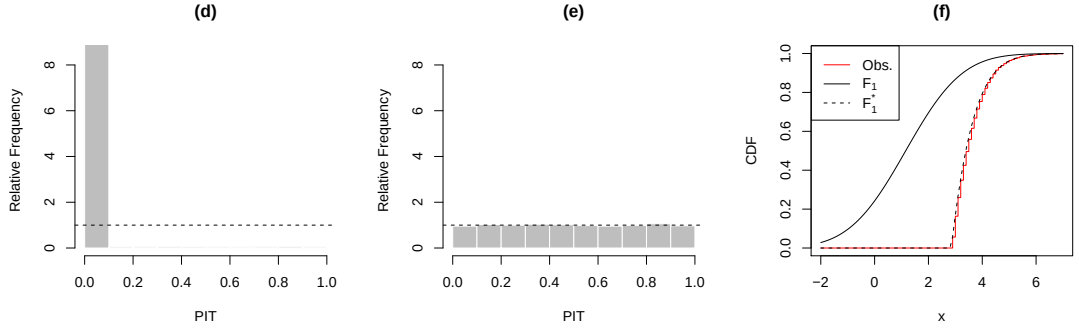

Figure 2.2.: PIT histograms for (d) F_1, (e) $F_1^\star$ and (f) plots of the empirical (red) and predictive (black) cumulative distribution functions for random samples of Y larger than the 95th percentile of the marginal distribution of Y.

ure 2.3. The triangle-shaped PIT histogram of F_p suggests that this predictive distribution is biased while $F_p^\star$ seems to be probabilistically calibrated.

Furthermore, a comparison of the predictive and the empirical cumulative distribution functions shows that unlike $F_p^\star$, F_p is also not marginally calibrated for the subset of extreme events.

Therefore, even the perfect forecaster who has all available information except for the knowledge of Y being an extreme event is neither probabilistically nor marginally calibrated any more if only extreme events are investigated.

In the latter situation where Ω and $\mathcal{A}$ remain unchanged after conditioning on Y being an extreme event, it is possible to show that there is no forecast which is both $\mathcal{Q}$-ideal and $\mathcal{Q}^\star$-ideal relative to a sub-σ-algebra $\mathcal{A}_0 \subset \mathcal{A} = \mathcal{A}^\star$.

Corollary 2.10. *If there exists a countable family of sets $\mathcal{E}$ generating the σ-algebra $\mathcal{A}_Y$ and there exists a sub-σ-algebra $\mathcal{A}_0$ with $\mathcal{A}_0 \subset \mathcal{A}$ and $\mathcal{A}_0 \subset \mathcal{A}^\star$ such that F is both $\mathcal{Q}_1$-ideal and $\mathcal{Q}_2$-ideal relative to this sub-σ-algebra for two measures*

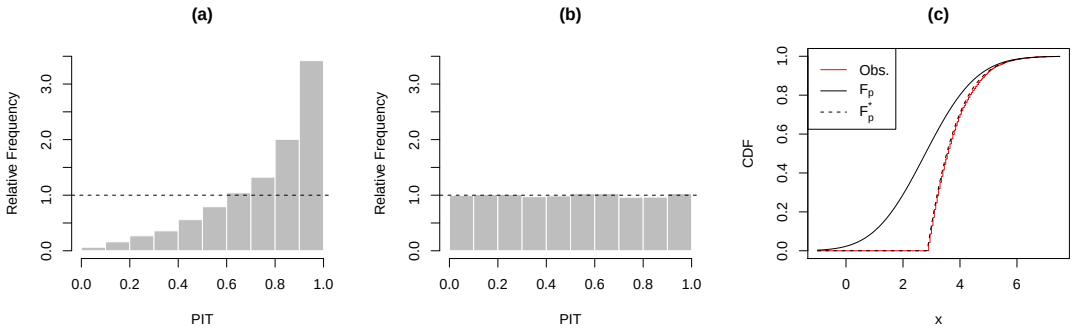

Figure 2.3.: PIT histograms for (a) F_p, (b) $F_p^\star$ and (c) plots of the empirical (red) and predictive (black) cumulative distribution functions for random samples of Y larger than the 95th percentile of the marginal distribution of Y.

$\mathcal{Q}_1$ and $\mathcal{Q}_2$, i.e.

$$F = \mathcal{L}(Y|\mathcal{A}_0) \quad \mathcal{Q}_1\text{-a.s.} \quad \text{and}$$
$$F = \mathcal{L}(Y|\mathcal{A}_0) \quad \mathcal{Q}_2\text{-a.s.},$$

then the two measures $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are identical on $\mathcal{A}_0$ almost surely.

Proof.

$$\begin{aligned} F = \mathcal{L}(Y|\mathcal{A}_0) \quad \mathcal{Q}_1\text{-a.s. and } F = \mathcal{L}(Y|\mathcal{A}_0) \quad \mathcal{Q}_2\text{-a.s.} \\ \Rightarrow \quad \mathcal{Q}_1(Y \leq y|\mathcal{A}_0) = \mathcal{Q}_2(Y \leq y|\mathcal{A}_0) \quad \forall y \in \mathbb{R} \\ \Rightarrow \quad \mathbb{E}_{\mathcal{Q}_1}(\mathbb{1}(Y \leq y)|\mathcal{A}_0) = \mathbb{E}_{\mathcal{Q}_2}(\mathbb{1}(Y \leq y)|\mathcal{A}_0). \end{aligned}$$

Theorem 44.3 of Bauer (2002) states that if $\mathcal{C}$ is a sub-σ-algebra of $\mathcal{A}$, if $\mathcal{L}(X|\mathcal{C})$ and $\mathcal{L}^\star(X|\mathcal{C})$ are two conditional distributions of a random variable $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$, and if there exists a countable family of sets $\mathcal{E}'$ generating the σ-algebra $\mathcal{A}'$, then there exists a set $N \in \mathcal{A}$ of measure 0 such that

$$\mathcal{L}(X|\mathcal{C})(\omega, A') = \mathcal{L}^\star(X|\mathcal{C})(\omega, A') \quad \forall \omega \in \bar{N}, A' \in \mathcal{A}',$$

where $\bar{N}$ is the complement of N in Ω.

Since $\mathcal{Q}_1(Y \leq y|\mathcal{A}_0) = \mathcal{Q}_2(Y \leq y|\mathcal{A}_0)$ holds for all $y \in \mathbb{R}$ which uniquely determines the distribution of $Y|\mathcal{A}_0$, it follows that $\mathcal{Q}_1 = \mathcal{Q}_2$ on $\mathcal{A}_0$ almost surely. □

If we consider two measures $\mathcal{Q}_1$ and $\mathcal{Q}_2$ which differ on a sub-σ-algebra $\mathcal{A}_0$ (for example by imposing the constraint on Y that the outcome is an extreme event), Corollary 2.10 implies that there is no forecast which is $\mathcal{Q}_1$-ideal and $\mathcal{Q}_2$-ideal relative to $\mathcal{A}_0$. Thus, any forecast which is $\mathcal{Q}_1$-ideal relative to $\mathcal{A}_0$ will not be $\mathcal{Q}_2$-ideal relative to $\mathcal{A}_0$ and may therefore no longer be probabilistically or marginally calibrated.

If F is $\mathcal{Q}_1$-ideal relative to $\mathcal{A}_0$ and not $\mathcal{Q}_2$-ideal relative to $\mathcal{A}_0$ there might of course still exist a different sub-σ-algebra $\mathcal{A}_1 \neq \mathcal{A}_0$ such that F is $\mathcal{Q}_2$-ideal relative to this σ-algebra and would therefore still be marginally and probabilistically calibrated. However, Example 2.9 showed that this may not even be the case for the perfect forecaster who has access to all available information except for knowing that the outcome of Y is an extreme event.

2.2. Forecast evaluation

Following Gneiting and Raftery (2007), we introduce proper scoring rules for the evaluation of probabilistic forecasts within the mathematical framework developed in 2.1.1. The evaluation of probabilistic forecasts follows the paradigm of maximizing the sharpness subject to calibration. Scoring rules assign a numerical score to pairs of forecasts and observations and provide summary measures of predictive performance by simultaneously addressing calibration and sharpness (Gneiting et al., 2007). Scoring rules can be either positively or negatively oriented. For the purpose of developing characterizations, we take scoring rules to be positively oriented functions taking values on the extended real line $\bar{\mathbb{R}} = [-\infty, \infty]$. If a forecaster states the prediction P and the outcome y is observed, the value of the positively oriented scoring rule $S(P, y)$ can be thought of as a reward the forecaster wishes to maximize.

2.2.1. Proper scoring rules

Based on the framework developed in Definition 2.1, consider an observation space $(\Omega_Y, \mathcal{A}_Y)$ and a convex class $\mathcal{P}$ of probability measures on $(\Omega_Y, \mathcal{A}_Y)$.

Definition 2.11. A function defined on Ω_Y which takes values on $\bar{\mathbb{R}} = [-\infty, \infty]$ is called *$\mathcal{P}$-quasi-integrable* if it is measurable with respect to $\mathcal{A}_Y$ and quasi-integrable with respect to every $P \in \mathcal{P}$, i.e. if there exists a real-valued integral for its positive or its negative part with respect to P (Bauer, 1992).

Definition 2.12. (a) A *scoring rule* is a function $S : \mathcal{P} \times \Omega_Y \longrightarrow \bar{\mathbb{R}}$ such that $S(P, \cdot)$ is $\mathcal{P}$-quasi-integrable for any $P \in \mathcal{P}$.

(b) The *expected score* of the probabilistic forecast P under the true distribution Q of Y is given by $\boldsymbol{S}(P, Q) = \int_{\Omega_Y} S(P, \omega)\, dQ(\omega)$.

(c) A positively oriented scoring rule is called *proper* relative to $\mathcal{P}$ if

$$\boldsymbol{S}(Q, Q) \geq \boldsymbol{S}(P, Q) \; \forall P, Q \in \mathcal{P}$$

and *strictly proper* relative to $\mathcal{P}$ if equality holds if and only if $P = Q$.

(d) A scoring rule is *regular* relative to the class $\mathcal{P}$ if $\boldsymbol{S}(P, P) \in \mathbb{R}$ for all $P \in \mathcal{P}$ and $\boldsymbol{S}(P, Q) \in [-\infty, \infty)$ for all $P, Q \in \mathcal{P}$.

Propriety of scoring rules is a desirable property as it encourages the forecaster to quote his or her true belief and they are designed not to provide incentive to digress from one's true belief (Gneiting and Raftery, 2007; Gneiting et al., 2008). Calibration, sharpness and the propriety of scoring rules correspond to the three types of goodness of performance for general forecasting systems identified by Murphy (1993). For details, see Friederichs and Thorarinsdottir (2012).

Proper scoring rules can be related to the theory of convex functions as follows.

Definition 2.13. A function $G : \mathcal{P} \longrightarrow \bar{\mathbb{R}}$ is *convex* if

$$G(\lambda P_0 + (1-\lambda)P_1) \leq \lambda G(P_0) + (1-\lambda)G(P_1)$$

for all $\lambda \in [0,1], P_0, P_1 \in \mathcal{P}$. G is furthermore *strictly convex* if equality holds if and only if $P_0 = P_1$.

A function $G^\star(P,\cdot) : \Omega_Y \longrightarrow \bar{\mathbb{R}}$ is a *subtangent* of G in $P \in \mathcal{P}$ if $G^\star$ is P-integrable and $\mathcal{P}$-quasi-integrable and

$$G(Q) \geq G(P) + \int_{\Omega_Y} G^\star(P,\omega)\,\mathrm{d}(Q-P)(\omega) \tag{2.1}$$

for all $Q \in \mathcal{P}$.

Theorem 2.14 (Gneiting and Raftery, 2007)**.** *A regular scoring rule $S : \mathcal{P} \times \Omega_Y \longrightarrow \bar{\mathbb{R}}$ is proper if and only if there exists a convex, real-valued function $G : \mathcal{P} \longrightarrow \mathbb{R}$ such that*

$$S(P,\omega) = G(P) - \int_{\Omega_Y} G^\star(P,\omega)\,\mathrm{d}P(\omega) + G^\star(P,\omega) \tag{2.2}$$

for $P \in \mathcal{P}$ and $\omega \in \Omega_Y$, where $G^\star(P,\cdot) : \Omega_Y \longrightarrow \bar{\mathbb{R}}$ is a subtangent of G in $P \in \mathcal{P}$. The statement still holds if proper is replaced by strictly proper and convex is replaced by strictly convex.

Proof. Theorem 1 of Gneiting and Raftery (2007, page 361). □

Definition 2.15. Suppose that S is a proper scoring rule relative to $\mathcal{P}$.

(a) The function $G : \mathcal{P} \longrightarrow \mathbb{R}$,

$$G(P) = \boldsymbol{S}(P,P) = \sup_{Q \in \mathcal{P}} \boldsymbol{S}(Q,P)$$

is the *information measure* or *generalized entropy function* associated with S.

(b) If S is regular and proper,

$$d : \mathcal{P} \times \mathcal{P} \longrightarrow \bar{\mathbb{R}},$$
$$d(P,Q) = \boldsymbol{S}(Q,Q) - \boldsymbol{S}(P,Q)$$

is the *divergence function* associated with S.

Note that the order of the arguments in the definition of $d(P, Q)$ differs from the previous practice. For $P, Q \in \mathcal{P}$, the divergence function $d(P, Q)$ is nonnegative and if S is strictly proper, $d(P, Q)$ is strictly positive except for $P = Q$. Using Definition 2.15, the theory of proper scoring rules can be related to the theory of *Bregman distances* (Bregman, 1967), which is widely used in convex optimization and machine learning (Boyd and Vandenberghe, 2004; Gneiting and Raftery, 2007).

Proper scoring rules also occur in the context of decision theoretical problems (Dawid, 1998). Let the scoring rule S be defined as

$$S(P, \omega) = U(\omega, a_P),$$

where ω is the observation, a_P is the Bayes act for $P \in \mathcal{P}$ and $U(\omega, a)$ is the utility for outcome ω and action a. Then S is proper relative to $\mathcal{P}$. This follows from

$$\begin{aligned} S(Q, Q) &= \int U(\omega, a_Q) \mathrm{d}Q(\omega) \\ &\geq \int U(\omega, a_P) \mathrm{d}Q(\omega) \\ &= S(P, Q) \end{aligned}$$

since the optimal Bayesian decision maximizes expected utility (Gneiting and Raftery, 2007).

We have seen that proper scoring rules can be characterized using the theory of convex functions and can be related to various other mathematical fields. In the following two sections, we will discuss combinations of proper scoring rules and examples of proper scoring rules for probabilistic forecasts of continuous real-valued variables.

2.2.2. Combining proper scoring rules

We explore combinations of proper scoring rules of the form

$$S(P, \omega) = h(S_1(P, \omega), S_2(P, \omega)),$$

where $S_i(P, \omega)$, $i = 1, 2$ is a proper scoring rule and $h : [-\infty, \infty) \times [-\infty, \infty) \longrightarrow [-\infty, \infty)$. In particular, we are interested in conditions on h such that S is a proper scoring rule.

Lemma 2.16 (Vector composition of convex functions)**.** *Consider* $g_i : \mathcal{P} \longrightarrow \mathbb{R}$, $i = 1, 2$ *and* $h : [-\infty, \infty) \times [-\infty, \infty) \longrightarrow [-\infty, \infty)$*. If*

(i) g_i *is convex for* $i = 1, 2$*,*
(ii) h *is convex and*
(iii) h *is nondecreasing in each argument,*

then

$$f(P) = h(g_1(P), g_2(P))$$

is convex.

Proof.

$$\begin{aligned}
f(\lambda P_0 + (1-\lambda)P_1) &= h(g_1(\lambda P_0 + (1-\lambda)P_1), g_2(\lambda P_0 + (1-\lambda)P_1)) \\
&\leq h(\lambda g_1(P_0) + (1-\lambda)g_1(P_1), \lambda g_2(P_0) + (1-\lambda)g_2(P_1)) \\
&\qquad \text{since } g_i \text{ is convex and h is nondecreasing} \\
&\overset{(a)}{=} h(\lambda[g_1(P_0), g_2(P_0)] + (1-\lambda)[g_1(P_2), g_2(P_2)]) \\
&\leq \lambda h(g_1(P_0), g_2(P_0)) + (1-\lambda)h(g_1(P_1), g_2(P_1)) \\
&\qquad \text{since } h \text{ is convex} \\
&= \lambda f(P_0) + (1-\lambda)f(P_1),
\end{aligned}$$

where (a) follows since

$$\begin{aligned}
(\lambda g_1(P_0) + (1-\lambda)g_1(P_1), \lambda g_2(P_0) + (1-\lambda)g_2(P_1)) = \\
(\lambda(g_1(P_0), g_2(P_0)) + (1-\lambda)(g_1(P_1), g_2(P_1))).
\end{aligned}$$

□

Lemma 2.17 (Ekeland and Temam (1976))**.** *If $(F_i)_{i\in I}$ is a family of convex functions of a vector space over $\mathbb{R}$ into $\bar{\mathbb{R}}$, their pointwise supremum $F = \sup_{i\in I} F_i$ is convex.*

Proof. Proposition 2.2 of Ekeland and Temam (1976, page 9). □

In the statement of Lemma 2.17, convexity of the functions $F_i, i \in I$ is a necessary, but not sufficient condition: Suppose

$$\begin{aligned}
f_1(x) &= \begin{cases} \cos(x) - 1, & x < 0 \\ x^2, & x \geq 0, \end{cases} \\
f_2(x) &= \begin{cases} x^2, & x \leq 0 \\ \cos(x) - 1, & x > 0. \end{cases}
\end{aligned}$$

Then the pointwise supremum, which is equal to the pointwise maximum

$$g(x) = \sup_{i=1,2} f_i(x) = \max_{i=1,2} f_i(x) = x^2$$

is convex although neither f_1 nor f_2 is convex.

The following lemma will allow for formulating necessary conditions to be imposed on h in order to retrieve a proper scoring rule by combining proper scoring rules.

Lemma 2.18. *$G(P) = \boldsymbol{S}(P, P)$ is convex in P if and only if S is a proper scoring rule.*

Proof. $G(P)$ is convex if and only if for all $P_1, P_2 \in \mathcal{P}$ and $\lambda \in [0,1]$ it holds that

$$
\begin{aligned}
& \lambda G(P_1) + (1-\lambda)G(P_2) - G(\lambda P_1 + (1-\lambda)P_2) \geq 0 \\
\Leftrightarrow\ & \lambda \boldsymbol{S}(P_1, P_1) + (1-\lambda)\boldsymbol{S}(P_2, P_2) \\
& \quad - \boldsymbol{S}(\lambda P_1 + (1-\lambda)P_2, \lambda P_1 + (1-\lambda)P_2) \geq 0 \ \forall P_1, P_2 \in \mathcal{P}, \lambda \in [0,1] \\
\overset{(a)}{\Leftrightarrow}\ & \lambda \boldsymbol{S}(P_1, P_1) + (1-\lambda)\boldsymbol{S}(P_2, P_2) - \big[\lambda \boldsymbol{S}(\lambda P_1 + (1-\lambda)P_2, P_1) \\
& \quad + (1-\lambda)\boldsymbol{S}(\lambda P_1 + (1-\lambda)P_2, P_2)\big] \geq 0 \ \forall P_1, P_2 \in \mathcal{P}, \lambda \in [0,1] \\
\Leftrightarrow\ & \lambda\big[\boldsymbol{S}(P_1, P_1) - \boldsymbol{S}(\lambda P_1 + (1-\lambda)P_2, P_1)\big] \\
& \quad + (1-\lambda)\big[\boldsymbol{S}(P_2, P_2) - \boldsymbol{S}(\lambda P_1 + (1-\lambda)P_2, P_2)\big] \geq 0 \ \forall P_1, P_2 \in \mathcal{P}, \lambda \in [0,1] \\
\overset{(b)}{\Leftrightarrow}\ & \boldsymbol{S}(P_1, P_1) - \boldsymbol{S}(\lambda P_1 + (1-\lambda)P_2, P_1) \geq 0 \\
& \quad \text{and } \boldsymbol{S}(P_2, P_2) - \boldsymbol{S}(\lambda P_1 + (1-\lambda)P_2, P_2) \geq 0 \ \forall P_1, P_2 \in \mathcal{P}, \lambda \in [0,1] \\
\Leftrightarrow\ & \boldsymbol{S}(P_1, P_1) \geq \boldsymbol{S}(\lambda P_1 + (1-\lambda)P_2, P_1) \\
& \quad \text{and } \boldsymbol{S}(P_2, P_2) \geq \boldsymbol{S}(\lambda P_1 + (1-\lambda)P_2, P_2) \ \forall P_1, P_2 \in \mathcal{P}, \lambda \in [0,1] \\
\overset{(c)}{\Leftrightarrow}\ & \boldsymbol{S}(P, P) \geq \boldsymbol{S}(Q, P) \ \forall P, Q \in \mathcal{P}, \lambda \in [0,1]
\end{aligned}
$$

where (a) follows since $\boldsymbol{S}$ is linear in the second component:

$$
\begin{aligned}
\boldsymbol{S}_i(Q, \lambda P_1 + (1-\lambda)P_2) &= \int_\Omega S_i(Q, \omega)\, \mathrm{d}\left[\lambda P_1 + (1-\lambda)P_2\right](\omega) \\
&= \int_\Omega S_i(Q, \omega)\left[\mathrm{d}\lambda P_1(\omega) + \mathrm{d}(1-\lambda)P_2(\omega)\right] \\
&= \lambda \int_\Omega S_i(Q, \omega)\, \mathrm{d}P_1(\omega) + (1-\lambda)\int_\Omega S_i(Q, \omega)\, \mathrm{d}P_2(\omega) \\
&= \lambda \boldsymbol{S}_i(Q, P_1) + (1-\lambda)\boldsymbol{S}_i(Q, P_2)
\end{aligned}
$$

and (b) follows since the sum in the equation above can only be nonnegative if both summands are nonnegative. This follows from the fact that the statement holds for any $P_1, P_2 \in \mathcal{P}$ and $\lambda \in (0,1)$ and furthermore implies that if one summand is smaller than 0, the other one is as well and the equation above cannot hold.

The last statement (c) is equivalent to S being a proper scoring rule and follows since the equation above holds for all $Q = \lambda P_1 + (1-\lambda)P_2$ with $P_1, P_2 \in \mathcal{P}$. Allowing λ to be 0 and 1 as well does not change the statements above, as for any $Q \in \mathcal{P}$ one might just set $\lambda = 1$ and $P_1 = Q$. □

Dawid (1998) discusses more general versions of Lemma 2.18. S being a proper scoring rule always implies convexity of G, but the converse does not necessarily hold. Dawid (1998) refers to Example 4.1 of Hendrickson and Buehler (1971) as a counter example but also mentions that the converse will hold "in suitable spaces under suitable continuity conditions" (Dawid, 1998, page 28). For a discussion of these conditions he refers to Johnson (1991). The proof of Lemma 2.18 shows that these conditions are met here.

If $G(P)$ is convex, then S is a proper scoring rule and $G(P) = S(P, P) =$

$\sup_{Q\in\mathcal{P}} S(Q,P)$. This can be used to find necessary conditions on h such that

$$S(P,\omega) = h(S_1(P,\omega), S_2(P,\omega)),$$

is a proper scoring rule.

Theorem 2.19. *If*

(i) S_i *is a proper scoring rule for* $i = 1, 2$,
(ii) $h : [-\infty,\infty) \times [-\infty,\infty) \longrightarrow [-\infty,\infty)$ *is convex and*
(iii) h *is non-decreasing in each component,*

then

$$S(P,\omega) = h(S_1(P,\omega), S_2(P,\omega)),$$

is a proper scoring rule.

Proof. In order to use the statement of Lemma 2.18, we show that $G(P) = \boldsymbol{S}(P,P) = h(\boldsymbol{S}_1(P,P), \boldsymbol{S}_2(P,P))$ is convex. This follows directly from Lemma 2.16: Condition (i) implies that $\boldsymbol{S}_i(Q,P)$ for $i = 1,2$ is convex in P for all $Q \in \mathcal{P}$ which was shown in the proof of Lemma 2.18 or can be obtained directly from Lemma 2.17. Together with (ii), (iii), Lemma 2.16, we can conclude that $G(P)$ is convex in P and therefore the propriety of S follows from Lemma 2.18. □

Example 2.20. Possible choices of h include

- linear functions $S(P,\omega) = h(S_1(P,\omega), S_2(P,\omega)) = aS_1(P,\omega) + bS_2(P,\omega)$ with $a, b > 0$,
- $S(P,\omega) = h(S_1(P,\omega), S_2(P,\omega)) = \left[S_1(P,\omega)^q + S_2(P,\omega)^q)\right]^{\frac{1}{q}}$ for $q \geq 1$,
- any norm $\|\cdot\|$ on $\mathbb{R}^2$ such that $S(P,\omega) = \|(S_1(P,\omega), S_2(P,\omega))^T\|$,
- $S(P,\omega) = h(S_1(P,\omega), S_2(P,\omega)) = \left[\max\{S_1(P,\omega), 0\}^q + \max\{S_2(P,\omega), 0\}^q\right]^{\frac{1}{q}}$ for $q \geq 1$.

To obtain propriety of $S(P,\omega) = h(S_1(P,\omega), S_2(P,\omega))$, h must be non-decreasing in each component. Otherwise there may exist $Q, P \in \mathcal{P}$ with $Q \neq P$ such that

$$\boldsymbol{S}(Q,P) = h(\boldsymbol{S}_1(Q,P), \boldsymbol{S}_2(Q,P)) > h(\boldsymbol{S}_1(P,P), \boldsymbol{S}_2(P,P)) = \boldsymbol{S}(P,P).$$

Not all choices of S_1 and S_2 yield meaningful proper scoring rules. Dawid (2007) defines equivalence of scoring rules.

Definition 2.21 (Dawid, 2007)**.** Let $c \in \mathbb{R}, c > 0$. A scoring rule S_1 is *equivalent* to another scoring rule S_2 if

$$S_1(P,\omega) = cS_2(P,\omega) + k(\omega)$$

for all $P \in \mathcal{P}, \omega \in \Omega_Y$ and a real-valued function k on Ω_Y.

Corollary 2.22. *Combinations of equivalent proper scoring rules do not provide any additional information compared to using only one of the equivalent proper scoring rules.*

Proof. If S_1 is equivalent to S_2, it holds that

$$\boldsymbol{S}_1(Q, P) = c\boldsymbol{S}_2(Q, P) + k(P),$$

where $k(P) = \int_{\Omega_Y} k(\omega) dP(\omega)$. Thus, $S_2(Q_1, P) \leq S_2(Q_2, P)$ implies that

$$\boldsymbol{S}_1(Q_1, P) = c\boldsymbol{S}_2(Q_1, P) + k(P) \leq c\boldsymbol{S}_2(Q_2, P) + k(P) = \boldsymbol{S}_1(Q_2, P)$$

for all $Q_1, Q_2, P \in \mathcal{P}$.

Therefore, S_1 does not provide any additional information compared to S_2 if for example the predictive performance of two competing predictive distributions is compared, and neither will any combination of S_1 and S_2. □

We will now introduce examples of proper scoring rules for probabilistic forecasts of continuous real-valued variables such as temperature or wind speed.

2.2.3. Proper scoring rules for continuous real-valued variables

In order to use results from the theory of convex functions, we have only considered positively oriented proper scoring rules so far. From now on, we will take scoring rules to be negatively oriented, following the common practice of forecast evaluation (Gneiting and Raftery, 2007). An negatively oriented scoring rule $S_0^{(n)}$ can easily be obtained from any positively oriented scoring rule $S_0^{(p)}$ by setting $S_0^{(n)} = -S_0^{(p)}$. If the forecaster states the prediction P and y is observed, $S_0^{(n)}(P, y)$ can be thought of as a penalty the forecaster wishes to minimize.

Analogous to the definition of propriety for positively oriented scoring rules we define:

Definition 2.23. A negatively oriented scoring rule is called *proper* relative to $\mathcal{P}$ if

$$\boldsymbol{S}(Q, Q) \leq \boldsymbol{S}(P, Q) \ \forall P, Q \in \mathcal{P}$$

and *strictly proper* relative to $\mathcal{P}$ if equality holds if and only if $P = Q$.

Gneiting and Raftery (2007) introduce several examples of scoring rules for forecasts of continuous variables.

Definition 2.24. For a σ-finite measure μ on $(\Omega_Y, \mathcal{A}_Y)$ and $\alpha > 1$, let $\mathcal{L}_\alpha$ denote the class of probability measures on $(\Omega_Y, \mathcal{A}_Y)$ that are absolutely continuous with respect to μ and have μ-density f such that

$$\|f\|_\alpha = \left(\int f(\omega)^\alpha \mu(\mathrm{d}\omega) \right)^{1/\alpha}$$

exists.

A probabilistic forecast $P \in \mathcal{L}_\alpha$ can be identified with its μ-density f which is a *predictive density* or *density forecast*.

Example 2.25. (a) The *quadratic score*,

$$\mathrm{QuadrS}(f, \omega) = \|f\|_2^2 - 2f(\omega),$$

is strictly proper relative to $\mathcal{L}_2$.

(b) The *pseudospherical score*,

$$\mathrm{PseudoS}(f, \omega) = -\frac{f(\omega)^{\alpha-1}}{\|f\|_\alpha^{\alpha-1}},$$

is strictly proper relative to $\mathcal{L}_\alpha$ and reduces to the spherical score for $\alpha = 2$.

(c) The *logarithmic score* or *ignorance score*,

$$\mathrm{LogS}(f, \omega) = -\log f(\omega),$$

is strictly proper relative to $\mathcal{L}_1$.

(d) The *linear score*,

$$\mathrm{LinS}(f, \omega) = -f(\omega),$$

is not a proper scoring rule (Gneiting and Raftery, 2007).

Gneiting and Raftery (2007) argue that in many applications, the restriction to density forecasts is impractical and scoring rules should be defined directly in terms of predictive cumulative distribution functions instead. They propose the continuous ranked probability score, which is defined as the integral over the Brier score (Brier, 1950) for the associated binary probability forecasts at all real-valued thresholds (Matheson and Winkler, 1976; Hersbach, 2000). Here, we consider $(\Omega_Y, \mathcal{A}_Y) = (\mathbb{R}, \mathcal{B})$ and identify a probabilistic forecast with the corresponding cumulative distribution function F.

Definition 2.26. The *continuous ranked probability score* (CRPS) is defined as

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

For some distributions, closed form expressions for the integral can be obtained. For normal distributions with mean μ and variance σ^2, the CRPS is given by

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

where φ and Φ denote the probability density function and cumulative distribution function of a standard Gaussian variable respectively (Gneiting and Raftery, 2007).

Results of Baringhaus and Franz (2004) show that

$$\mathrm{CRPS}(F, y) = \mathbb{E}_F \left| Y - y \right| - \frac{1}{2} \mathbb{E}_F \left| Y - Y' \right|,$$

where Y and Y' are two independent random variables with cumulative distribution function F and finite first moment (Gneiting and Raftery, 2007).

The CRPS is proper relative to the class of Borel probability measures on $\mathbb{R}$ and strictly proper relative to the subclass of Borel probability measures with finite first moment. It can be reported in the same unit as the observation and can be used to directly compare deterministic and probabilistic forecasts since it reduces to the absolute error if F is a deterministic forecast which corresponds to F being a point measure (Gneiting and Raftery, 2007).

2.3. Evaluation of probabilistic forecasts for extreme events

We have seen that there is a large variety of theoretically justifiable methods for the evaluation of probabilistic forecasts, for example by following the paradigm of maximizing the sharpness of the predictive distribution subject to calibration. However, this is not the case for the evaluation in case of extreme events. The natural approach would be to select extreme outcomes while discarding non-extreme outcomes, and to apply proper scoring rules restricted to this subset of extreme events. The following will show that the conditioning on extreme events has unexpected and unwanted effects and that restricting proper scoring rules to subsets of extreme events yields improper scoring rules. However, proper scoring rules for extreme events can be defined using weighted scoring rules to emphasize regions of interest. Two methods have recently been proposed in the economic literature (Gneiting and Ranjan, 2011b; Diks et al., 2011).

2.3.1. Restricting proper scoring rules to extreme events

Since we take scoring rules to be negatively oriented, a scoring rule $S(f, y)$ for a density forecast is proper if

$$\begin{aligned} \boldsymbol{S}(f, Y) = \mathbb{E}_f S(f, Y) &= \int f(y) S(f, y) \mathrm{d}y \\ &\leq \int f(y) S(g, y) \mathrm{d}y \\ &= \mathbb{E}_f S(g, Y) = \boldsymbol{S}(g, Y) \end{aligned}$$

for all probability density functions f and g. S is strictly proper if equality only holds for $f = g$.

Gneiting and Ranjan (2011b) show that the product of a proper scoring rule and a weight function $w(y)$ which depends on the observation y yields an improper scoring rule.

Theorem 2.27 (Gneiting and Ranjan, 2011b)**.** *Suppose that f is the sampling density of the random variable Y. Let S_0 be any proper scoring rule and let w be a weight function such that $0 < \int w(y)f(y)\mathrm{d}y < \infty$. Then the expected value of the score*

$$\boldsymbol{S}(h, Y) = w(Y)\boldsymbol{S}_0(h, Y)$$

is minimized by the density forecast

$$g(y) = \frac{w(y)f(y)}{\int w(y)f(y)\mathrm{d}y}.$$

Proof. Theorem 1 of Gneiting and Ranjan (2011b, page 415). □

Thus, $\boldsymbol{S}(h, Y) = w(Y)\boldsymbol{S}_0(h, Y)$ is improper except for constant weight functions w since the expected score is minimized by issuing the density forecast g which differs from the true sampling density f. By choosing $w(y) = \mathbb{1}(y \geq t)$ for a threshold $t \in \mathbb{R}$, Theorem 2.27 implies that restricting proper scoring rules to extreme events corresponds to the use of improper scoring rules. Thus, computing proper scoring rules for subsets of extreme events will discredit skillful and calibrated forecasting procedures.

This can be illustrated using an example given by Diks et al. (2011).

Example 2.28. The logarithmic score restricted to the set of observations larger than $t \in \mathbb{R}$ is given by

$$\mathrm{LogS}^\star(f, y) = -\mathbb{1}(y \geq t) \log f(y)$$

as proposed by Amisano and Giacomini (2007), who use this scoring rule in the context of a weighted likelihood ratio test. If it holds that

$$g(y) > f(y)$$

for all $y \geq t$ and two competing density forecasts f and g, then

$$\mathbb{E}\mathrm{LogS}^\star(g, y) < \mathbb{E}\mathrm{LogS}^\star(f, y)$$

independent of the true sampling density.

This may have far-reaching consequences: A fat-tailed density forecast g might be preferred over a thin-tailed density f, even if f is the true sampling density. Obviously, the restricted logarithmic score $\mathrm{LogS}^\star$ is not proper.

Theorem 2.27 furthermore suggests a hedging strategy if the improper scoring rule $S(h, y) = w(y)S_0(h, y)$ is used. The forecaster will minimize the expected score by issuing the density forecast g which is proportional to his true belief f and the weight function w (Gneiting and Ranjan, 2011b). In the situation of Example 2.28, the optimal density forecast is given by

$$g(y) = \frac{f(y)}{\int_{[t,\infty)} f(y)\mathrm{d}y}\mathbb{1}(y \geq t).$$

We will now discuss the observation of Theorem 2.27 in view of the mathematical framework developed in Section 2.1.1. For observations

$$Y : (\Omega, \mathcal{A}, \mathcal{Q}) \longrightarrow (\Omega_Y, \mathcal{A}_Y)$$

and probabilistic forecasts in the form of predictive densities

$$f : (\Omega, \mathcal{A}) \longrightarrow (\mathcal{P}, \mathcal{A}_{\mathcal{P}}),$$

a proper scoring rule S is a mapping

$$S : \mathcal{P} \times \Omega_Y \longrightarrow \bar{\mathbb{R}}.$$

Conditioning Y on being an extreme event corresponds to restricting the observation space $(\Omega_Y, \mathcal{A}_Y)$ to a subspace $(\Omega_Y^\star, \mathcal{A}_Y^\star)$. Therefore, the scoring rule restricted to the subset of extreme events, $S^\star$, is a mapping

$$S^\star : \mathcal{P} \times \Omega_Y^\star \longrightarrow \bar{\mathbb{R}}$$

and is thus minimized by

$$\mathcal{L}(Y|\mathcal{A}^\star) = \mathcal{L}(Y^\star) = F^\star \neq F = \mathcal{L}(F) = \mathcal{L}(Y|\mathcal{A}).$$

$S^\star$ is minimized by a different element of $\mathcal{P}$ and therefore an improper scoring rule, even if S is proper.

2.3.2. Proper scoring rules for extreme events

The previous section showed that the natural approach of restricting proper scoring rules to subsets of extreme events yields improper scoring rules which allow for a simple hedging strategy. Therefore, proper scoring rules restricted to subsets of extreme events should not be used to evaluate probabilistic forecasts and new methods for the evaluation of probabilistic forecasts have to be developed. Here, we investigate two methods which have recently been proposed in the economic literature (Gneiting and Ranjan, 2011b; Diks et al., 2011).

We consider density forecasts in a time series context, where density forecasts for an observation which lies k time steps ahead are issued. Competing density forecasts $\hat{f}_{t+k}$ and $\hat{g}_{t+k}$ are generated at times $t = 1, \ldots, n-k$, where n is the total number of observations. Gneiting and Ranjan (2011b) furthermore require the forecasts to be produced only depending on the data in a rolling estimation window which consists of the past m observations. However, for the purpose of comparing the approaches of Gneiting and Ranjan (2011b) and Diks et al. (2011), we will omit this. The different size of information bases of different forecasters might of course explain differences of the predictive performance. Therefore, we will always point out if different amounts of data are taken into account to produce the forecasts.

Threshold- and quantile-weighted scoring CRPS

We now turn to appropriately weighted, proper versions of the continuous ranked probability score as proposed by Gneiting and Ranjan (2011b). While in Definition 2.26, the CRPS was defined in terms of a predictive cumulative distribution function F, this can easily be extended to predictive densities. Any density forecast f can be identified with the corresponding predictive cumulative distribution function F. Therefore, we can simply define

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

where F is the predictive CDF corresponding to the density forecast f and y is observed.

The CRPS can be represented in three equivalent ways,

$$\begin{aligned} \mathrm{CRPS}(f, y) &= \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 \mathrm{d}z && (2.3) \\ &= \mathbb{E}_f |Y - y| - \frac{1}{2} \mathbb{E}_f |Y - Y'| && (2.4) \\ &= 2 \int_0^1 (\mathbb{1}\{y \leq F^{-1}(\alpha)\} - \alpha)(F^{-1}(\alpha) - y) \mathrm{d}\alpha, && (2.5) \end{aligned}$$

where Y and Y' are independent random variables with sampling density f. The *kernel score representation* (2.4) was proved by Gneiting and Raftery (2007) and already discussed above. Laio and Tamea (2007) showed the equivalence of the *quantile score representation* (2.5) and the *threshold decomposition* (2.3) (Gneiting and Ranjan, 2011b). The quantile score representation corresponds to the integral over the *quantile score*

$$\mathrm{QS}_\alpha(F^{-1}(\alpha), y) = 2(\mathbb{1}\{y \leq F^{-1}(\alpha)\} - \alpha)(F^{-1}(\alpha) - y)$$

over all $\alpha \in [0, 1]$, where $F^{-1}(\alpha)$ is the quantile forecast for α corresponding to the predictive CDF F.

The threshold decomposition and the quantile score representation can be used to construct weighted versions of the continuous ranked probability score.

Definition 2.29. (a) The *threshold-weighted continuous ranked probability score* is given by

$$\mathrm{CRPS}^{\mathrm{t}}(f, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 u(z) \mathrm{d}z, \qquad (2.6)$$

where u is a nonnegative weight function on the real line.

(b) The *quantile-weighted continuous ranked probability score* is given by

$$\mathrm{CRPS}^{\mathrm{q}}(f, y) = 2 \int_0^1 (\mathbb{1}\{y \leq F^{-1}(\alpha)\} - \alpha)(F^{-1}(\alpha) - y) v(\alpha) \mathrm{d}\alpha, \qquad (2.7)$$

where v is a nonnegative weight function on $[0, 1]$.

Matheson and Winkler (1976) prove that the threshold-weighted CRPS and the quantile-weighted CRPS are both proper scoring rules (Gneiting and Ranjan, 2011b).

For $u \equiv 1$ and $v \equiv 1$, the usual unweighted CRPS is obtained. Different weight functions can be used to emphasize specific regions of interest. If the interest lies in the predictive performance in case of extreme events, u might be chosen as an indicator function $u(z) = \mathbb{1}(z \geq r)$, which is 1 for values larger than a threshold $r \in \mathbb{R}$ and v might be chosen as an indicator function $v(\alpha) = \mathbb{1}(\alpha \geq q)$, which is 1 for quantiles larger than a threshold $q \in [0, 1]$. Furthermore, Gneiting and Ranjan (2011b) propose potential weight functions if the interest lies in the center, tails, right or left tail of the distribution which are listed in Table 2.1.

Table 2.1.: Table 4 of Gneiting and Ranjan (2011b, page 416). Proposed weight functions for threshold- and quantile-weighted CRPS. $\varphi_{a,b}$ and $\Phi_{a,b}$ denote density function and CDF of a normal distribution with mean a and standard deviation b.

Emphasis	Threshold weight function	Quantile weight function
Center	$u_1(z) = \varphi_{a,b}(z)$	$v_1(\alpha) = \alpha(1-\alpha)$
Tails	$u_2(z) = 1 - \varphi_{a,b}(z)/\varphi_{a.b}(0)$	$v_2(\alpha) = (2\alpha - 1)^2$
Right tail	$u_3(z) = \Phi_{a,b}(z)$	$v_3(\alpha) = \alpha^2$
Left tail	$u_4(z) = 1 - \Phi_{a,b}(z)$	$v_4(\alpha) = (1-\alpha)^2$

Alternative weight functions may be suggested by specific applications at hand. Gneiting and Ranjan (2011b) illustrate this using a data example of wind speed forecasts. Optimal wind speed point forecasts are α-quantiles of the predictive distribution, where α depends on market conditions. Typical market conditions and costs associated with over- or underpredictions suggest a triangular quantile weight function $v(\alpha) = \triangle_{0.73}(\alpha)$, which has a peak of height 1 at $\alpha = 0.73$ and decays to 0 at $\alpha = 0$ and $\alpha = 1$.

If no closed-form expressions of (2.6) and (2.7) are available, discretized approximations can be computed. The threshold-weighted CRPS (2.6) can be approximated by

$$\mathrm{CRPS}^{\mathrm{t}}(f, y) \approx \frac{y_u - y_l}{I - 1} \sum_{i=1}^{I} (F(y_i) - \mathbb{1}\{y \leq y_i\})^2 u(y_i),$$

where (y_l, y_u) is the range of interest and $y_i = y_l + i\frac{y_u - y_l}{I}$. The quantile-weighted CRPS (2.7) can be approximated by

$$\mathrm{CRPS}^{\mathrm{q}}(f, y) \approx \frac{1}{J - 1} \sum_{j=1}^{J-1} 2(\mathbb{1}\{y \leq F^{-1}(\alpha_j)\} - \alpha_j)(F^{-1}(\alpha_j) - y)v(\alpha_j),$$

where $\alpha_j = \frac{j}{J}$. These discretizations are feasible since discrete versions of proper scoring rules are themselves proper scoring rules (Gneiting and Ranjan, 2011b).

Conditional and censored likelihood scoring rules

While Gneiting and Ranjan (2011b) generalize the continuous ranked probability score in order to emphasize different regions of interest, the approach of Diks et al. (2011) is based on the comparison of density forecasts using the Kullback-Leibler information criterion (KLIC) as suggested by Amisano and Giacomini (2007). The Kullback-Leibler information criterion for a density forecast $\hat{f}$ is defined as

$$\begin{aligned}\mathrm{KLIC}(\hat{f}) &= \mathbb{E}_p(\log p(Y) - \log \hat{f}(Y)) \\ &= \int_{-\infty}^{\infty} p(y) \log\left(\frac{p(y)}{\hat{f}(y)}\right) \mathrm{d}y,\end{aligned}$$

where p denotes the true sampling density. Lower expected values of the logarithmic score with respect to the true density p,

$$\boldsymbol{S}(f, y) = \mathbb{E}_p[-\log f(Y)],$$

are equivalent to lower values of the KLIC. The KLIC has absolute lower bound 0 which is achieved if and only if $\hat{f} = p$ and thus can be used as a measure of the divergence between $\hat{f}$ and p (Diks et al., 2011). If p is unknown, the KLIC cannot be evaluated directly, but can still be used to measure the relative accuracy of two competing density forecasts $\hat{f}$ and $\hat{g}$ by using the difference $\mathrm{KLIC}(\hat{f}) - \mathrm{KLIC}(\hat{g})$. The term $\mathbb{E}_p[\log p(Y)]$ drops out which yields the logarithmic score difference $d = \log \hat{g}(y) - \log \hat{f}(y)$. However, the KLIC cannot be used directly to measure the accuracy of a density forecast in a specific region of interest (Diks et al., 2011). This corresponds to the results and observations of Theorem 2.27 and Example 2.28. Restricting the logarithmic score to a specific region of interest, which is of course equivalent to restricting the KLIC to this region, corresponds to the use of an improper scoring rule.

The aim of Diks et al. (2011) is to address this deficiency of the approach of Amisano and Giacomini (2007) and to generalize likelihood-based scoring rules in order to compare density forecasts on a specific region of interest. The main reasons for this pursuing are that likelihood-based scoring rules are invariant under transformations of the observation space and that they lead to likelihood ratio tests which have are known to perform well in many statistical settings (Diks et al., 2011).

This aim is achieved by replacing the full likelihood with the conditional likelihood given that the observation lies in the region of interest or with the censored likelihood.

Definition 2.30. Given a region of interest $A \subset \mathbb{R}$, the *conditional likelihood (CL) scoring rule* is defined as

$$\mathrm{CL}(f, y) = -\mathbb{1}(y \in A) \log\left(\frac{f(y)}{\int_A f(s)\mathrm{d}s}\right), \tag{2.8}$$

where f is a density forecast and y is observed.

Using the CL scoring rule, density forecasts are evaluated only based on their behavior in the region of interest. Density forecasts are normalized on the region of interest, such that they can be compared in terms of their relative KLIC (Diks et al., 2011). However, this normalization also gives rise to the problem that the total probability assigned to the region of interest is not taken into account. Therefore, the CL scoring rule should only be used to compare density forecasts which assign similar probabilities to the region of interest. For the purpose of taking the tail probability into account, Diks et al. (2011) propose the censored likelihood scoring rule which censors observations outside the region of interest.

Definition 2.31. Given a region of interest $A \subset \mathbb{R}$, the *censored likelihood (CSL) scoring rule* is defined as

$$\mathrm{CSL}(f, y) = -\left[\mathbb{1}(y \in A) \log f(y) + \mathbb{1}(y \in A^c) \log\left(\int_{A^c} f(s)\mathrm{d}s\right)\right], \tag{2.9}$$

where f is a density forecast, y is observed and A^c is the complement of A.

Note that both (2.8) and (2.9) differ from the definitions of Diks et al. (2011) in their signs because we take scoring rules to be negatively oriented.

The previous definitions of the conditional and censored likelihood scoring rules depend on a specific region of interest. However, they can easily be generalized by replacing the indicator functions $\mathbb{1}(y \in A)$ with a weight function $w(y)$. This allows for emphasizing regions of interest in a more general way. Therefore, we redefine the conditional and censored likelihood scoring rule.

Definition 2.32. (a) Given a weight function $w(y)$ on the real line, the *generalized conditional likelihood (CL) scoring rule* is defined as

$$\mathrm{CL}(f, y) = -w(y) \log\left(\frac{f(y)}{\int w(s) f(s)\mathrm{d}s}\right), \tag{2.10}$$

where f is a density forecast and y is observed.

(b) Given a weight function $w(y)$ on the real line, the *generalized censored likelihood (CSL) scoring rule* is defined as

$$\mathrm{CSL}(f, y) = -\left[w(y) \log f(y) + (1 - w(y)) \log\left(1 - \int w(s) f(s)\mathrm{d}s\right)\right], \tag{2.11}$$

where f is a density forecast and y is observed.

Obviously, the original definition follows as a special case for $w(y) = \mathbb{1}(y \in A)$. As for the threshold- and quantile-weighted versions of the CRPS, the computation of discretized versions of the CL and CSL scoring rule is feasible.

Theorem 2.33 (Diks et al., 2011)**.** *If*

(i) the two density forecasts f and g satisfy $\mathrm{KLIC}(f) < \infty$ *and* $\mathrm{KLIC}(g) < \infty$*, where* $\mathrm{KLIC}(h) = \int p(y) \log(p(y)/h(y))\mathrm{d}y$ *is the Kullback-Leibler divergence between the density forecast h and the true density p, and*

(ii) (a) the weight function w is only determined by the information available at the time when the outcome is observed, (b) $0 \leq w(y) \leq 1$ and (c) $\int w(y)p(y)\mathrm{d}y > 0$,

the generalized conditional likelihood (CL) scoring rule given in (2.10) *and the generalized censored likelihood (CSL) scoring rule given in* (2.11) *are proper.*

Proof. Lemma 1 of Diks et al. (2011, page 220) □

Tests of equal performance

For the purpose of comparing the predictive performance of competing density forecasts on several specific regions of interest, for example in case of more and more extreme events, the values of the corresponding proper scoring rules are of different magnitudes and cannot be directly compared. If, for example, we use two threshold-weighted versions of the continuous ranked probability score, $\mathrm{CRPS}_i^{\mathrm{t}}$ for $i = 1, 2$ with weight functions $w_i(y) = \mathbb{1}(y \geq r_i)$, where $r_1 < r_2$, then

$$\mathrm{CRPS}_1^{\mathrm{t}}(f, y) > \mathrm{CRPS}_2^{\mathrm{t}}(f, y)$$

for all density forecasts f and observations y. Different thresholds r_i result in expected scores of different magnitudes.

To compare such results for competing density forecasts, formal tests of equal performance can be applied. Both Gneiting and Ranjan (2011b) and Diks et al. (2011) follow Amisano and Giacomini (2007) in using test statistics proposed by Diebold and Mariano (1995).

As before, we consider density forecasts for an observation which lies k time steps ahead. Competing density forecasts $\hat{f}_{t+k}$ and $\hat{g}_{t+k}$ are generated at times $t = 1, \ldots, n-k$ and ranked by comparing their average scores. For

$$\bar{S}_n^f = \frac{1}{n-k+1} \sum_{t=1}^{n-k} S(\hat{f}_{t+k}, y_{t+k}) \text{ and}$$
$$\bar{S}_n^g = \frac{1}{n-k+1} \sum_{t=1}^{n-k} S(\hat{g}_{t+k}, y_{t+k}),$$

we prefer f if $\bar{S}_n^f < \bar{S}_n^g$ and g otherwise (Gneiting and Ranjan, 2011b). We will consider tests of equal performance based on the *Diebold-Mariano-type* test statistic

$$t_n = \sqrt{n}\, \frac{\bar{S}_n^f - \bar{S}_n^g}{\hat{\sigma}_n}, \tag{2.12}$$

where $\hat{\sigma}_n^2$ is an estimator of the asymptotic variance of the score difference. Under certain regularity conditions, t_n asymptotically follows a standard normal distribution under the null hypothesis of equal performance. This allows for interpretations of the observed average score differences in terms of significance. The approaches of Gneiting and Ranjan (2011b) and Diks et al. (2011) differ not only in their

choice of the scoring rule $S(f, y)$, but also in the way of estimating the asymptotic variance of the score difference.

Gneiting and Ranjan (2011b) propose threshold- and quantile-weighted versions of the CRPS. They estimate the asymptotic variance by

$$\hat{\sigma}_n^2 = \frac{1}{n+k-1} \sum_{j=-(k-1)}^{k-1} \sum_{t=1}^{n-k-|j|} \Delta_{t,k} \Delta_{t+|j|,k},$$

where $\Delta_{t,k} = S(\hat{f}_{t+k}, y_{t+k}) - S(\hat{g}_{t+k}, y_{t+k})$, as proposed by Diebold and Mariano (1995). This variance estimator only considers autocorrelation of forecast errors up to lag $k-1$. Diebold and Mariano (1995) note that *optimal* k-step-ahead forecast errors are at most $(k-1)$-dependent. However, in practical applications, $(k-1)$-dependence might be violated due to various reasons. Even so, they still suggest to take $(k-1)$-dependence as a "reasonable benchmark for a k-step-ahead forecast error" (Diebold and Mariano, 1995, page 254). Therefore, the variance estimator chosen by Gneiting and Ranjan (2011b) seems to be a reasonable choice. Gneiting and Ranjan (2011b) prove the asymptotic normality of the test statistic t_n under a moment condition which can be reduced to the rule of thumb that "the normal approximation for t_n is appropriate unless the forecast densities have infinite moments of low order" (Gneiting and Ranjan, 2011b, page 417).

Diks et al. (2011) only consider 1-step-ahead forecasts and use the conditional (CL) and censored (CSL) likelihood scoring rules to rank competing density forecasts. They propose the use of a heteroskedasticity and autocorrelation-consistent (HAC) variance estimator $\hat{\sigma}_n^2$, as for example

$$\hat{\sigma}_n^2 = \hat{\gamma}_0 + 2 \sum_{j=1}^{J-1} a_j \hat{\gamma}_j,$$

where $\hat{\gamma}_j$ denotes the lag-j-sample covariance of the sequence of score differences, $a_j = \frac{1-j}{J}$ are the Bartlett weights and $J = \lfloor n^{1/4} \rfloor$. This estimation procedure takes into account autocorrelation up to a lag larger than $k-1$ dependent on the sample size n.

Taking into account autocorrelation up to a lag larger than $k-1$ following Diks et al. (2011) will in general lead to larger estimates of the asymptotic variance than those obtained following Gneiting and Ranjan (2011b). Given that the same scoring rules are used, this results in smaller absolute values of the test statistic t_n and suggests less significant differences of the predictive performance of two competing forecasting procedures. Therefore, we expect a tendency to find more significant score differences when following Gneiting and Ranjan (2011b), whose approach is based on the assumption of at most $(k-1)$-dependence. This assumption was motivated by theoretical results and can be readily assessed empirically (Diebold and Mariano, 1995).

In the current chapter, we have seen that during the last two decades, a large variety of methods to assess the predictive performance of probabilistic forecasts of continuous, real-valued variables has been developed. The evaluation of prob-

abilistic forecasts should follow the paradigm of maximizing the sharpness of the predictive distribution subject to calibration. Proper scoring rules provide summary measures of predictive performance by simultaneously addressing calibration and sharpness (Gneiting and Ranjan, 2011b). The literature concerned with evaluation of probabilistic forecasts of rare and extreme events is more sparse. The natural approach of applying standard evaluation procedures after selecting extreme events and discarding non-extreme events is bound to fail since it results in the use of improper scoring rules. Therefore, forecast evaluation should not be carried out conditional on the outcome being an extreme event. However, proper scoring rules such as the CRPS or the logarithmic score can be extended by using weight functions which emphasize specific regions of interest. In the following chapter, we illustrate these results using a simulation study. Furthermore, we will compare the approaches to forecast evaluation for extreme events of Gneiting and Ranjan (2011b) and Diks et al. (2011).

3. Simulation study

3.1. Mathematical framework

Based on the simulation study of Gneiting et al. (2007), we consider a simulation study where at each time $t = 1, 2, \ldots, T$, nature chooses a distribution G_t which we think of as the true data-generating process and each forecaster picks a predictive distribution F_t. Here, $G_t = \mathcal{N}(\mu_t, 1)$, where μ_t is drawn from a standard normal distribution for each t. In a weather-forecasting context, μ_t might be thought of as an "accurate description of the latest observable state of the atmosphere, summarizing all information that a forecaster might possibly have access to" (Gneiting et al., 2007, page 244).

We compare different forecasting procedures which are summarized in Table 3.1. The ideal forecaster has complete knowledge of the current state μ_t. In the framework of Section 2.1.1, his information basis contains all available information and he issues the ideal probabilistic forecast relative to this information basis which is the true distribution of the data-generating process. Thus, this forecast is ideal relative to the sub-σ-algebra generated by μ_t and the probabilistic forecast is probabilistically and marginally calibrated. The climatological forecaster predicts the unconditional distribution $F_t = \mathcal{N}(0, 2)$ which is ideal relative to the trivial sub-σ-algebra and is thus probabilistically and marginally calibrated as well. The unfocused forecaster observes μ_t, but adds a bias in form of a mixture component which is a normal distribution with standard deviation 1 and mean value 1 or -1 with equal probability. Gneiting et al. (2007) show that the unfocused forecaster is probabilistically calibrated, but not marginally calibrated. The sign-biased forecaster correctly observes μ_t except for the sign and is marginally, but not probabilistically calibrated (Gneiting et al., 2007). Therefore, for both the unfocused and the sign-biased forecaster, there exists no sub-σ-algebra such that they are ideal relative to this sub-σ-algebra. This follows directly from Theorem 2.6

Table 3.1.: Mathematical scenario for the simulation study with $G_t = \mathcal{N}(\mu_t, 1)$, $\mu_t \sim \mathcal{N}(0, 1)$, $a = 2.5$ and $\tau_t = \pm 1$ with probability $\frac{1}{2}$ each.

Forecaster	F_t
Ideal	$\mathcal{N}(\mu_t, 1)$
Climatological	$\mathcal{N}(0, 2)$
Unfocused	$\frac{1}{2}\{\mathcal{N}(\mu_t, 1) + \mathcal{N}(\mu_t + \tau_t, 1)\}$
Sign-biased	$\mathcal{N}(-\mu_t, 1)$
Biased	$\mathcal{N}(\mu_t + a, 1)$

(Gneiting and Ranjan, 2011a). The biased forecaster here adds a constant bias of 2.5 to the true mean value μ_t. Due to this constant bias, the biased forecaster will obviously be neither probabilistically nor marginally calibrated.

Gneiting et al. (2007) use their simulation study for the purpose of illustrating the disconcerting result that forecast evaluation solely based on checking the PIT histograms for uniformity is unable to distinguish between the ideal, the climatological and the unfocused forecaster. Despite predicting the true CDF, the ideal forecaster would not be preferred over these competitors. The authors address this deficiency by proposing the paradigm of maximizing the sharpness of the predictive distributions subject to calibration and show that proper scoring rules such as the CRPS or the logarithmic score are able to distinguish between the ideal forecaster and his competitors.

Here, we are interested in the predictive performance of the competing forecasting procedures in case of extreme events. We will briefly summarize the results of Gneiting et al. (2007) and then turn to an application of the evaluation procedures for probabilistic forecasts discussed in the previous chapter. This will provide further insight into differences, advantages and disadvantages of the two approaches proposed by Gneiting and Ranjan (2011b) and Diks et al. (2011).

3.2. Results for all events

3.2.1. Calibration

We repeat the prediction experiment $T = 10\,000$ times. As noted above, theoretical results suggest that the ideal, the climatological and the unfocused forecaster are probabilistically calibrated. Figure 3.1 shows PIT histograms for all forecasters. While only minor deviations from uniformity for the ideal, the climatological and the unfocused forecaster empirically confirm these findings, the U-shaped PIT histogram of the sign-biased forecaster indicates underdispersion and prediction intervals that are too narrow on average. The triangle-shaped PIT histogram of the biased forecaster is caused by the constant bias and indicates lack of probabilistic calibration.

Marginal calibration can be assessed not only by comparing predictive and observed cumulative distribution functions, but also by comparing marginal predictive densities, as shown in Figure 3.2. The marginal density curves of the climatological and the sign-biased forecaster are identical to the curve of the ideal forecaster, who predicts the true density of the observations. Therefore, the ideal, the climatological and the sign-biased forecaster are marginally calibrated. The marginal density curve of the unfocused forecaster matches the true marginal density in location, but not in shape, with a standard deviation which seems slightly too high. Due to the constant mean bias, the marginal density curve of the biased forecaster matches the true marginal density in shape, but not in location, being shifted to the right by 2.5. In accordance with the theoretical results, both the unfocused and the biased forecaster appear to be not marginally calibrated.

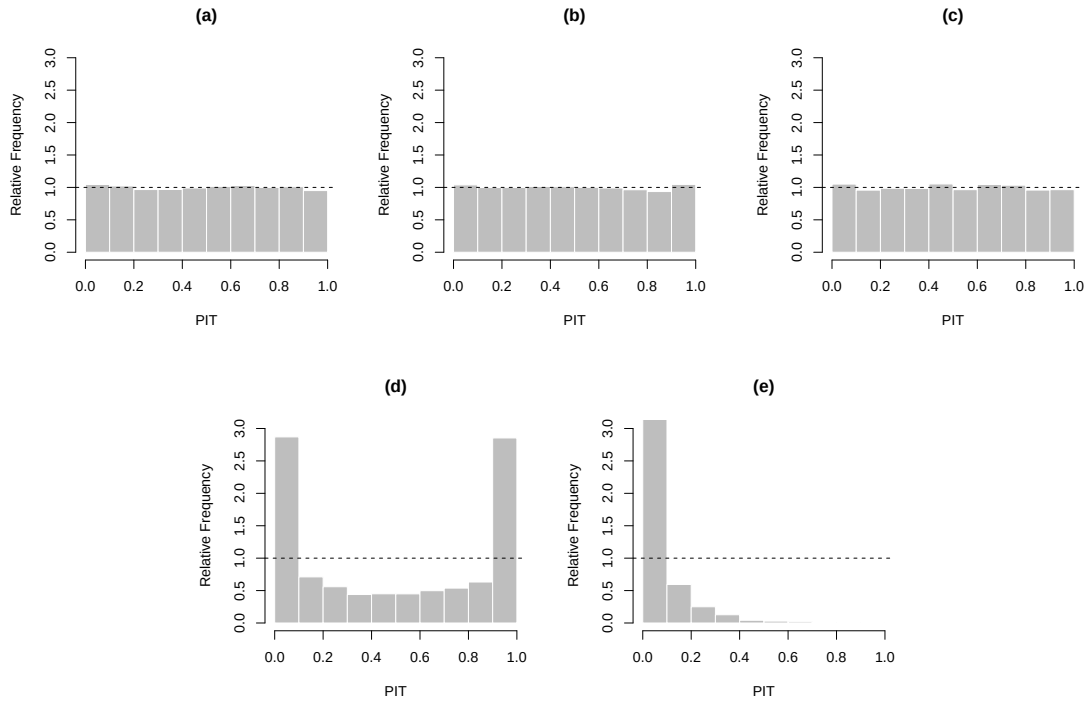

Figure 3.1.: PIT histograms for (a) the ideal forecaster, (b) the climatological forecaster, (c) the unfocused forecaster, (d) the sign-biased forecaster and (e) the biased forecaster for 10 000 repetitions of the prediction experiment.

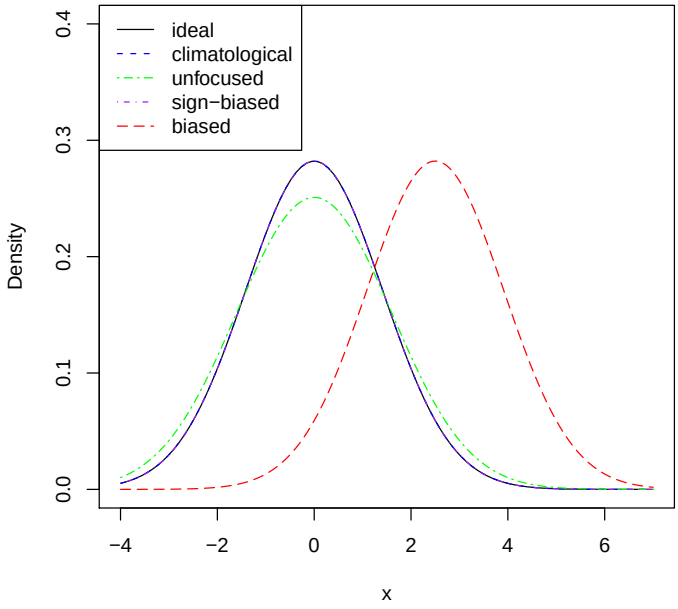

Figure 3.2.: Marginal predictive densities of the forecasters given in Table 3.1.

Table 3.2.: Average values of the CRPS, the LogS, the mean absolute error (MAE) and the empirical coverage (in %) of 80% prediction intervals.

Forecaster	CRPS	LogS	MAE	Coverage
Ideal	**0.56**	**1.42**	**0.80**	**80.1**
Climatological	0.79	1.76	1.12	80.4
Unfocused	0.63	1.53	0.89	80.3
Sign-biased	1.38	3.39	1.77	43.6
Biased	1.97	4.52	2.49	11.1

3.2.2. Summary measures

Table 3.2 shows values of various summary measures of predictive performance for the competing forecasters. Here, CRPS and LogS denote the mean continuous ranked probability score, $\frac{1}{T}\sum_{t=1}^{T}\mathrm{CRPS}(F_t, y_t)$, and the mean logarithmic score, $\frac{1}{T}\sum_{t=1}^{T}\mathrm{LogS}(F_t, y_t)$, respectively. The mean absolute error (MAE) is the mean absolute difference between the mean value of the predictive distribution and the observation. The average coverage of, for example, 80% prediction intervals is the relative frequency with which observations fall into 80% prediction intervals centered around the mean value of the predictive distribution. The average coverage can be used to assess probabilistic calibration which is indicated by values close to the theoretical coverage of 80%. As expected, the ideal forecaster outperforms any other forecasting procedure in terms of these summary measures. Being neither probabilistically nor marginally calibrated, the biased forecaster performs by far the worst. However, as we see below, these results change dramatically if only subsets of extreme events are considered.

3.3. Results for extreme events

3.3.1. Calibration

Here, we condition the observations on being extreme events. More precisely, we select all observations larger than the 99th percentile of the marginal distribution of the observations and discard the rest. PIT histograms for this subset of extreme events are shown in Figure 3.3. Although the ideal, the climatological and the unfocused forecaster are probabilistically calibrated, they appear not to be probabilistically calibrated for the extreme events. The same holds for the sign-biased forecaster, who is not probabilistically calibrated, neither for all events nor for extreme events. The biased forecaster still appears not to be probabilistically calibrated. However, the deviations from uniformity are much smaller compared to the competing forecasting procedures.

A similar observation can be made when marginal calibration is examined. Figure 3.4 shows marginal predictive densities conditional on the observation being an extreme event. Although there was a large concordance between the marginal densities of the ideal, the climatological, the unfocused and the sign-biased fore-

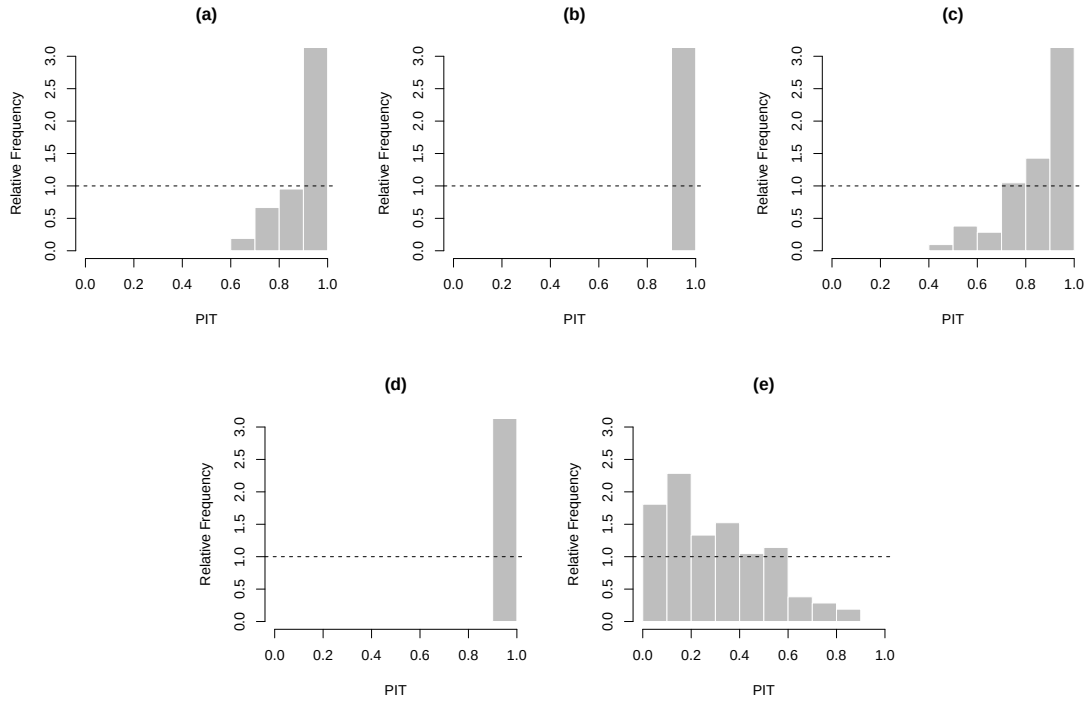

Figure 3.3.: PIT histograms for (a) the ideal forecaster, (b) the climatological forecaster, (c) the unfocused forecaster, (d) the sign-biased forecaster and (e) the biased forecaster using all observations larger than the 99th percentile of the marginal distribution of the observations.

caster in the unconditional case, none of these competing conditional predictive densities matches the true marginal conditional density of the observation in location or shape. Note that this in particular holds for the ideal forecaster due to the difference between the unconditional and the conditional distribution. The conditional predictive density of the biased forecaster seems to have large parts of its mass in the correct location, while the shape strongly differs from the true marginal conditional density.

3.3.2. Proper scoring rules restricted to extreme events

Table 3.3 shows the results of the summary measures for the subset of extreme events. The biased forecaster outperforms all competing forecasting procedures in terms of all summary measures and would be preferred over the ideal forecaster although the ideal forecaster issues the true unconditional distribution of the observations. Therefore, if the forecast evaluation was only based on this subset of extreme events, skillful and calibrated forecasting procedures would be discredited.

Figure 3.5 shows plots of the summary measures as functions of the threshold which defines extreme events. For example, the value of the mean restricted CRPS for the threshold 1.5 is the value of the mean CRPS computed only based on all observations larger than 1.5. The values of mean CRPS, mean LogS and

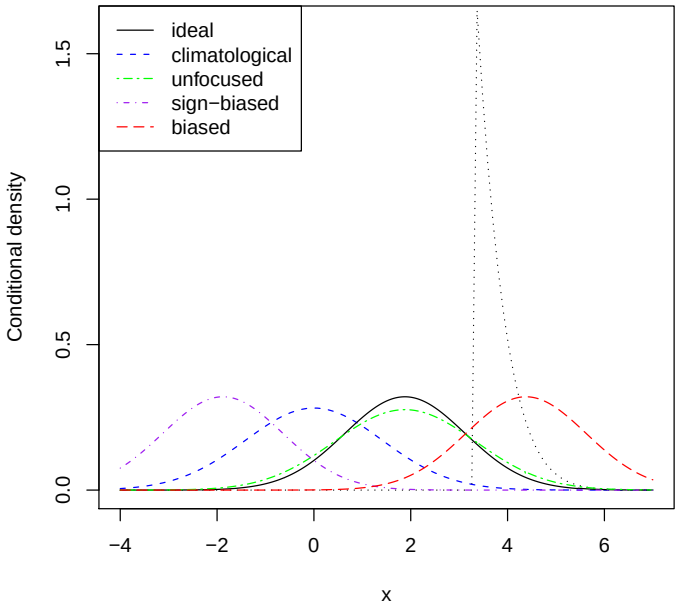

Figure 3.4.: Marginal predictive densities of the forecasters given in Table 3.1 given that the observation is larger than the 99th percentile of the marginal distribution, for which the corresponding conditional density is indicated by the dotted black line.

Table 3.3.: Average values of the CRPS, the LogS, the mean absolute error (MAE) and the empirical coverage (in %) of 80% prediction intervals for the subset of observations larger than the 99th percentile of the marginal distribution of the observations.

Forecaster	CRPS	LogS	MAE	Coverage
Ideal	1.36	8.47	1.86	18.1
Climatological	2.92	4.75	3.72	0.0
Unfocused	1.34	2.69	1.84	30.0
Sign-biased	5.01	16.87	5.58	0.0
Biased	**0.55**	**1.38**	**0.79**	**81.9**

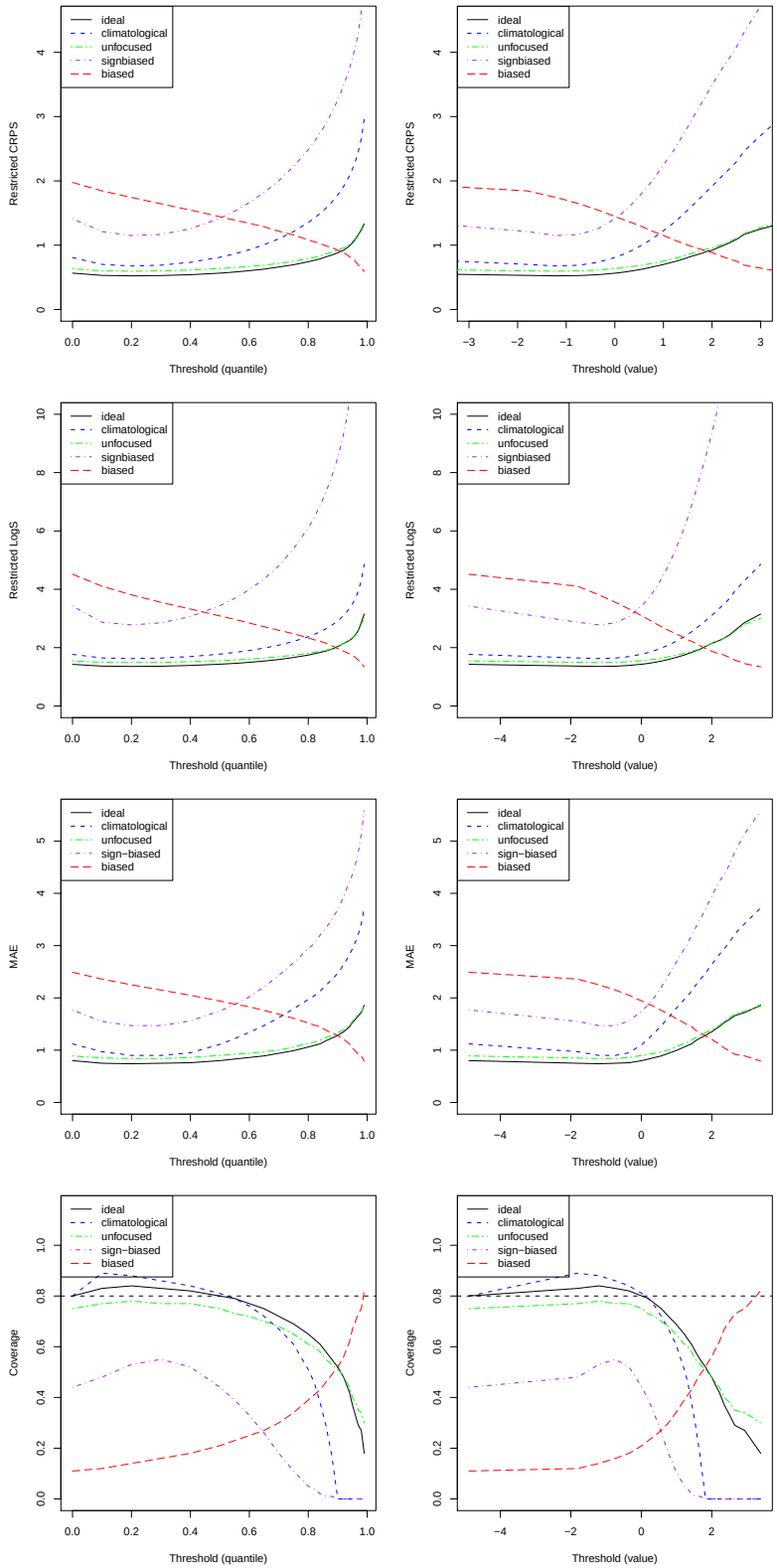

Figure 3.5.: Summary measures for subsets of extreme values as functions of the threshold defining extreme values in terms of quantiles of the marginal distribution (left column) or values (right column).

MAE increase for larger thresholds for all forecasting procedures except the biased forecaster, for which they decrease. For thresholds larger than around 2, which approximately corresponds to the 90th percentile of the marginal distribution of the observations, the biased forecaster outperforms any competing forecasting procedure. In particular, the biased forecaster outperforms the ideal forecaster although the ideal forecaster predicts the true unconditional density. Therefore, we observe what was to be expected from the theoretical results of Theorem 2.27: Restricting proper scoring rules to extreme events corresponds to the use of an improper scoring rule. Forecast evaluation based on subsets of extreme events will discredit skillful and calibrated forecasting procedures, even the ideal forecaster who predicts the true unconditional density.

Qualitatively, the same behavior can be observed for the average coverage of 80% prediction intervals, see bottom row of Figure 3.5. For mostly non-extreme events, the ideal, the climatological and the unfocused forecaster are probabilistically calibrated, the average coverage is close to the theoretical coverage of 80%. For subsets of more and more extreme events, the coverage progressively decreases. On the contrary, the coverage of 80% prediction intervals for the biased forecaster approaches the theoretical value of 80% for more and more extreme thresholds.

Instead of plotting the values of scoring rules as functions of the threshold defining extreme events, we may also use test of equal performance by computing values of Diebold-Mariano-type test statistics

$$t_n = \sqrt{n}\,\frac{\bar{S}_n^f - \bar{S}_n^g}{\hat{\sigma}_n} \tag{3.1}$$

depending on this threshold, where S denotes the corresponding restricted scoring rule. Computing p-values associated with the values of the test statistic under the standard normal hypothesis gives us further insight into the significance of the observed score differences.

Figure 3.6 shows the test statistics of the test of equal performance comparing the ideal and the biased forecaster for the restricted CRPS (left column) and the restricted LogS (right column) together with p-values associated with the values of the test statistic under the standard normal hypothesis. If the test statistic attains values smaller than 0, we prefer the ideal forecaster over the biased forecaster, otherwise we prefer the biased forecaster over the ideal forecaster.

However, note that the standard normal hypothesis might be violated for larger thresholds due to the increasingly smaller sample sizes. For the restricted CRPS, $\hat{\sigma}_n$ was estimated as proposed by Gneiting and Ranjan (2011b), for the restricted LogS, the HAC estimator proposed by Diks et al. (2011) was used. Qualitatively comparable results were obtained if the respectively different variance estimation procedure was used. However, since we are interested in studying proper scoring rules for the evaluation of probabilistic forecasts for extreme events and we already saw that restricted versions of the CRPS and the LogS are improper, we postpone the analysis of variance estimation effects to the next section where we will investigate the threshold-weighted CRPS as well as the conditional and censored likelihood scoring rule.

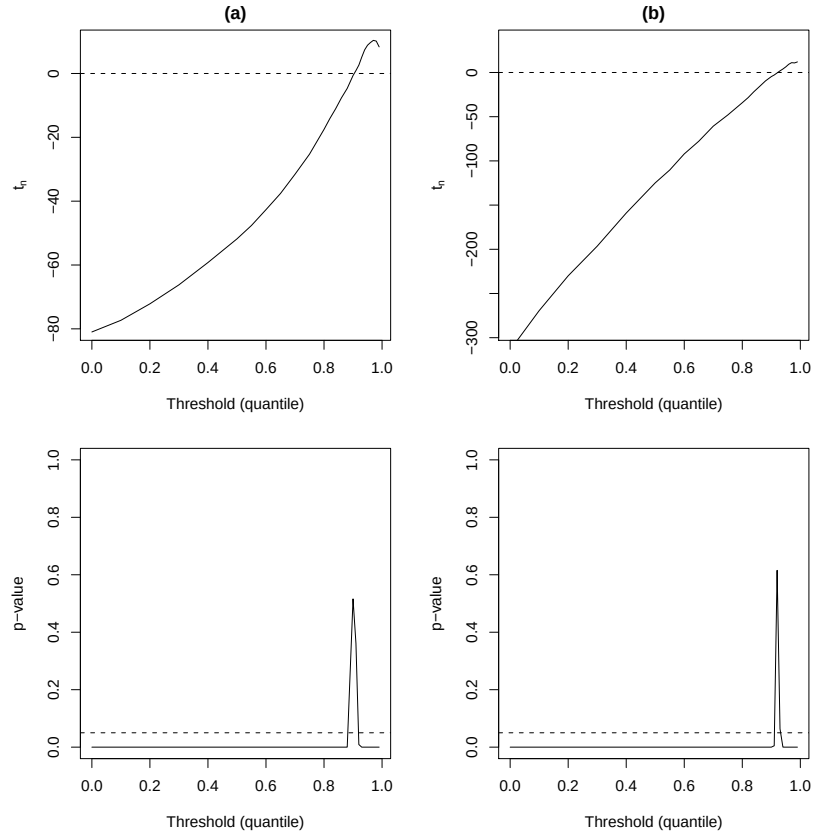

Figure 3.6.: Top row: Diebold-Mariano-type test statistic (3.1) for test of equal performance comparing the ideal and the biased forecaster as functions of the threshold which defines extreme events, using (a) the restricted CRPS and (b) the restricted LogS. Bottom row: Corresponding p-values under the standard normal hypothesis, the dashed lines indicate a 5% significance level.

For both restricted CRPS and restricted LogS, the Diebold-Mariano-type test statistics for the comparison of the ideal and the biased forecaster shown in Figure 3.6 are negative up to a threshold value of approximately the 90th percentile of the marginal distribution of the observations, which corresponds to a value of around 2, and positive for larger threshold values. The measured score differences are significant, except for a small interval of thresholds around the 90th percentile of the marginal distribution as indicated by the corresponding p-value plots. In particular, the tests of equal performance prefer the biased forecaster over the ideal forecaster for extreme events and the measured score differences are significant even for small significance levels. This observation again confirms our finding of the impropriety of the restricted CRPS and the restricted LogS.

3.3.3. Proper scoring rules for extreme events

Gneiting and Ranjan (2011b) and Diks et al. (2011) propose proper scoring rules for extreme events which were discussed in 2.3.2. We apply the threshold-weighted

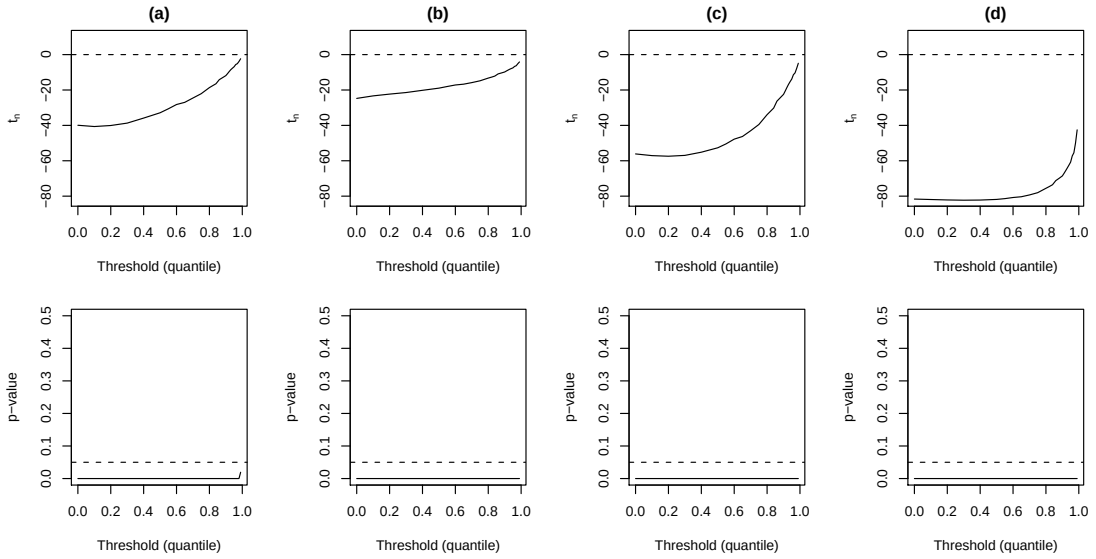

Figure 3.7.: Diebold-Mariano-type test statistics (3.1) for tests of equal performance using the threshold-weighted CRPS to compare the ideal forecaster with (a) the climatological, (b) the unfocused, (c) the sign-biased and (d) the biased forecaster as functions of the threshold r in (3.2). The second row shows corresponding p-values under the standard normal hypothesis, with dashed lines indicating a 5% significance level.

continuous ranked probability score,

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

the conditional likelihood scoring rule,

$$\mathrm{CL}(f, y) = -w(y) \log \left(\frac{f(y)}{\int w(s) f(s) \mathrm{d}s} \right),$$

and the censored likelihood scoring rule,

$$\mathrm{CSL}(f, y) = -\left[w(y) \log f(y) + (1 - w(y)) \log \left(1 - \int w(s) f(s) \mathrm{d}s \right) \right],$$

to the competing forecasting procedures and study effects of sample size, variance estimation and the choice of weight functions. At first, we will assume the weight function w to be an indicator function,

$$w_r(x) = \mathbb{1}(x \geq r), \tag{3.2}$$

where $r \in \mathbb{R}$ is a real-valued threshold. Later, we will study different weight functions.

Figure 3.7 shows plots of the test statistic t_n as a function of r, in terms of

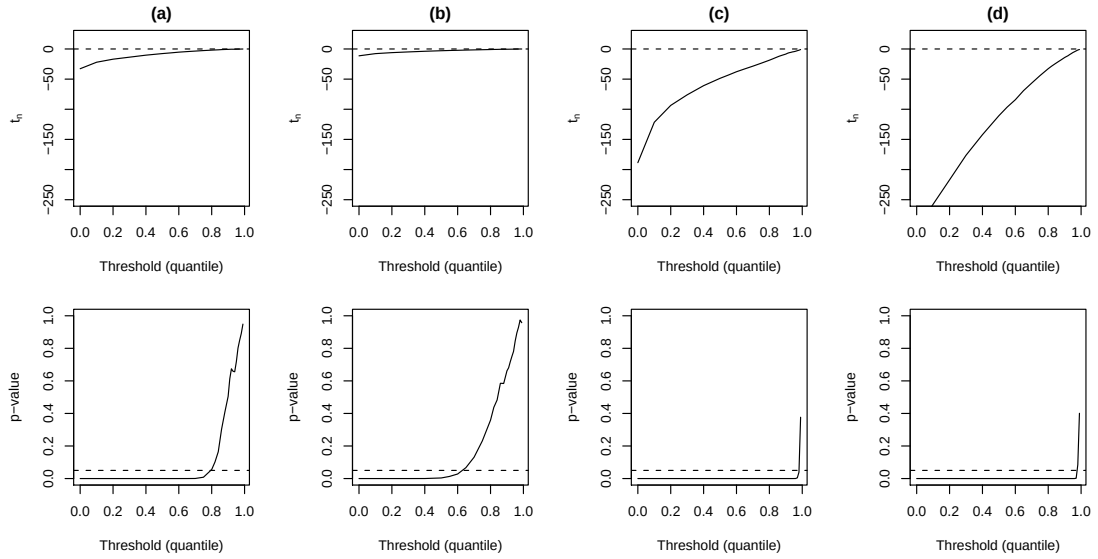

Figure 3.8.: Diebold-Mariano-type test statistics (3.1) for tests of equal performance using the conditional likelihood scoring rule to compare the ideal forecaster with (a) the climatological, (b) the unfocused, (c) the sign-biased and (d) the biased forecaster as functions of the threshold r in (3.2). The second row shows corresponding p-values under the standard normal hypothesis, with dashed lines indicating a 5% significance level.

quantiles of the marginal distribution of the observations, for the comparison of the ideal forecaster and the competing forecasting procedures using the threshold-weighted CRPS. In case of negative values of the test statistic the ideal forecaster is preferred over the competitor, for positive values the competitor is preferred over the ideal forecaster. The asymptotic variance of the score difference was estimated as proposed by Gneiting and Ranjan (2011b). We will investigate the effects of different variance estimation procedures below. In all four cases, the ideal forecaster is preferred over his competitor for all choices of r, all observed score differences are significant under the standard normal hypothesis. Note that here, the sample size equals T for any choice of r, which was not the case for the restricted scoring rules used in the previous section. Thus, the standard normal hypothesis is much more likely to hold in the case of larger values of r as well.

Figures 3.8 and 3.9 show analogous plots for the conditional and the censored likelihood scoring rule. Here, the asymptotic variance of the score differences was estimated as proposed by Diks et al. (2011). For the conditional likelihood scoring rule, all observed values of the test statistics are negative, however, the observed differences are much less significant under the standard normal hypothesis. The score differences between the ideal and the unfocused forecaster are only significant up to approximately $r = 0.5$ which corresponds to the 60th percentile of the marginal distribution of the observations. The conditional likelihood scoring rule is furthermore unable to significantly distinguish between the predictive performance of the ideal and the biased forecaster for thresholds larger than the

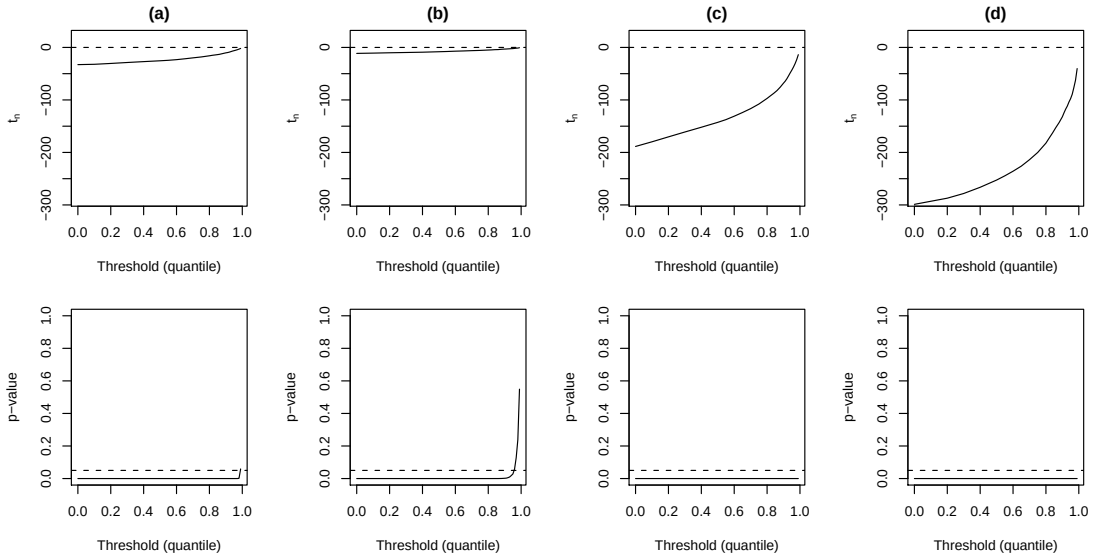

Figure 3.9.: Diebold-Mariano-type test statistics (3.1) for tests of equal performance using the censored likelihood scoring rule to compare the ideal forecaster with (a) the climatological, (b) the unfocused, (c) the sign-biased and (d) the biased forecaster as functions of the threshold r in (3.2). The second row shows corresponding p-values under the standard normal hypothesis, with dashed lines indicating a 5% significance level.

97th percentile of the marginal distribution of the observations. The results for the censored likelihood scoring rule are qualitatively equivalent to those for the threshold-weighted CRPS. However, in case of the unfocused forecaster, the score differences measured by the censored likelihood scoring rule are not significant for values of r larger than the 95th percentile of the marginal distribution of the observations. The corresponding score differences measured by the threshold-weighted CRPS are significant for all thresholds.

To summarize, the results of the simulation study are coherent with the theoretical results of Gneiting and Ranjan (2011b) and Diks et al. (2011) in that the three scoring rules indeed appear to be proper and are well able to distinguish between the predictive performance of the ideal forecaster and various competitors. However, for predictive distributions with similar tail behavior, the score differences measured by the conditional likelihood scoring rule are much less significant than the results for the other two scores. This holds, for example, for the comparison of the ideal and the unfocused forecaster. Note that Diks et al. (2011) suggest to only use the CL scoring rule if similar probabilities are assigned to the region of interest. However, that is the case here.

The remainder of this chapter is dedicated to a detailed comparison of the effects of variations of sample sizes, variance estimation procedures and weight functions on the ability of the different proper scoring rules to significantly distinguish between the ideal forecaster and his competitors.

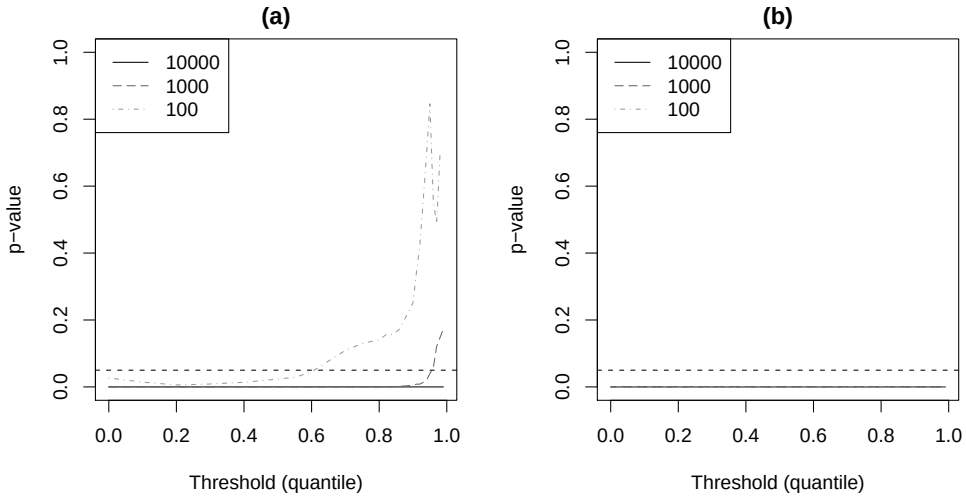

Figure 3.10.: P-values associated with the observed values of the Diebold-Mariano-type test statistics for tests of equal performance using the threshold-weighted CRPS to compare the ideal with (a) the unfocused and (b) the biased forecaster as functions of the threshold r, for different sample sizes. The dashed lines indicate a 5% significance level.

Sample size

To begin with effects of the sample size, we focus on tests of equal performance of ideal and unfocused, and ideal and biased forecaster. This seems to be a reasonable choice because on the one hand, the unfocused forecaster is only outperformed by the ideal forecaster and Figures 3.7, 3.8 and 3.9 indicate that the least significant score differences are measured between the ideal and the unfocused forecaster. On the other hand, we also want to compare the ideal and the biased forecaster in order to see if any disconcerting results as for the restricted CRPS and LogS occur.

We repeat the simulation study with $T = 1\,000$ and $T = 100$, the asymptotic variance of the score differences is estimated as above. For all choices of r and every proper scoring rule, the test statistics only attain strictly negative values in favor of the ideal forecaster. Thus, the proper scoring rules can also be used to distinguish predictive performance for smaller sample sizes. However, the measured skill differences are less significant for smaller sample sizes as indicated by the plots of p-values as functions of r in terms of quantiles of the marginal distribution of the observations, which are shown in Figures 3.10 (threshold-weighted CRPS), 3.11 (CL scoring rule) and 3.12 (CSL scoring rule).

The threshold-weighted CRPS differences between the ideal and the unfocused forecaster are significant under the standard normal hypothesis up the 60th percentile of the marginal distribution of the observations for sample size 100 and up to the 90th percentile for sample size 1 000. The threshold-weighted CRPS differences between the ideal and the biased forecaster are significant for all thresholds and sample sizes. If the CL scoring rule is used, basically none of the observed score differences between the ideal and the unfocused forecaster are significant for T = 100 and T = 1 000, even when comparing the ideal and the biased forecaster,

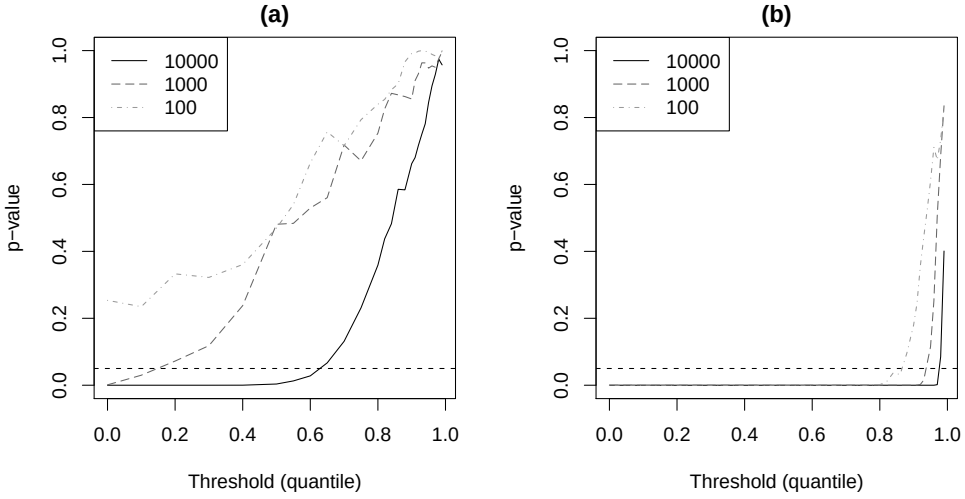

Figure 3.11.: P-values associated with the observed values of the Diebold-Mariano-type test statistics for tests of equal performance using the CL scoring rule to compare the ideal with (a) the unfocused and (b) the biased forecaster as functions of the threshold r, for different sample sizes. The dashed lines indicate a 5% significance level.

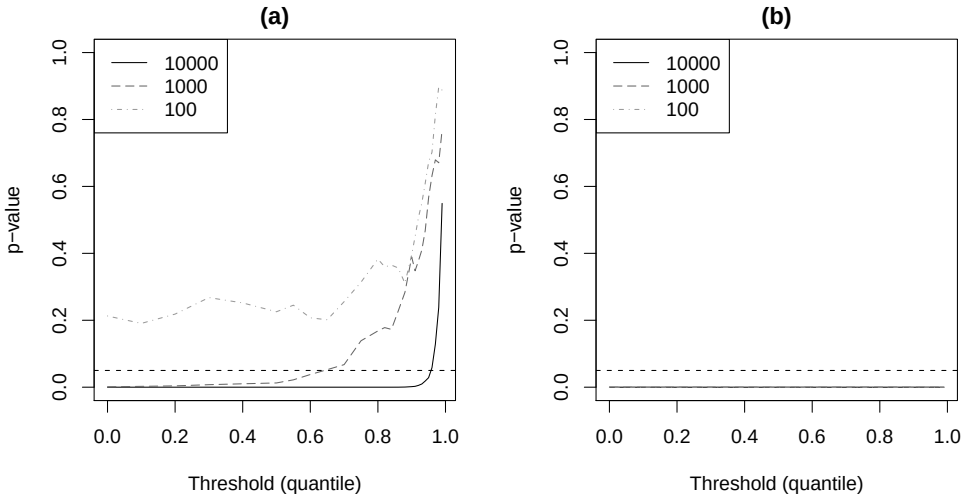

Figure 3.12.: P-values associated with the observed values of the Diebold-Mariano-type test statistics for tests of equal performance using the CSL scoring rule to compare the ideal with (a) the unfocused and (b) the biased forecaster as functions of the threshold r, for different sample sizes. The dashed lines indicate a 5% significance level.

the observed score differences are insignificant for thresholds larger than the 85th (sample size 100) or the 95th (sample size 1 000) percentile of the marginal distribution of the observations. The CSL scoring rule performs better in comparing the ideal and the biased forecaster, all observed score differences are significant. However, if the ideal and the unfocused forecaster are compared for smaller sample sizes, the score differences are not significant for all thresholds (sample size 100) or thresholds larger than 0.55 (sample size 1 000) which corresponds the to 65th percentile of the marginal distribution of the observations.

Thus, CL and CSL scoring rule seem to be much more prone to score differences being insignificant due to smaller sample sizes than the threshold-weighted CRPS. This effect might be caused by the differences in the variance estimation. If the variance estimation procedure proposed by Gneiting and Ranjan (2011b) is used for the CL and CSL scoring rule as well, the results for the CSL scoring rule are qualitatively equivalent to those for the threshold-weighted CRPS. The CL scoring rule, however, is still not able to significantly distinguish between the ideal and the unfocused forecaster for thresholds larger than the 70th percentile of the marginal distribution of the observations and furthermore fails to detect significant score differences for thresholds larger than the 90th percentile if the ideal and the biased forecaster are compared. In the following, variance estimation effects will be discussed in more detail.

Variance estimation

The approaches to forecast evaluation for extreme events of Gneiting and Ranjan (2011b) and Diks et al. (2011) differ not only in their choice of the scoring rule S, but also in how the asymptotic variance of the score differences is estimated. As discussed in Section 2.3.2, Gneiting and Ranjan (2011b) follow Diebold and Mariano (1995) by only taking into account auto-correlation up to a lag of at most $k-1$. Diks et al. (2011), on the contrary, take into account autocorrelation up to a lag larger than $k-1$.

Figures 3.13 and 3.14 illustrate the effect of using these two different variance estimation procedures. For larger differences in forecasting quality, as for example if the ideal and the biased forecaster are compared (Figure 3.13), the variance estimation procedure does not effect the significance of the observed score differences. Except for the CL scoring rule and very large thresholds, any combination of scoring rule and variance estimation procedure yields significant score differences for all thresholds. The ideal forecaster is always preferred over the biased forecaster.

However, if forecasters of similar predictive performance are compared, larger dissimilarities in the significance of the score differences can be observed. Figure 3.14 shows test statistics and corresponding p-values for the three scoring rules and the two different variance estimation procedures comparing the ideal and the unfocused forecaster. As expected, by taking into account autocorrelation up to a larger lag, the values of t_n are much smaller if the asymptotic variance is estimated as proposed by Diks et al. (2011). Although still preferring the ideal forecaster over the unfocused forecaster, the observed score differences thus are much less significant compared to those obtained by employing the variance estimation pro-

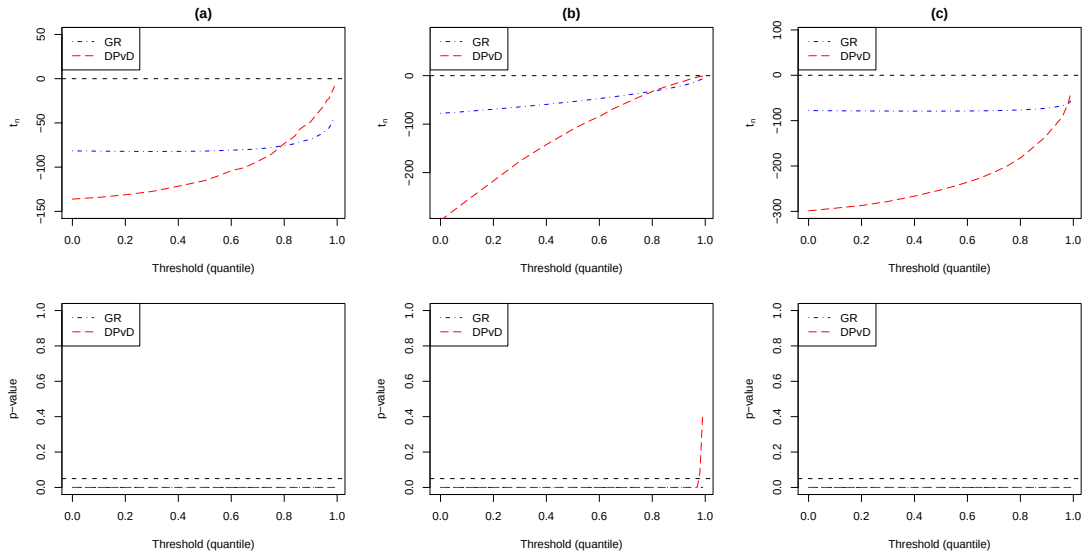

Figure 3.13.: Test statistics t_n for tests of equal performance using (a) the threshold-weighted CRPS, (b) CL and (c) CSL scoring rule to compare the ideal and the biased forecaster as functions of the threshold r. The asymptotic variance was estimated as proposed by Gneiting and Ranjan (2011b) (GR, blue) and Diks et al. (2011) (DPvD, red). The bottom row shows the corresponding p-values under the standard normal hypothesis. Note that for clarity of the presentation, the scale of the test statistics varies in the plots in the top row.

cedure proposed by Gneiting and Ranjan (2011b). For both the threshold-weighted CRPS and the CL scoring rule, the observed score differences are not significant for thresholds larger than approximately the 70th percentile of the marginal distribution of the observations if the HAC estimator of Diks et al. (2011) is used. Neither are the censored likelihood score differences for thresholds larger than the corresponding 95th percentile. As opposed to this, the score differences for all scoring rules are significant for all thresholds if the variance is estimated as proposed by Gneiting and Ranjan (2011b), except for the CL scoring rule and very large thresholds.

The estimator of the asymptotic variance proposed by Gneiting and Ranjan (2011b) is based on the theoretical assumption of Diebold and Mariano (1995) that optimal k-step-forecast errors are at most $(k-1)$-dependent. However, this assumption of $(k-1)$-dependence might be violated in practice due to various reasons. In our simulation study, the values of the scoring rules, which can be interpreted as forecast errors, are 0-dependent by construction. This can empirically be confirmed by computing estimates of the autocorrelation, as shown in Figure 3.15 for the ideal forecaster. Qualitatively equivalent results can be obtained for all other forecasting procedures. Thus, the approach to variance estimation used by Gneiting and Ranjan (2011b) seems to be a reasonable choice for our simulation study, while the approach proposed by Diks et al. (2011) takes into account unnecessary large amounts of autocorrelation up to lag 10.

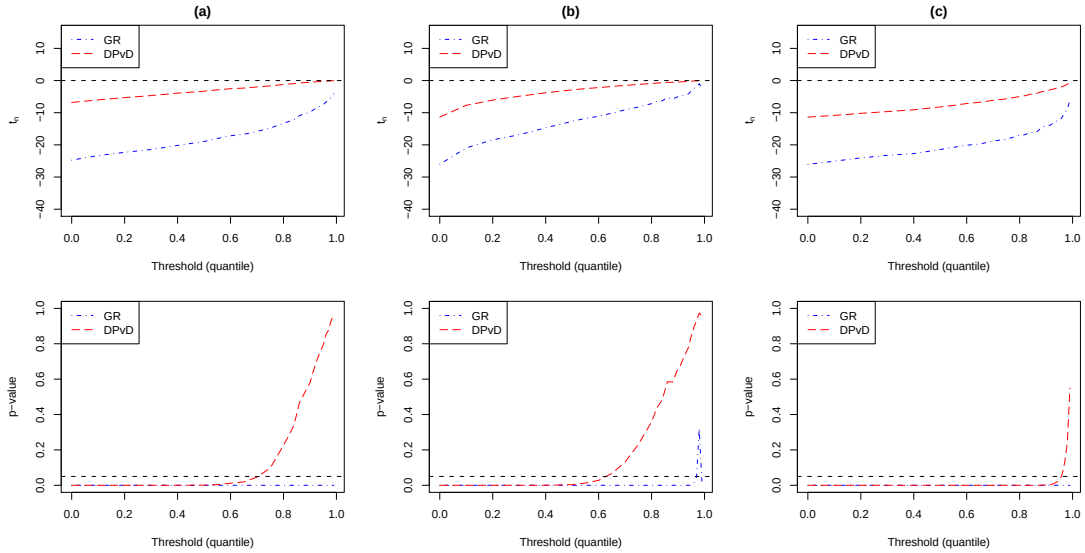

Figure 3.14.: Test statistics t_n for tests of equal performance using (a) the threshold-weighted CRPS, (b) CL and (c) CSL scoring rule to compare the ideal and the unfocused forecaster as functions of the threshold r. The asymptotic variance was estimated as proposed by Gneiting and Ranjan (2011b) (GR, blue) and Diks et al. (2011) (DPvD, red). The bottom row shows the corresponding p-values under the standard normal hypothesis.

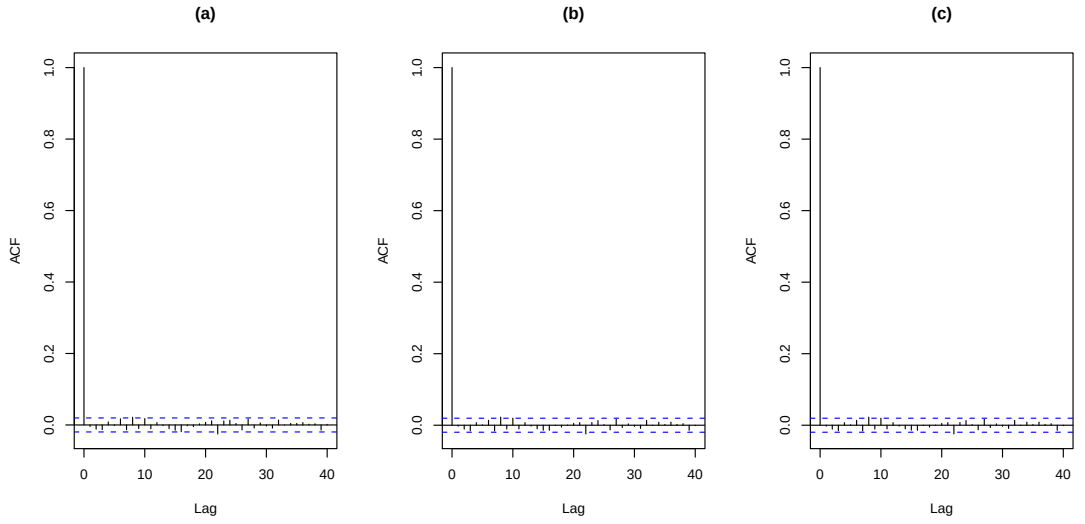

Figure 3.15.: Autocorrelation of the values of (a) threshold-weighted CRPS, (b) CL and (c) CSL scoring rule for the ideal forecaster.

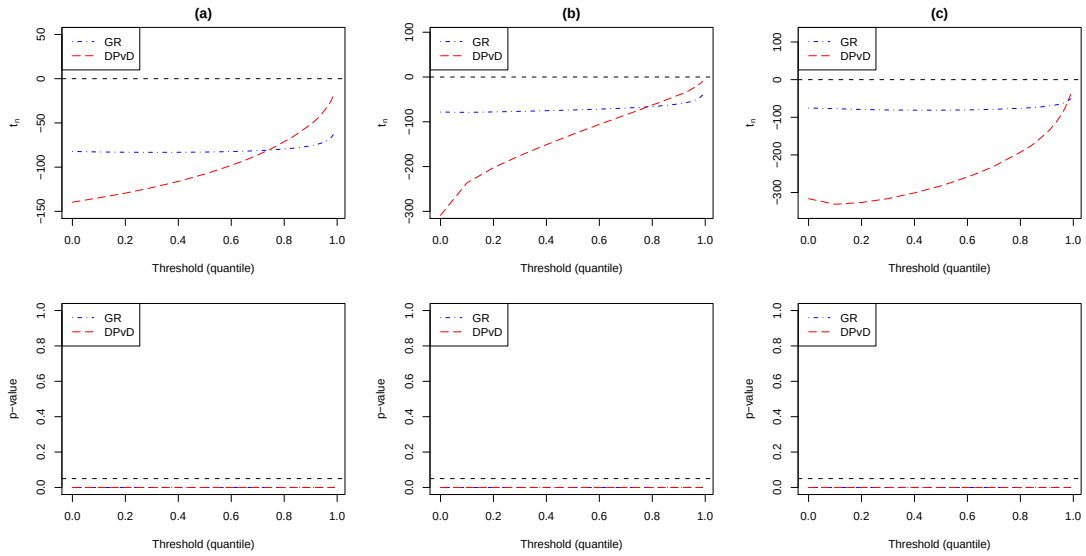

Figure 3.16.: Top row: Test statistics of the test of equal performance of the ideal and the biased forecaster as functions of the threshold r for (a) threshold-weighted CRPS, (b) CL and (c) CSL scoring rule using the normal CDF weight function (3.3). The asymptotic variance was estimated as proposed by Gneiting and Ranjan (2011b) (GR, blue) and Diks et al. (2011) (DPvD, red). Bottom row: Corresponding p-values under the standard normal hypothesis.

Weight functions

Thus far, we only used weight functions of the form

$$w_r(x) = \mathbb{1}(x \geq r)$$

with $r \in \mathbb{R}$ in order to emphasize extreme events. However, proper scoring rules based on this weight function are not able to distinguish between two competing density forecasts with identical tail behavior on $[r, \infty)$, but different behavior on $(-\infty, r)$. In order to emphasize the right tail of a distribution, Gneiting and Ranjan (2011b) propose the weight function

$$w(x) = \Phi_{a,b}(x),$$

where $\Phi_{a,b}$ denotes the CDF of a normal distribution with mean a and standard deviation b. In the remainder of this chapter, we investigate the use of this weight function in comparison to the indicator function used before. In order to obtain a similar threshold dependence, we set $a = r$ and $b = 1$ so that

$$w_r(x) = \Phi_{r,1}(x). \tag{3.3}$$

Figures 3.16 and 3.17 show the test statistics as functions of r in terms of quan-

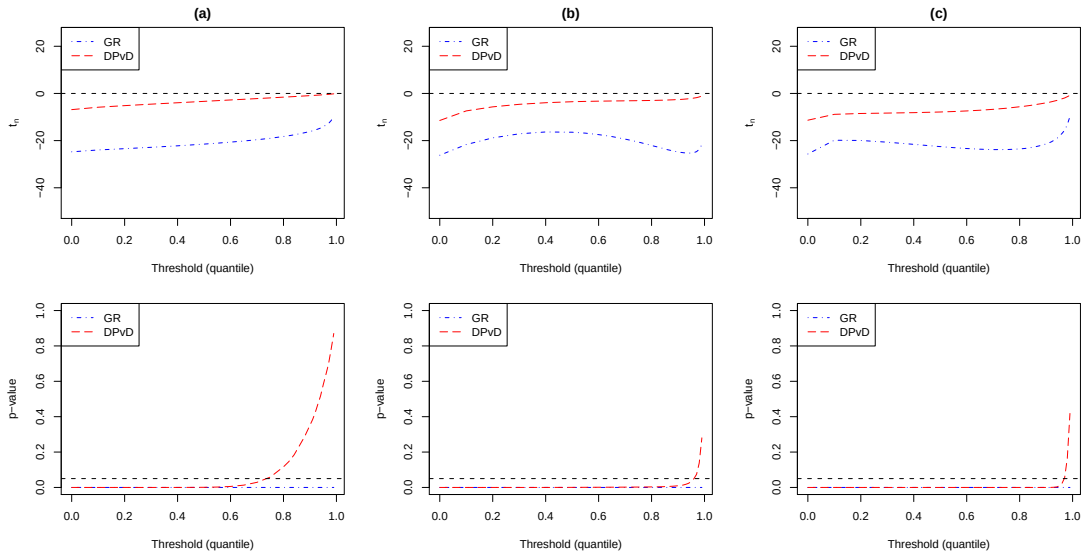

Figure 3.17.: Top row: Test statistics of the test of equal performance of the ideal and the unfocused forecaster as functions of the threshold r for (a) threshold-weighted CRPS, (b) CL and (c) CSL scoring rule using the normal CDF weight function (3.3). The asymptotic variance was estimated as proposed by Gneiting and Ranjan (2011b) (GR, blue) and Diks et al. (2011) (DPvD, red). Bottom row: Corresponding p-values under the standard normal hypothesis.

tiles of the marginal distribution of the observations using the normal CDF weight function as defined in (3.3) to compare the ideal and the biased forecaster (Figure 3.16), and the ideal and the unfocused forecaster (Figure 3.17). The results are equivalent to those obtained if the indicator weight function $w_r(x) = \mathbb{1}(x \geq r)$ is used. The ideal forecaster is preferred over the competitor for all choices of the threshold r. The observed score differences appear to be slightly more significant if the asymptotic variance is estimated as proposed by Diks et al. (2011) compared to the results for this estimation procedure and the indicator weight function. If the asymptotic variance is estimated as proposed by Gneiting and Ranjan (2011b), all observed score differences are significant for all scoring rules.

We can conclude that here, the choice of an indicator weight function $w_r(x) = \mathbb{1}(x \geq r)$ is admissible as well. However, if density forecasts with identical tail behavior, but different behavior on the remaining part of the real line are compared, the normal CDF weight function $w_r(x) = \Phi_{r,1}(x)$ should be used instead.

3.4. Summary

Our simulation study empirically confirms the theoretical results of the previous chapter and suggests that the discussed proper scoring rules for extreme events work well. In our simulation study, all proper scoring rules for extreme events prefer the ideal forecaster over the biased forecaster which is not the case if proper

scoring rules are restricted to subsets of extreme events. However, the observed score differences between the ideal forecaster and his competitors are not always significant.

In general, the score differences measured by the conditional likelihood scoring rule tend to be less significant if forecasters with similar predictive performance are compared. Diks et al. (2011) note that the CL scoring rule should only be used to compare forecasters who assign similar probabilities to the region of interest. However, we found that the CL scoring rule is unable to distinguish between the ideal and the unfocused forecaster despite the similarity of their marginal densities.

A comparison of the approaches of Gneiting and Ranjan (2011b) and Diks et al. (2011) indicates that the smaller the sample size, the less significant the observed score differences. The threshold-weighted CRPS seems to be less prone to insignificant score differences due to smaller sample sizes. However, this is mainly an effect of the different approaches to estimating the asymptotic variance of the score differences. The variance estimation procedure proposed by Diks et al. (2011) generally produces less significant score differences by taking into account autocorrelation up to a larger lag. The variance estimation procedure proposed by Gneiting and Ranjan (2011b) is based on the theoretical result of at most $(k-1)$-dependence of the forecast errors which we empirically confirmed here. Therefore, in this situation, the variance estimation procedure proposed by Gneiting and Ranjan (2011b) seems to be the better choice. Using this estimation procedure, all proper scoring rules yield significant score differences for nearly all thresholds. In the situation of our simulation study, different choices of the weight function to emphasize extreme events result in equivalent decisions.

Both the threshold-weighted CRPS and the censored likelihood scoring rule are well able to distinguish between the ideal forecaster and all competitors. The variance estimation procedure should be chosen dependent on the amount of autocorrelation empirically found in the vector of score differences.

4. Case study

4.1. Introduction

In an application to wind gust forecasting, we investigate the forecaster's dilemma and the proper scoring rules for extreme events proposed by Gneiting and Ranjan (2011b) and Diks et al. (2011) in a real-world example. A wind gust is a sudden, brief increase in wind speed during a specific time interval. The exact technical definition varies dependent on the observation system at hand. Accurate forecasts of wind speeds are necessary in many applications such as agriculture, aviation or energy production. In particular, accurate and reliable forecasts of wind gust are of importance for minimizing damages and human losses in high-impact extreme weather events. Wind gusts furthermore influence the design of buildings and bridges (Friederichs et al., 2009; Thorarinsdottir and Johnson, 2012).

However, wind gust observations are sparse both on the spatial and the temporal scale, and are not a standard output from numerical weather prediction (NWP) models. Our case study is based on the work of Thorarinsdottir and Johnson (2012) who propose a forecasting framework for gust speed when maximum wind speed forecasts are available. In a first step, an ensemble forecast utilizing multiple runs of an NWP model with different initial conditions and numerical representation of the atmosphere is obtained (Gneiting and Raftery, 2005). Calibrated and sharp probabilistic 48-hour-ahead forecasts of maximum gust speed are then obtained using the ensemble predictions of maximum wind speeds, gust factors, and nonhomogeneous Gaussian regression (NGR) (Thorarinsdottir and Gneiting, 2010).

We propose competing forecasting procedures with the objective of accurate predictions of extreme gust speeds by employing generalized extreme value (GEV) and generalized Pareto (GP) distributions and investigate the forecast evaluation procedures discussed in Chapter 2. In addition, we propose an improvement of the NGR forecasting procedure for wind speed of Thorarinsdottir and Gneiting (2010) where we use a regime-switching forecasting procedure dependent on the median ensemble forecast.

4.2. Gust speed forecasting

4.2.1. Nonhomogeneous Gaussian Regression

The gust speed forecasting procedure proposed by Thorarinsdottir and Johnson (2012) consists of three parts. Based on ensemble postprocessing procedures leading to probabilistic forecasts of daily maximum wind speed, and estimates of the

probability of gust, probabilistic forecasts of maximum gust speed conditional on gust being observed are obtained.

Predictive distributions of wind speed can be modeled as a mixture of gamma distributions using Bayesian model averaging (BMA) as proposed by Sloughter et al. (2010) or as normal distributions truncated at 0 using nonhomogeneous Gaussian regression (NGR) as proposed by Thorarinsdottir and Gneiting (2010). Both methods provide calibrated and sharp probabilistic wind speed forecasts. For the purpose of gust speed forecasting, the NGR approach is more appealing if the assumption of a multiplicative relationship between wind speed and gust speed is made. Following the meteorological tradition of estimating gust speed by multiplying wind speed forecasts with gust factors, predictive truncated normal distributions for gust speed can immediately be obtained since the family of truncated normal distributions is closed under affine transformations.

We will now briefly summarize the NGR approach to gust speed forecasting proposed by Thorarinsdottir and Johnson (2012). Let Y denote wind speed and let $X_1, \ldots, X_k$ denote distinguishable ensemble member forecasts for Y. Following the NGR approach of Thorarinsdottir and Gneiting (2010), the predictive distribution for wind speed is modeled by a truncated normal distribution with a cutoff at 0,

$$Y|X_1, \ldots, X_k \sim \mathcal{N}_{[0,\infty)}(\mu, \sigma^2),$$

where the location parameter $\mu = a + b_1 X_1 + \cdots + b_k X_k$ is an affine function of the ensemble forecasts and the variance $\sigma^2 = c + dS^2$ is an affine function of the ensemble variance $S^2 = \frac{1}{k}\sum_{i=1}^{k}(X_i - \bar{X})^2$. The density of this truncated normal distribution is given by

$$f(y) = \frac{1}{\Phi\left(\frac{\mu}{\sigma}\right)}\left[\frac{1}{\sigma}\varphi\left(\frac{y-\mu}{\sigma}\right)\right]$$

for $y > 0$ and 0 otherwise, where φ and Φ denote the density and the cumulative distribution function of the standard normal distribution, respectively.

The parameters $a, b_1, \ldots, b_k, c, d$ can be estimated using minimum CRPS estimation over a rolling training period which consists of the observations of the last m days. The parameters are estimated regionally in that training data from all stations are pooled together. Under the assumption of a multiplicative relationship between the wind speed Y and the gust speed Z,

$$Z = \gamma Y$$

for a gust factor $\gamma \geq 1$, we obtain a predictive distribution

$$Z|X_1, \ldots, X_k \sim \mathcal{N}_{[0,\infty)}(\gamma\mu, [\gamma\sigma]^2).$$

The estimation of gust factors and maximum gust speed depends on the definition of wind gust. Thorarinsdottir and Johnson (2012) use data from the North American Automated Surface Observation System (ASOS) network, where only gust speeds of at least 14 knots (1 kt = 0.514 $\mathrm{m\,s^{-1}}$) are reported. Two separate

gust factors for the probability of gust and the gust speed forecasts conditional on gust being observed are estimated. The probability of gust being observed is

$$\mathbb{P}(Z \geq 14|X_1, \ldots, X_k) = \frac{1}{\Phi\left(\frac{\mu}{\sigma}\right)} \left[1 - \Phi\left(\frac{14 - \gamma_1\mu}{\gamma_1\sigma}\right)\right].$$

Given NGR parameter estimates for μ and σ, the parameter γ_1 is estimated by minimizing the Brier score (Brier, 1950),

$$\frac{1}{J}\sum_{j=1}^{J}\left[\mathbb{P}(Z_j \geq 14|X_1^j, \ldots, X_k^j) - \mathbb{1}\{Z_j \text{ observed}\}\right]^2,$$

where the sum extends over all forecast cases in the training set.

The predictive distribution of gust speed conditional on gust being observed and given NGR parameter estimates for μ and σ is $\mathcal{N}_{[14,\infty)}(\gamma_2\mu, [\gamma_2\sigma]^2)$. The parameter γ_2 can be estimated using minimum CRPS estimation over the gust speed observations in the training set. Here, the CRPS becomes

$$\begin{aligned}\text{CRPS}(\mathcal{N}_{[14,\infty)}(\alpha,\beta), z) = \beta\lambda(\alpha,\beta)^{-2}\Bigg\{&\frac{z-\alpha}{\beta}\lambda(\alpha,\beta)\left[2\Phi\left(\frac{z-\alpha}{\beta}\right) + \lambda(\alpha,\beta) - 2\right]\\ &+ 2\varphi\left(\frac{z-\alpha}{\beta}\right)\lambda(\alpha,\beta) - \frac{1}{\sqrt{\pi}}\left[1 - \Phi\left(\sqrt{2}\frac{14-\alpha}{\beta}\right)\right]\Bigg\},\end{aligned}$$

where $\lambda(\alpha,\beta) = 1 - \Phi[(14-\alpha)/\beta]$ and $z \geq 14$. Thus, given NGR parameter estimates for wind speed, only the estimation of gust factors is required and the additional computational costs are very low.

4.2.2. Forecasting procedures based on extreme value theory

We will now use theoretical results from extreme value theory (EVT), the branch of probability theory and statistical science that is concerned with the modeling of extreme events (Coles, 2001), to obtain accurate and calibrated forecasts for subsets of extreme gust speeds. EVT uses subsets of large sample values to infer the extremal behavior of the underlying data-generating process (Coelho et al., 2008). Large values can be selected in various ways, for example as block maxima or as values exceeding a high threshold. Here, we fit parameters of generalized extreme value (GEV) and generalized Pareto (GP) distributions to extreme events exceeding a high threshold to obtain probabilistic forecasts of gust speed. In the following, we will only discuss predictive distributions of gust speed conditional on gust being observed. The probability of gust being observed can be estimated by minimizing the Brier score as discussed in Section 4.2.1.

Generalized extreme value distribution

The *Fisher-Tippett theorem* or *extreme value theorem* describes asymptotic distributions of extreme order statistics $M_n = \max\{X_1, \ldots, X_n\}$. It was named after

Fisher and Tippett (1928), for further details see Reiss and Thomas (2007). Three types of limits, known as the *Gumbel*, *Fréchet* and *Weibull* family, arise as limit distribution $F(z)$,

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq z\right) \longrightarrow F(z), \tag{4.1}$$

if suitable sequences of real-valued constants $a_n > 0$ and b_n exist. These families can be combined into a single family of models having distribution functions of the form

$$F_{GEV}(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\},$$

defined on the set $\{z \in \mathbb{R} : 1+\xi(z-\mu)/\sigma > 0\}$, where the parameters satisfy $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. F_{GEV} is the cumulative distribution function of the *generalized extreme value (GEV) distribution* with location parameter μ, scale parameter σ and shape parameter ξ (Coles, 2001). The cases $\xi > 0$ and $\xi < 0$ correspond to the Fréchet family and to the Weibull family in (4.1), respectively. The limit $\xi \to 0$ leads to the Gumbel family with CDF

$$F_{\text{Gumbel}}(z) = \exp\left[-\exp\left\{-\left(\frac{z-\mu}{\sigma}\right)\right\}\right],$$

where $-\infty < z < \infty$ (Coles, 2001).

GEV distributions can be used to model extreme values obtained as block maxima. The parameters of GEV distributions will be estimated over various training sets of extreme gust speed observations, using maximum likelihood (ML) estimation or minimum CRPS (minCRPS) estimation.

Maximum likelihood estimation corresponds to minimizing the mean logarithmic score over the training set. For the GEV distribution, there is no analytical solution (Coles, 2001), but numerical approximations can be obtained using standard algorithms (Hosking, 1985; Smith, 1985). Here, we apply the algorithm provided by the R package `ismev`.

Minimum CRPS parameter estimates are obtained by minimizing the mean CRPS over the training set. Friederichs and Thorarinsdottir (2012) prove a closed-form expression of the CRPS for the GEV distribution with shape parameter $\xi < 1$. For $\xi \neq 0$, the CRPS is given by

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

where $\Gamma(\cdot)$ denotes the gamma function and $\Gamma_l(\cdot)$ denotes the lower incomplete gamma function. For $\xi = 0$, the CRPS is given by

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

where $C \approx 0.5772$ is the Euler-Mascheroni constant and $Ei(x) = \int_{-\infty}^{x} \frac{e^t}{t}\mathrm{d}t$. The

minimization is carried out using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm as implemented in R (R Development Core Team, 2010). The BFGS algorithm is a quasi-Newton method, for further details see Bertsekas (1995). Due to reasons of numerical stability, values of the shape parameter ξ between -0.01 and 0.01 are rounded to 0. Since the closed form expression of the CRPS is only valid for $\xi < 1$, we furthermore restrict the shape parameter to the interval $(-\infty, 1)$. However, note that none of the corresponding ML estimates of the shape parameter takes a value larger than 1 for all forecasting procedures and training sets.

For the parameter estimation, three different kinds of training sets consisting of extreme gust speed observations are employed. The parameters can be assumed to be constant in terms of location and time, which is similar to standard climatological forecasting of marginal distributions. In this case, all gust speed observations from the years 2006 and 2007 which are at least 29 kt, the 90th percentile of the marginal distribution of the gust speed observations, are used to estimate the parameters of a GEV distribution. Due to the large size of the training set, numerical minimization of the CRPS is associated with large computational costs. Therefore, we use only maximum likelihood estimation for this training set.

As for the NGR method, the parameters can be estimated regionally using a rolling training period of 20 days. All observations of gust speed larger or equal to 29 kt during the last 20 days are pooled together and used to estimate the parameters of a GEV distribution which is issued as a probabilistic 48-h-ahead forecast. Both maximum likelihood and minimum CRPS estimation are applied to these training sets. Rolling training periods of different lengths were used as well but did not result in significant changes of the predictive performance which is in accordance with the results of Thorarinsdottir and Gneiting (2010).

Furthermore, the parameters can be estimated locally for each station. Due to the sparsity of extreme gust speeds, all gust speed observations larger or equal to 29 kt from the years 2006 and 2007 at the corresponding station were used to estimate the parameters. Here, both ML estimation and minimum CRPS estimation were applied as well.

The forecasting procedures are summarized in Table 4.1.

Generalized Pareto distribution

The *Pickands-Balkema-de Haan theorem* states that the tail of a distribution function can be well approximated by a generalized Pareto distribution for a large class of distributions. It was named after Balkema and de Haan (1974) and Pickands (1975), for further details see Reiss and Thomas (2007).

More precisely, the theorem states that if extreme events are regarded as those events that exceed some high threshold μ, the behavior of extreme events is given by the conditional excess probability

$$\mathbb{P}(Y > \mu + y | Y > \mu) = \frac{1 - F(\mu + y)}{1 - F(\mu)}, \qquad y > 0,$$

where Y is a random variable with CDF F (Coles, 2001). For large enough μ, the

Table 4.1.: Model choices, training sets and parameter estimation procedures used in this case study.

Abbreviation	Training set	Parameter estimates
GEV1	2006/7	Constant in location and time
GEV2ML	last 20 days	Time-varying, constant in location
GEV2minCRPS	last 20 days	Time-varying, constant in location
GEV3ML	2006/7	Varying over stations, constant in time
GEV3minCRPS	2006/7	Varying over stations, constant in time
GP1ML	2006/7	Constant in location and time
GP2ML	last 50 days	Time-varying, constant in location
GP3ML	2006/7	Varying over stations, constant in time

distribution function of Y conditional on $Y > \mu$ is approximately

$$F_{GP}(y) = \begin{cases} 1 - \left(1 + \frac{\xi(y-\mu)}{\sigma}\right)^{-1/\xi} & \text{for } \xi \neq 0, \\ 1 - \exp\left(-\frac{y-\mu}{\sigma}\right) & \text{for } \xi = 0, \end{cases}$$

defined on $\{y : y \geq \mu \text{ when } \xi \geq 0 \text{ and } \mu \leq y \leq \mu - \sigma/\xi \text{ when } \xi < 0\}$. F_{GP} defines the generalized Pareto distribution with location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma > 0$ and shape parameter $\xi \in \mathbb{R}$ (Embrechts et al., 1997).

Generalized Pareto distribution and generalized extreme value distribution are closely related. They differ in their characterization of extreme events as excesses over a large threshold or as block maxima, but it can be shown that if block maxima have approximating distribution F_{GEV}, then threshold excesses have a corresponding approximate distribution F_{GP} within the generalized Pareto family with the same shape parameter ξ (Coles, 2001). Therefore, we expect similar results for GEV and GP forecasting procedures.

The parameters of the GP distributions are estimated using maximum likelihood estimation over the same training sets as before. As for the GEV parameter estimation, analytical maximization of the log-likelihood is not possible and we use the numerical approximation algorithm provided by the R package `ismev`. For further details, see Smith (1985). Friederichs and Thorarinsdottir (2012) also prove a closed-form expression of the CRPS for the generalized Pareto distribution. However, minimum CRPS estimation suffered from numerical instability problems for our case study. Therefore, we will restrict our attention to maximum likelihood parameter estimates. This will suffice in order to demonstrate the forecaster's dilemma due to the similar predictive performance of GEV and GP distributions and maximum likelihood and minimum CRPS estimation. For the GPD2ML forecasting procedure, the training period had to be set to 50 days due to reasons of numerical stability. The various training sets and parameter estimation procedures are summarized in Table 4.1.

Figure 4.1.: Figure 2 of Thorarinsdottir and Johnson (2012, page 893). Locations of the 83 ASOS stations over the North American Pacific Northwest, including the Canadian provinces of British Columbia (BC) and Alberta (AB), and the U.S. states of Washington (WA), Oregon (OR), Idaho (ID), California (CA), and Nevada (NV).

4.3. Data

We use forecast and observation data from 83 stations of the North American Automated Surface Observation System (ASOS) network over the North American Pacific Northwest. The locations of the stations are shown in Figure 4.1.

At these stations, a 5-s average wind speed is stored in memory for 10 minutes and the maximum in memory is reported as gust if it is greater than the current 2-min average wind speed by at least 3 knots, the current wind speed is more than 2 kt, and the maximum gust in memory is at least 10 kt greater than the minimum 5-s average wind speed over the last 10 min (Thorarinsdottir and Johnson, 2012). Only gust speeds of at least 14 kt are reported. All observations are rounded to the nearest whole knot.

The wind speed forecasts are 48-h-ahead forecasts of maximum wind speed obtained from the University of Washington Mesoscale Ensemble (UWME) system using bilinear interpolation (Eckel and Mass, 2005). As Thorarinsdottir and Johnson (2012), we use data from 1 January to 31 December 2008. Observations are available for 291 days and a total of 22 863 individual forecast cases. Gust was observed in 8 324 forecast cases (36% of all forecast cases).

4.4. Results

4.4.1. Results for all events

Calibration

To assess probabilistic calibration, PIT histograms for the NGR model and the GEV2minCRPS forecasting procedure are shown in Figure 4.2 together with a ver-

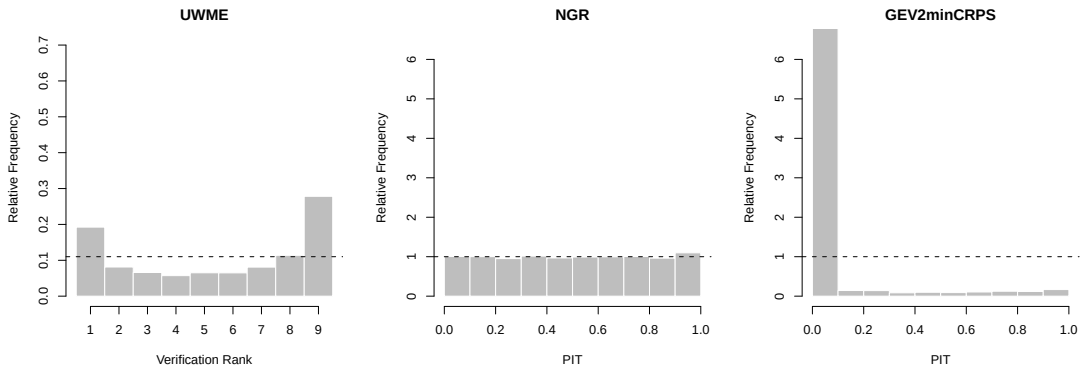

Figure 4.2.: Verification rank histogram for the ensemble and PIT histograms for the NGR and the GEV2minCRPS forecasting procedure.

ification rank histogram for the ensemble predictions. *Verification rank histograms* or *Talagrand diagrams* have been proposed, among others, by Anderson (1996) and Hamill and Colucci (1997). They are histograms of the rank of the observations when pooled within the ordered ensemble predictions (Gneiting et al., 2007). PIT histograms can be seen as continuous analogs of verification rank histograms.

The verification rank histogram indicates underdispersive UMWE predictions, too many observations fall outside the ensemble range. The NGR forecasting procedure significantly improves the ensemble forecasts. The PIT histogram shows only minor deviations from uniformity and the NGR forecasts appear to be probabilistically calibrated. The PIT histogram of the GEV2minCRPS forecasting procedure basically consists of only one bar which indicates the bias arising from the construction of the training sets used for the parameter estimation. PIT histograms for the other EVT forecasting procedures look almost identical with none of these forecasting procedures being probabilistically calibrated.

For the purpose of assessing marginal calibration, it is possible to compare the empirical density of the gust speed observations with marginal predictive densities of the different forecasting procedures as shown in Figure 4.3. Marginal predictive densities are obtained by averaging over all density forecasts over all days and stations in 2008 where gust was observed. While the NGR method appears to be marginally calibrated, the GEV2minCRPS marginal predictive density strongly differs from the empirical density of observed gust speed. With almost identical marginal predictive densities, none of the GEV and GP forecasting procedures seems to be marginally calibrated.

Summary measures

Table 4.2 summarizes values of various summary measures of predictive performance for the competing forecasting procedures. Mean CRPS, MAE and average width of 77.8% prediction intervals are given in kt. This specific prediction interval was chosen because it corresponds to the probability that the observation falls within the range of a perfectly calibrated ensemble. The MAE is the mean absolute error of point forecasts given by the median of the corresponding predictive

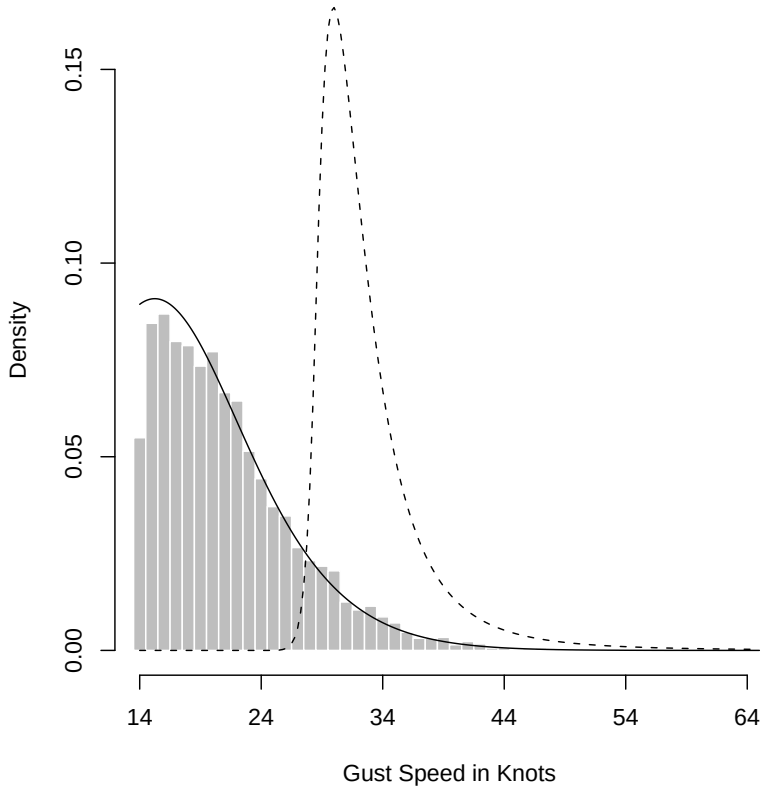

Figure 4.3.: Histogram of observed gust speed and marginal predictive densities under the NGR (solid line) and the GEV2minCRPS (dashed line) model.

Table 4.2.: Mean CRPS, MAE, average coverage (in %) and width of 77.8% prediction intervals for various probabilistic gust speed forecasts.

Forecast	CRPS	MAE	Coverage	Width
Climatology	3.08	4.33	81.6	12.81
UWME with gust factors	3.72	4.68	52.9	**8.61**
NGR	**2.56**	**3.60**	**76.9**	10.50
GEV1	10.35	11.16	8.5	10.00
GEV2ML	10.16	10.87	7.9	9.07
GEV2minCRPS	9.98	10.92	8.5	8.51
GEV3ML	10.40	11.22	8.1	9.59
GEV3minCRPS	10.32	11.38	8.2	8.87
GP1ML	10.36	12.00	8.9	11.62
GP2ML	10.68	11.63	8.5	10.09
GP3ML	11.29	12.64	7.4	9.24

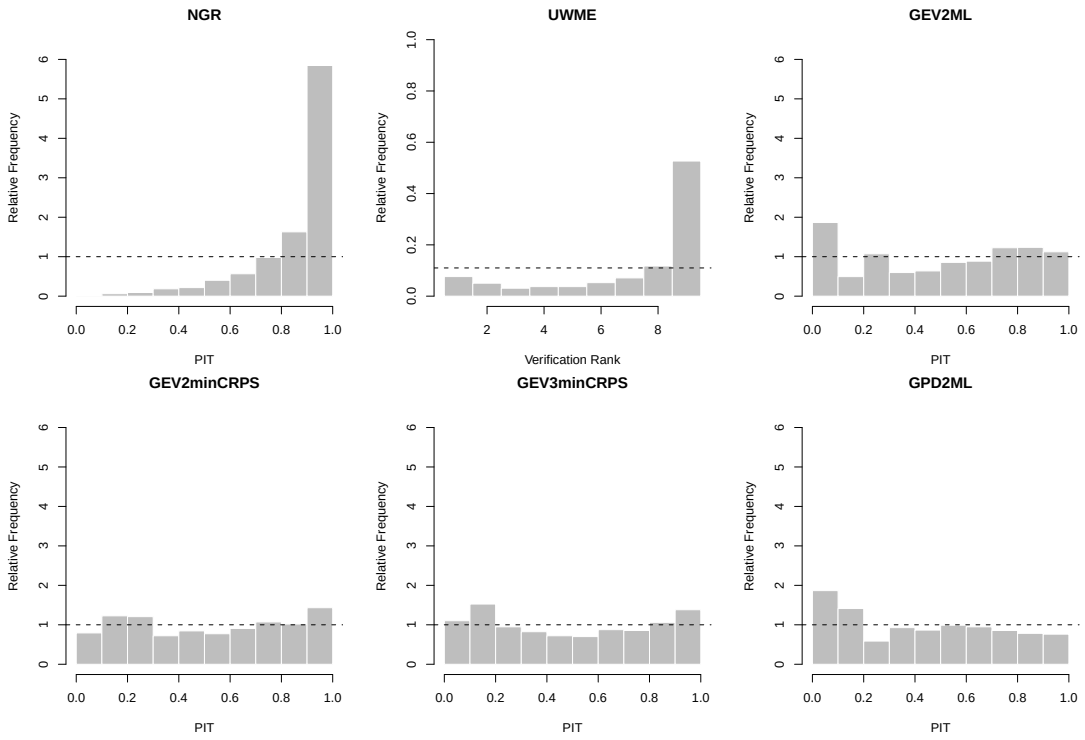

Figure 4.4.: Verification rank histogram for the ensemble and PIT histograms for various forecasting procedure based on the gust speed observations of at least 29 kt.

distribution. For densities that are not strictly positive, for example the predictive generalized Pareto densities, the logarithmic score does not attain finite values. Therefore, no values of the logarithmic score were computed.

The NGR method significantly improves the ensemble predictions and clearly outperforms any GEV or GP forecasting procedure. All forecasting procedures based on EVT exhibit poor predictive performance. On the contrary, the NGR method provides calibrated, sharp and accurate predictions of gust speed. However, this changes dramatically if only subsets of extreme events are regarded.

4.4.2. Results for subsets of extreme events

Calibration

We restrict our attention to subsets of extreme events by selecting all gust speed observations of at least 29 kt, which corresponds to the 90th percentile of the marginal distribution of gust speeds, and discarding the rest. Figure 4.4 shows the verification rank histogram for the ensemble predictions and PIT histograms for selected forecasting procedures based on these extreme events.

For this subset of extreme events, neither the ensemble nor the NGR predictions are probabilistically calibrated. The corresponding rank verification histogram and PIT histogram indicate an apparent bias. The PIT histograms of the forecasting procedures based on EVT exhibit smaller deviations from uniformity. How-

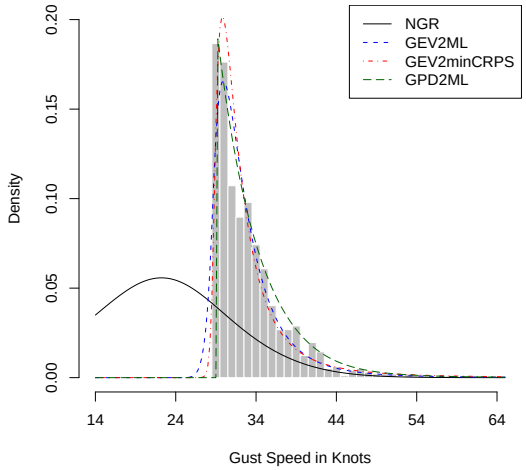

Figure 4.5.: Histogram of observed extreme gust speed and marginal predictive densities under the NGR and competing EVT models.

ever, none of them appears to be perfectly probabilistically calibrated, but both GEV2minCRPS and GEV3minCRPS perform reasonably well. Note that the sample size of 970 differs from the number of gust speed observations and thus, due to random effects, larger deviations from uniformity are to be expected even for perfectly calibrated forecasting procedures.

Figure 4.5 shows a histogram of the extreme gust speed observations together with marginal predictive densities of different forecasting procedures. The displayed marginal predictive densities of all forecasting procedures based on EVT seem to provide good fits to the marginal density of extreme gust speed observations while there is no concordance to the marginal NGR predictive density. Thus, unlike the NGR forecasting procedure, all forecasting procedures based on EVT appear to provide marginally calibrated probabilistic forecasts for this subset of extreme events.

Summary Measures

Table 4.3 summarizes values of the summary measures discussed before, now restricted to the subset of extreme events. Neither the gust speed climatology (based on all gust speed observations), nor the ensemble or the NGR forecasting procedure perform well in predicting these extreme events.

On the contrary, all forecasting procedures based on EVT provide accurate and sharp probabilistic forecasts for these events. Only minor differences in their predictive performances can be observed. Since extreme observations are reported as those observations larger than a high threshold and since the parameters of the GEV and GP models are estimated using training sets consisting of observations exceeding the same threshold, the Pickands-Balkema-de Haan theorem suggest that the GP models should outperform the GEV models. However, despite these

Table 4.3.: Mean CRPS, MAE, average coverage (in %) and width of 77.8% prediction intervals for various probabilistic gust speed forecasts and gust speed observations of at least 29 kt.

Forecast	CRPS	MAE	Coverage	Width
Climatology	10.10	13.79	9.7	13.21
UWME with gust factors	6.36	7.60	39.7	9.88
NGR	6.44	8.82	39.2	13.22
GEV1	**2.34**	3.26	72.9	10.00
GEV2ML	2.41	**3.23**	68.0	8.96
GEV2minCRPS	2.40	3.31	72.6	**8.54**
GEV3ML	2.40	3.31	69.2	9.51
GEV3minCRPS	2.46	3.40	70.2	8.74
GP1ML	2.43	3.35	**76.1**	11.62
GP2ML	2.37	3.29	73.1	10.35
GP3ML	2.58	3.61	63.7	8.74

theoretical results, the GEV models appear to provide better calibrated and more accurate forecasts than GP models. Furthermore, minimum CRPS parameter estimation appears to produce slightly better calibrated and sharper, but less accurate probabilistic forecasts than ML parameter estimation.

To summarize, if the forecast evaluation was only based on extreme gust speed observations, any of the forecasting procedures based on EVT would be preferred over the NGR forecasting procedure although the latter provides much more accurate and calibrated forecasts of gust speed. Again, the forecaster's dilemma can be observed: Forecast evaluation only based on subsets of extreme events discredits skillful and calibrated forecasting procedures such as the NGR forecasting procedure.

Proper scoring rules restricted to subsets of extreme events

Table 4.3 shows values of the CRPS and the MAE if the 90th percentile of the marginal distribution of gust speed observations is used as threshold defining extreme events. As for the simulation study discussed in Chapter 3, it is possible to plot the summary measures as functions of the threshold which defines extreme events.

Figure 4.6 shows the mean restricted CRPS as a function of this threshold. The average values of the restricted CRPS for the ensemble, the climatological forecaster and the NGR forecasting procedure increase for subsets of more and more extreme events. On the contrary, the values for the GEV and GP forecasting procedures decrease for larger thresholds and increase again for thresholds close to the 99th percentile of the marginal distribution of the gust speed observations. For thresholds larger than 24 knots, which approximately corresponds to the 70th percentile of the marginal distribution, all forecasting procedures based on EVT

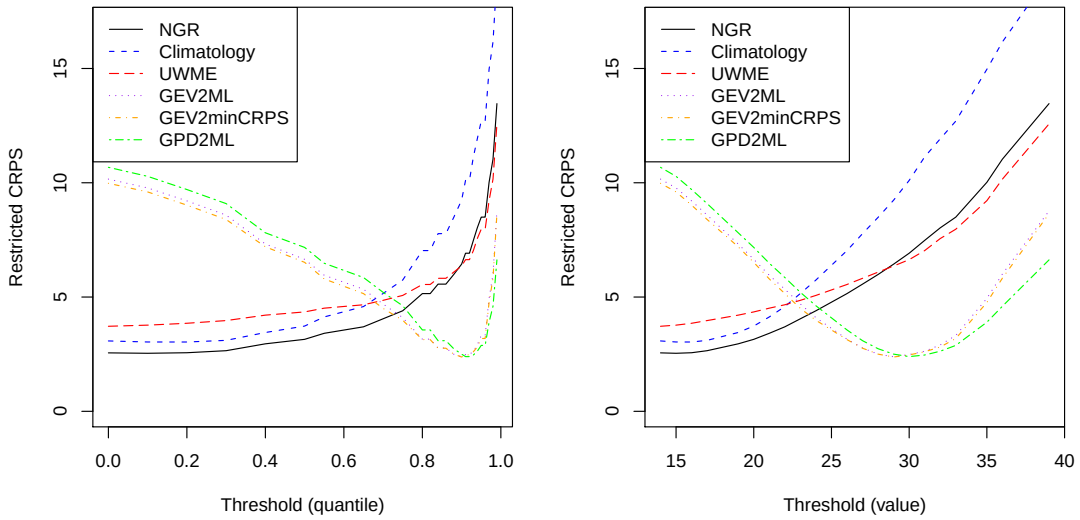

Figure 4.6.: CRPS restricted to subsets of extreme events as function of the threshold which defines extreme events in terms of quantiles of the marginal distribution of the observations (left) and values in knots (right).

outperform the NGR forecasting procedure. Again, it can be observed that restricting proper scoring rules to subsets of extreme events yields improper scoring rules and discredits skillful and calibrated forecasting systems. Plots of the forecasting procedures not shown in Figure 4.6 and plots of the restricted MAE yield qualitatively equivalent results and are omitted here.

Furthermore, the restricted CRPS can be used to test for equal performance using the Diebold-Marino-type test statistic

$$t_n = \sqrt{n}\,\frac{\bar{S}_n^f - \bar{S}_n^g}{\hat{\sigma}_n}$$

depending on the threshold defining extreme events (Diebold and Mariano, 1995). Computing p-values associated with the values of the test statistic under the standard normal hypothesis allows us to gain further insight into the significance of the observed score differences. Note that the standard normal assumption might be violated due to the small sample sizes for large thresholds.

Figure 4.7 shows the test statistics of the test of equal performance comparing the NGR and the GEV2minCRPS forecasting procedure as well as the associated p-values. The asymptotic variance was estimated as proposed by Gneiting and Ranjan (2011b). If the test statistic attains values smaller than 0, the NGR forecaster is preferred over the GEV2minCRPS forecaster, otherwise the GEV2minCRPS forecaster is preferred over the NGR forecaster. The test statistic is negative up to a threshold of approximately the 70th percentile of the marginal distribution of the observations. For larger thresholds, the test statistic is positive and prefers the GEV2minCRPS forecaster over the NGR forecaster. In the simulation study discussed in Chapter 3, a small interval of insignificant score differences around 0

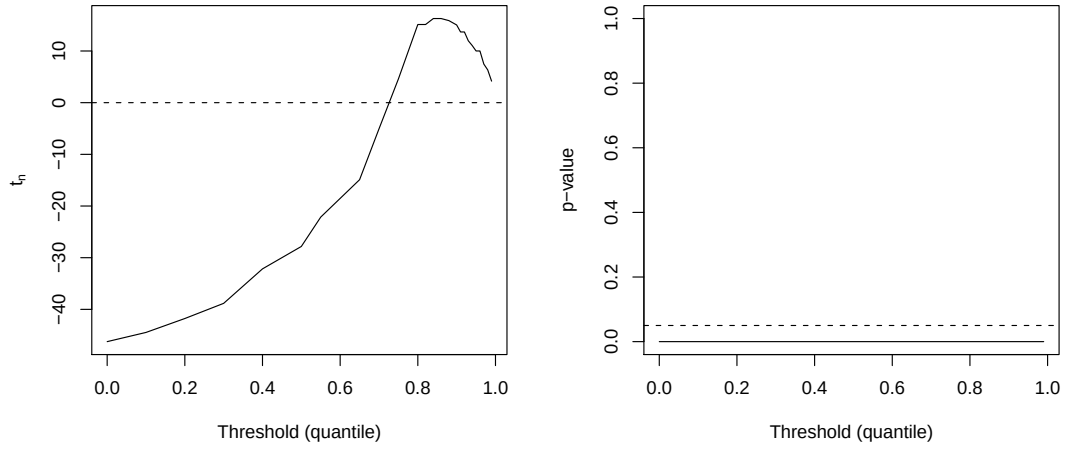

Figure 4.7.: Left: Diebold-Mariano-type test statistic for test of equal performance comparing the NGR and the GEV2minCRPS forecasting procedure as functions of the threshold defining extreme events, using the restricted CRPS. Right: Corresponding p-values under the standard normal hypothesis, the dashed line indicates a 5% significance level.

could be observed (c.f. Figure 3.6). Here, all observed score differences are significant under the standard normal hypothesis because all observations are rounded to the nearest whole knot. For the threshold $r = 23$, the value $t_n = -5.1$ significantly favors the NGR forecasting procedure, while for the threshold $r = 24$, the value $t_n = 4.6$ significantly favors the GEV2minCRPS forecasting procedure. Therefore, no interval of insignificant score differences is observed.

The test of equal performance based on the restricted CRPS thus prefers the GEV2minCRPS forecasting procedure over the NGR forecasting procedure for thresholds larger than 23 which again confirms the impropriety of the restricted CRPS. For all other forecasting procedures based on EVT, qualitatively equivalent results can be obtained.

4.4.3. Results for proper scoring rules for extreme events

Turning to proper scoring rules for extreme events as proposed by Gneiting and Ranjan (2011b) and Diks et al. (2011), a deficiency of the conditional and censored likelihood scoring rules becomes obvious. Meaningful forecast evaluation using these proper scoring rules proposed by Diks et al. (2011) requires the predictive densities to attain strictly positive values at all observations. Therefore, applying them to the predictive GEV and GP densities results in infinite mean scores, no longer allowing us to distinguish predictive performance. Therefore, we focus on the threshold weighted CRPS,

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

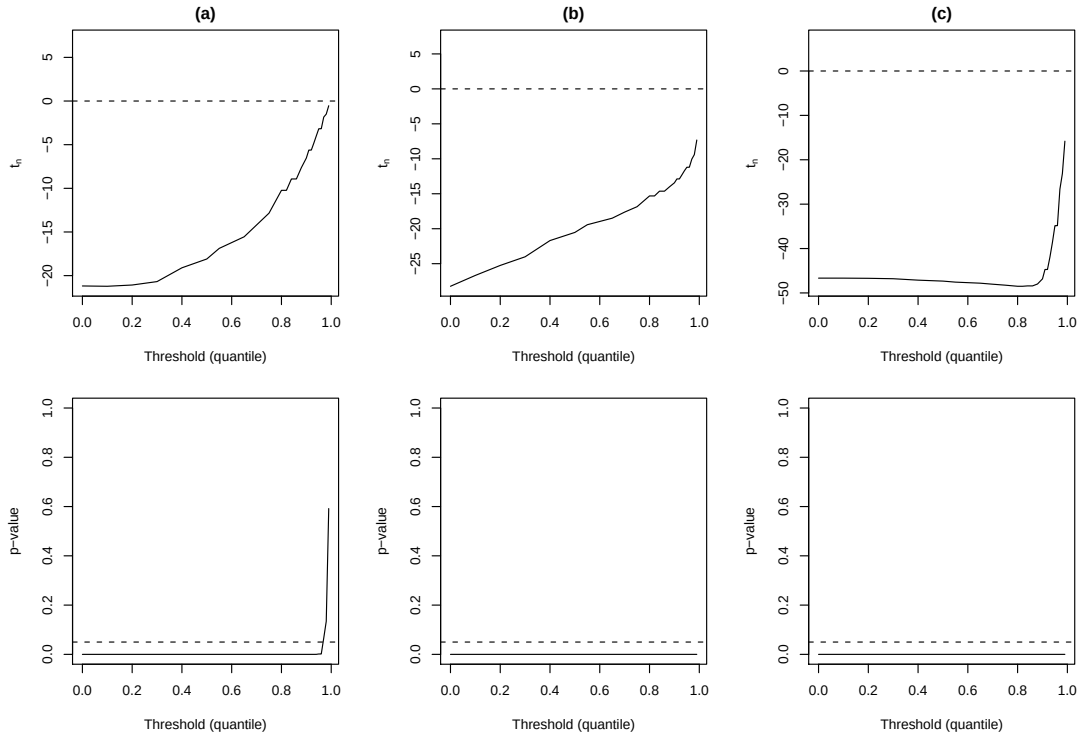

Figure 4.8.: Diebold-Mariano-type test statistics for test of equal performance using the threshold-weighted CRPS to compare the NGR model and (a) the climatological forecaster, (b) the raw ensemble, and (c) the GEV2minCRPS forecasting procedure as functions of the threshold which defines extreme events. The bottom row shows the corresponding p-values under the standard normal hypothesis, the dashed lines indicate a 5% significance level.

which does not suffer from this deficiency.

Figure 4.8 shows plots of the Diebold-Mariano-type test statistics using the threshold-weighted CRPS to compare the NGR model and the raw ensemble, the climatological, and the GEV2minCRPS forecasting procedure. Here, the indicator weight function $w_r(z) = \mathbb{1}(z \geq r)$ was used. In all three cases, the test statistics attain only negative values and always prefer the NGR forecaster over the respective competitor. All observed score differences are significant, except for the score differences between the NGR and the climatological forecasting procedure for very large thresholds. Note that due to the dependence of the NGR forecasting procedure on the poorly performing ensemble, this might be a consequence of the lack of correlation between ensemble predictions and observations for extreme events. For details, see Figure 4.13.

Thus, using the proper scoring rule proposed by Gneiting and Ranjan (2011b) shows that the NGR forecaster outperforms any forecasting procedure based on EVT if the right tail of the marginal distribution of gust speed observations is emphasized. Again, qualitatively equivalent results are obtained for all other forecasting procedures based on EVT and for the normal CDF weight function

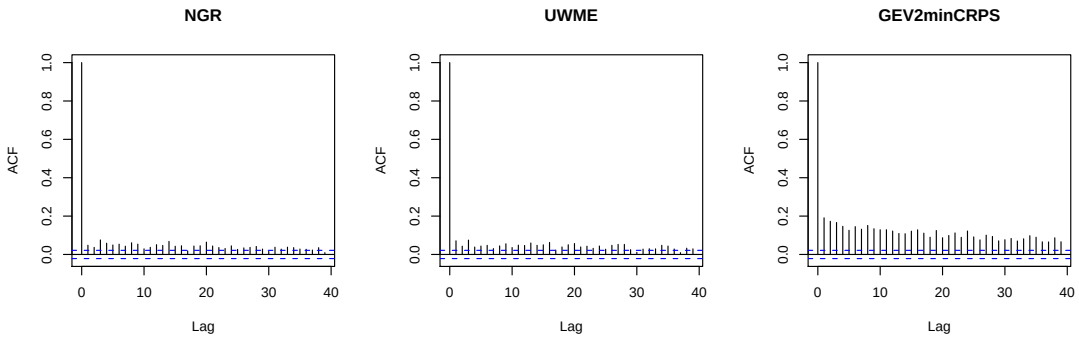

Figure 4.9.: Autocorrelation of the values of the threshold-weighted CRPS for the NGR model, the raw ensemble and the GEV2minCRPS forecaster.

$w_r(z) = \Phi_{r,1}(z)$.

The asymptotic variance was estimated as proposed by Gneiting and Ranjan (2011b). Plots of empirical estimates of the autocorrelation of the score vectors for different forecasters as displayed in Figure 4.9 show that the theoretical assumption of at most $(k-1)$-dependence is not violated here and that the applied variance estimation procedure thus is appropriate. Here, $k = 2$ since the forecasts are issued 48-h-ahead and the gust speeds are observed as daily maximum. For the GEV2minCRPS forecasting procedure, a larger amount of autocorrelation can be observed which is due to the constant bias towards larger values.

These empirical results confirm the observations made in the simulation study of Chapter 3. Restricting the evaluation to subsets of extreme events corresponds to the use of improper scoring rules and discredits skillful and calibrated forecasting procedures. If, however, proper scoring rules as proposed by Gneiting and Ranjan (2011b) are used, the NGR forecasting procedure is preferred over the competitors based on EVT.

4.5. Extreme ensemble forecasts

In the situation of our case study, it might be interesting to assess which forecasting procedure performs best in case of extreme ensemble predictions. Note that conditioning on extreme ensemble predictions significantly differs from conditioning the observations on being extreme events.

Conditioning Y on being an extreme event, $Y^\star = Y|Y \geq r$ with $r \in \mathbb{R}$, corresponds to restricting the observation space $(\Omega_Y, \mathcal{A}_Y)$ to a subspace $(\Omega_Y^\star, \mathcal{A}_Y^\star)$. Therefore, the scoring rule restricted to the subset of extreme events, $S^\star$, is a mapping

$$S^\star : \mathcal{P} \times \Omega_Y^\star \longrightarrow \bar{\mathbb{R}}$$

which is minimized by

$$\mathcal{L}(Y|\mathcal{A}^\star) = \mathcal{L}(Y^\star) = F^\star \neq F = \mathcal{L}(F) = \mathcal{L}(Y|\mathcal{A}).$$

Therefore, $S^\star$ is an improper scoring rule, even if S is proper.

On the contrary, conditioning Y on the corresponding median ensemble prediction being an extreme event, $\tilde{Y} = Y|\text{med}\{X_1, \ldots, X_k\} \geq r$ with $r \in \mathbb{R}$, does not effect the observation space $(\Omega_Y, \mathcal{A}_Y)$ because Y does not depend on the ensemble predictions $X_1, \ldots, X_k$. Computing the value of a proper scoring rule S restricted to the subset of observations made after the median ensemble prediction was an extreme event corresponds to the use of the scoring rule

$$\tilde{S} : \tilde{\mathcal{P}} \times \Omega_Y \longrightarrow \bar{\mathbb{R}},$$

where $\tilde{\mathcal{P}}$ is the class of probabilistic forecasts of forecasters with information bases containing the information that the median ensemble prediction is an extreme event. $\tilde{S}$ is minimized by $\mathcal{L}(\tilde{Y}) = \mathcal{L}(Y|\tilde{\mathcal{A}})$, where

$$\mathcal{L}(Y|\tilde{\mathcal{A}}) = \begin{cases} G_0 = \mathcal{L}(Y) & \text{if } \mathcal{L}(Y) \in \tilde{\mathcal{P}}, \\ \tilde{G} = \min_{G \in \tilde{\mathcal{P}}} \tilde{\boldsymbol{S}}(\tilde{G}, \mathcal{L}(\tilde{Y})) & \text{if } \mathcal{L}(Y) \notin \tilde{\mathcal{P}}. \end{cases}$$

Unlike $S^\star$, $\tilde{S}$ is a proper scoring rule since

$$\tilde{\boldsymbol{S}}(\tilde{F}, G_0) \geq \tilde{\boldsymbol{S}}(\tilde{G}, G_0) = \min_{G \in \tilde{\mathcal{P}}} \tilde{\boldsymbol{S}}(\tilde{G}, G_0) \geq \tilde{\boldsymbol{S}}(G_0, G_0)$$

for all $\tilde{F} \in \tilde{\mathcal{P}}$.

Note that in particular, the conditioning on the ensemble prediction cannot be achieved by employing a weight function $w(y)$ on the observations and the result of Theorem 2.27 does not hold here.

Results

Here, we set $r = 29$ and consider only observations which were made after the median ensemble prediction for gust speed was at least 29 kt. Ensemble predictions for gust speed are obtained as products of ensemble predictions for wind speed and gust factors as discussed in 4.2.1.

Figure 4.10 shows the verification rank histogram for the ensemble and PIT histograms for the NGR and the GEV2minCRPS forecasting procedure restricted to the subset of observations made after the median ensemble prediction was an extreme event. The ensemble and the GEV2minCRPS forecasting procedure are biased. Both models overestimate the gust speed and are not probabilistically calibrated. Similar results can be obtained for all other GEV and GP forecasting procedures. With only minor deviations from uniformity, the NGR forecasting procedure significantly improves the ensemble predictions and appears to be probabilistically calibrated for this subset of gust speed observations.

Qualitatively, the same can be observed when marginal calibration is examined. While the NGR forecasting procedure appears to provide a good fit to the subset of gust speed observations obtained after extreme ensemble predictions, none of the competing forecasting procedures based on EVT is marginally calibrated, see Figure 4.11.

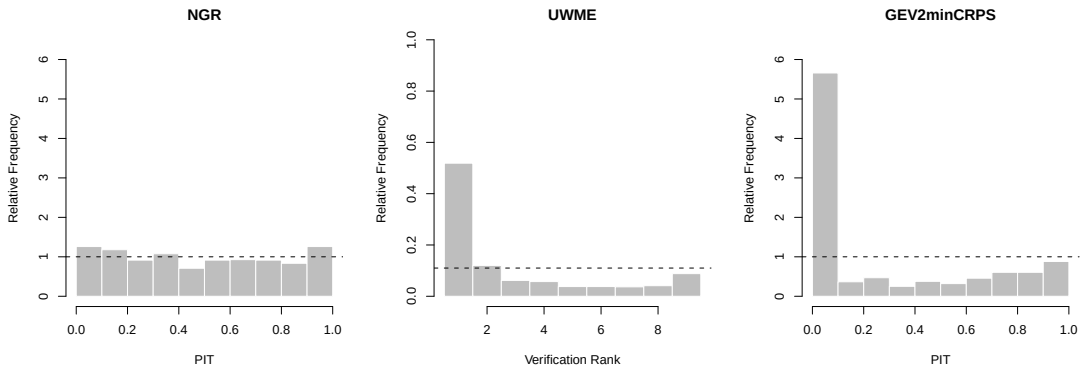

Figure 4.10.: Verification rank histogram for the ensemble and PIT histograms for the NGR and the GEV2minCRPS forecasting procedure based on the gust speed observations made after the median ensemble prediction was at least 29 kt.

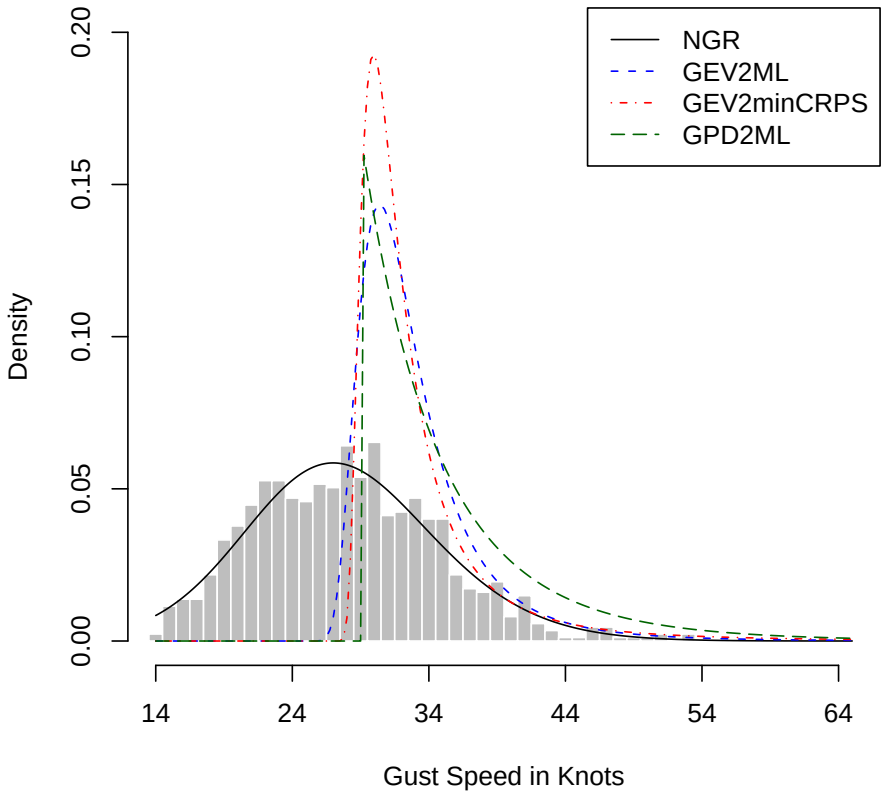

Figure 4.11.: Histogram of gust speeds observed after extreme median ensemble predictions and marginal predictive densities under the NGR and the competing models based on EVT.

Table 4.4.: Mean CRPS, MAE, average coverage (in %) and width of 77.8% prediction intervals for various probabilistic gust speed forecasts and gust speed observations made after the median ensemble prediction was at least 29 kt.

Forecast	CRPS	MAE	Coverage	Width
Climatology	6.11	8.50	54.2	13.14
UWME with gust factors	6.15	7.55	39.3	10.41
NGR	**3.88**	**5.45**	**72.3**	14.65
GEV1	5.57	5.89	35.1	10.00
GEV2ML	5.49	5.56	32.0	8.87
GEV2minCRPS	5.39	5.42	32.8	**8.46**
GEV3ML	5.61	5.64	33.1	9.54
GEV3minCRPS	5.57	5.64	33.1	8.73
GP1ML	5.92	6.17	37.3	11.62
GP2ML	5.73	5.92	35.1	10.12
GP3ML	6.10	6.41	30.5	9.25

Table 4.4 shows the values of the summary measures of predictive performance for the subset of observations made after extreme median ensemble predictions. In the case of large ensemble predictions, the ensemble performs far worse compared to the set of all gust speed observations and is outperformed by all competing forecasting procedures. Despite the dependence on the ensemble predictions, the NGR forecasting procedure significantly improves the ensemble predictions and performs best in terms of mean CRPS, MAE and average coverage of 77.8% prediction intervals. In particular, the NGR forecasting procedure outperforms all forecasting procedures based on EVT. The NGR forecasting procedure is thus well able to correct the biased and uncalibrated extreme ensemble predictions. Note that the average width of 77.8% prediction intervals is significantly larger compared to the subset of extreme events or all events for both the ensemble predictions and the NGR forecasting procedure. Therefore, larger median ensemble predictions are associated with larger uncertainty.

Figure 4.12 shows the value of the mean CRPS restricted to subsets of gust speed observations made after extreme ensemble predictions as a function of the threshold r which defines extreme events. For increasing values of r, the mean CRPS values of the climatological forecaster, the ensemble and the NGR forecasting procedure increase. The mean CRPS curves of the climatological forecaster and the ensemble exhibit a significantly larger slope than the curve of the NGR forecasting procedure. On the contrary, the mean CRPS values of the GEV and GP forecasting procedures decrease for larger thresholds. For thresholds larger than around 27, which approximately corresponds to the 80th percentile of the gust speed observations, all forecasting procedures based on EVT and the climatological forecaster outperform the ensemble. However, even for large thresholds, the NGR forecasting procedure still outperforms all GEV and GP forecasting procedures.

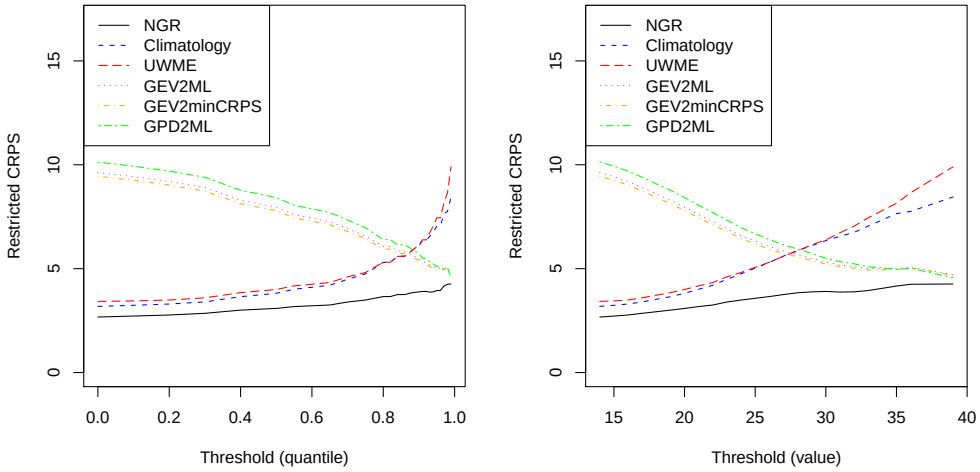

Figure 4.12.: CRPS restricted to subsets of observations made after the median ensemble prediction was larger than a threshold as function of this threshold in terms of quantiles of the marginal distribution of the observations (left) and values in knots (right).

These empirical results confirm the theoretical results from above in that restricting the CRPS to subsets of the observations conditional on the ensemble predictions does not discredit skillful and calibrated forecasting procedures as the NGR model.

4.6. Regime-switching combination of NGR and GEV forecasting procedures

4.6.1. Wind gust

The theoretical results from EVT summarized in Section 4.2.2 show that extreme values can be modeled using GEV or GP distributions. Assuming that the ensemble predictions and the observations are highly correlated, it should thus be possible to improve the NGR forecasting procedure by combining it with a forecasting procedure based on EVT, depending on the ensemble predictions. In particular, we use the NGR model if the median ensemble prediction is smaller than a threshold $r \in \mathbb{R}$ and a forecasting procedure based on EVT otherwise.

However, no combination of thresholds and EVT forecasting procedures leads to any improvements of the results of the NGR forecasts for the gust speed observations. This might be a consequence of the lack of correlation between the median ensemble predictions and the observations. If no significant positive correlation between the ensemble predictions and the observations is observed, the observations made after extreme ensemble predictions do not follow the theoretical extreme value distribution. Figure 4.13 shows pairs of observations and point predictions, obtained as median ensemble predictions, for all gust speed observations, and for

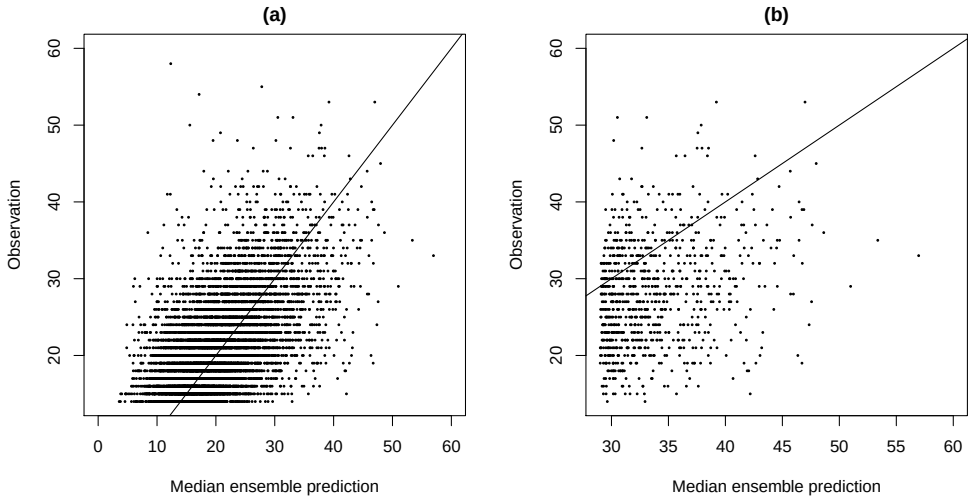

Figure 4.13.: Scatterplot of the median ensemble predictions for gust speed and the corresponding gust speed observations for (a) all observations and (b) the subset of observations conditional on the median ensemble prediction being an extreme event. The solid line indicates an angle of 45 degrees which corresponds to a perfect linear dependence.

gust speed observations made after the median ensemble prediction was at least 29 kt. While there seems to be a relatively large correlation for all observations, the ensemble clearly overestimates the gust speed in the case of extreme ensemble predictions of gust speed. The correlation of 0.52 for all observations reduces to 0.20 for the restricted sample. This deficiency of the ensemble forecasts might arise from the approach of estimating gust speed assuming a simple multiplicative relationship between wind speed and gust speed. The lack of correspondence between the ensemble predictions and the gust speed observations furthermore explains the comparatively bad results for the ensemble and the NGR forecasting procedure in case of extreme ensemble predictions as summarized in Table 4.4. Improvements might be achieved by more elaborated models for gust speed, for example by modeling gust speed as a more general function of wind speed. For details, see Thorarinsdottir and Johnson (2012) and the references therein.

4.6.2. Wind speed

Here, we turn to the corresponding ensemble predictions and observations of wind speed. Scatterplots of the median ensemble predictions and the wind speed observations are shown in Figure 4.14. Here, extreme predictions are identified as predictions of at least 17 kt, the 90th percentile of the wind speed observations. Compared to the ensemble predictions for gust speed, a larger concordance between predictions and observations can be observed, although the ensemble still appears to overestimate the wind speed in case of large values of the median ensemble prediction. The correlations of 0.60 for all observations and 0.29 for observations made after extreme median ensemble predictions are improved as well.

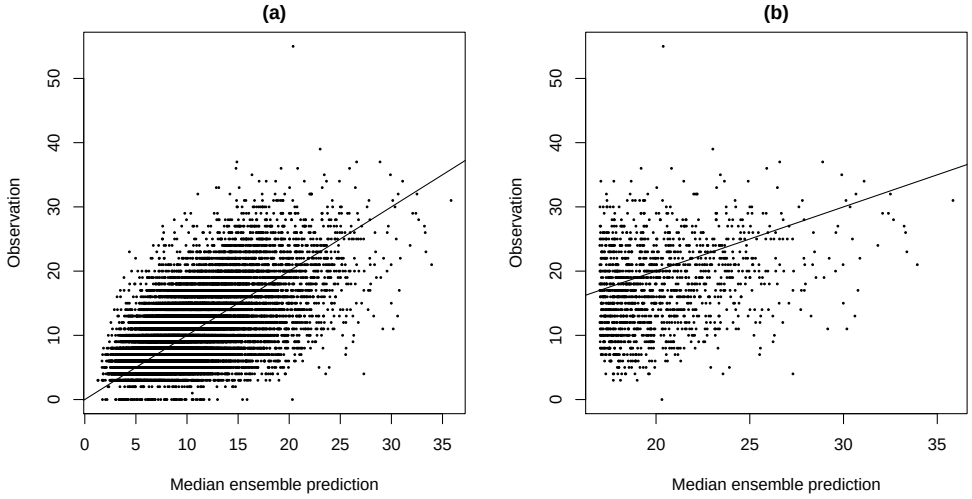

Figure 4.14.: Scatterplot of the median ensemble predictions for wind speed and the corresponding wind speed observations for (a) all observations and (b) the subset of observations conditional on the median ensemble prediction being an extreme event. The solid line indicates an angle of 45 degrees which corresponds to a perfect linear dependence.

None of the GEV and GP forecasting procedures introduced in Section 4.2.2 depends on the ensemble predictions. In order to improve the predictive performance of the forecasting procedures based on EVT, we use a Bayesian covariate selection method as described in Section 3.3 of Friederichs and Thorarinsdottir (2012). Details on the regression variable selection algorithm can be found in Stephenson and Tawn (2004), Galiatsatou et al. (2008) and Hoff (2009). We estimate the parameters of a GEV distribution, where the location parameter $\mu^{\text{GEV}} = \mu_0 + \alpha\bar{X}$ is given as a linear function of the ensemble mean $\bar{X}$, and the scale and shape parameter $\sigma^{\text{GEV}} = \sigma_0$ and $\xi^{\text{GEV}} = \xi_0$ are estimated independent of the ensemble predictions.

For the model selection, we focused on the GEV forecasting procedures and model location and scale parameter of the GEV distributions as linear functions of the ensemble mean and the ensemble variance. However, using data from 2006 and 2007 as training data, the Markov chains of the latent variables associated with the inclusion of the covariates failed to converge except for the chains associated with the constant parts for both the location and the shape parameter, and the ensemble mean for the location parameter, which yields the model from above.

The parameters μ_0, α, σ_0 and ξ_0 are estimated using minimum CRPS estimation over different training sets as described in Section 4.2.2. For different thresholds r and different training sets, the best results were obtained by using a threshold of 13 kt and an estimation procedure similar to the GEV3minCRPS procedure. The parameters of the GEV distribution are thus estimated station-wise, assuming to be constant in time, using all wind speed observations from 2006 and 2007 made at the specific station after the median ensemble prediction was at least 13 kt. This forecast model for wind speed will be referred to as the GEV3minCRPS* model. The threshold value of 13 kt approximately corresponds to the 78th percentile of

Table 4.5.: Mean CRPS, MAE, average coverage (in %) and width of 77.8% prediction intervals for various probabilistic wind speed forecasting procedures.

Forecast	CRPS	MAE	Coverage	Width
UWME	2.64	3.23	45.3	**4.94**
NGR	2.20	3.06	**78.6**	9.16
GEV3minCRPS*	2.34	3.29	78.7	10.51
NGR + GEV3minCRPS*	**2.06**	**2.88**	81.1	9.24

the observations and the 80th percentile of the median ensemble predictions. Note that the results for the GEV3minCRPS* forecasting procedure for all wind speed observations are worse than those of the NGR forecasting procedure as the wind speed observations obviously do not follow an extreme value distribution.

We consider the following regime-switching combination of probabilistic forecasts for wind speed produced by the NGR and GEV3minCRPS* models, depending on the median ensemble prediction. For the individual forecast cases $t = 1, \ldots, 22\,863$, the predictive CDF is given by

$$\hat{F}_t = \begin{cases} \Phi_{[0,\infty)}(\mu_t^{\mathcal{N}}, \sigma_t^{\mathcal{N}}), & \text{if } X_m^t < 13 \text{ kt}, \\ F_{\text{GEV}}(\mu_t^{\text{GEV}}, \sigma_t^{\text{GEV}}, \xi_t^{\text{GEV}}), & \text{if } X_m^t \geq 13 \text{ kt}, \end{cases} \tag{4.2}$$

where X_m^t denotes the corresponding median ensemble prediction for wind speed and $\Phi_{[0,\infty)}(\mu_t, \sigma_t)$ denotes the CDF of a truncated normal distribution with a cutoff at zero, $\mathcal{N}_{[0,\infty)}(\mu_t^{\mathcal{N}}, \sigma_t^{\mathcal{N}})$. Note that $\mu_t^{\mathcal{N}}$ and $\sigma_t^{\mathcal{N}}$ denote mean and standard deviation for the truncated normal distribution, and μ_t^{GEV} and σ_t^{GEV} denote location and scale for the GEV distribution. The NGR model and the GEV model do not share any parameter values, the common symbols are used for consistency of the notation. The parameters of the truncated normal distribution are estimated using the NGR approach as described in Section 4.2.1, the parameters of the GEV distribution are estimated using the GEV3minCRPS* approach described above. Results for the summary measures of predictive performance are given in Table 4.5.

The combination of the NGR and the GEV3minCRPS* forecasting procedures significantly improves the results of the NGR forecasting procedure. The mean CRPS is reduced from 2.20 to 2.06 (by approximately 6.1%). The mean CRPS for observations made after the median ensemble prediction was at least 13 kt is improved from 3.24 to 2.58 (by 20.4 %). For different choices of thresholds, training sets and parameter estimation procedures for the GEV distribution, the combined forecasting procedures also outperform the NGR forecasting procedure. However, the improvements are considerably smaller than for the combination as defined in (4.2). For example, if the same threshold is chosen and the parameters are estimated independent of the ensemble predictions using the same training sets as above and ML estimation, the mean CRPS of the NGR method is reduced from 2.20 to 2.11. Note that these improvements of the combined forecasting procedures compared to the NGR forecasting procedure are not due to the location-specific

estimation of the GEV parameters. The local approach for the NGR method results in lower general predictive performance (Thorarinsdottir and Johnson, 2012).

For the purpose of assessing probabilistic calibration, we use probability plots as suggested by Coles (2001). Let $y_t, t = 1, \ldots, T$ denote the ordered observations, then the empirical cumulative distribution function evaluated at y_t is given by

$$\bar{F}(y_t) = \frac{t}{T+1}.$$

The corresponding model-based estimates are given by $\hat{F}_t(y_t)$ for $\hat{F}$ as defined in (4.2). For a probabilistically calibrated model, the probability plot consisting of the points

$$\left\{\left(\bar{F}(y_t), \hat{F}_t(y_t)\right), t = 1, \ldots, T\right\}$$

should lie close to the unit diagonal (Coles, 2001). Figure 4.15 shows the probability plots for the NGR model and for the combination of the NGR and the GEV3minCRPS* forecasting procedure defined in (4.2). Both methods appear to be quite well calibrated. For larger wind speed observations, indicated by larger values of the empirical CDF, the curve corresponding to the combination of the NGR and the GEV3minCRPS* forecasting procedure is slightly closer to the unit diagonal. Therefore, the combined model is slightly better calibrated than the NGR model for larger observations. However, the NGR model is slightly better calibrated for lower values. The corresponding PIT histograms shown in Figure 4.16 suggest the same interpretation. The verification rank histogram for the raw ensemble forecasts indicates underdispersion and is omitted here.

To assess marginal calibration, we examine the marginal predictive densities of the models. Figure 4.17 shows a histogram of the observed wind speeds together with marginal predictive densities for the NGR model, the GEV3minCRPS* model and the combination of those two models. Compared to the NGR model, the combined model appears to improve the marginal calibration, in particular for larger wind speed observations. Note that both the NGR model and the combined model overestimate the probability of very small wind speed observations.

To summarize, the predictive performance of the NGR model can be improved by combining it with models based on EVT dependent on the median ensemble predictions. The combination of the NGR model and the GEV3minCRPS* model results in lower values of the mean CRPS and the MAE and exhibits better probabilistic and marginal calibration, especially for large wind speed observations. If there was a larger correlation between high values of the median ensemble predictions and the wind speed observations, the results for the combined model would possibly be even more improved.

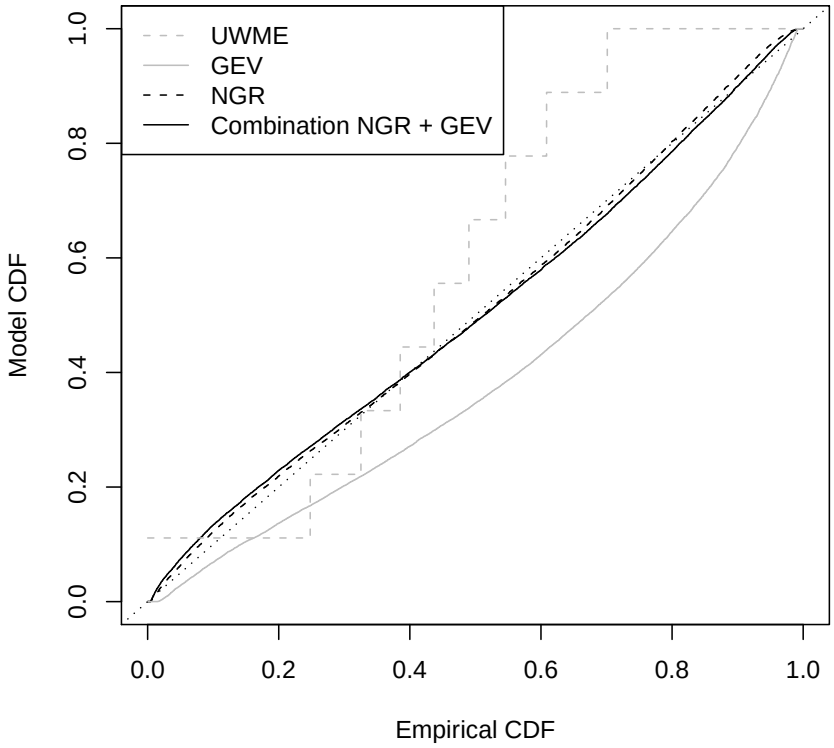

Figure 4.15.: Probability plot comparing the empirical CDF of the wind speed observations with the model CDF of the raw ensemble (gray dashed line), the GEV3minCRPS* model (gray solid line), the NGR model (black dashed line) and the forecasting procedure combining the NGR and the GEV3minCRPS* model (black solid line). The dotted line indicates the unit diagonal.

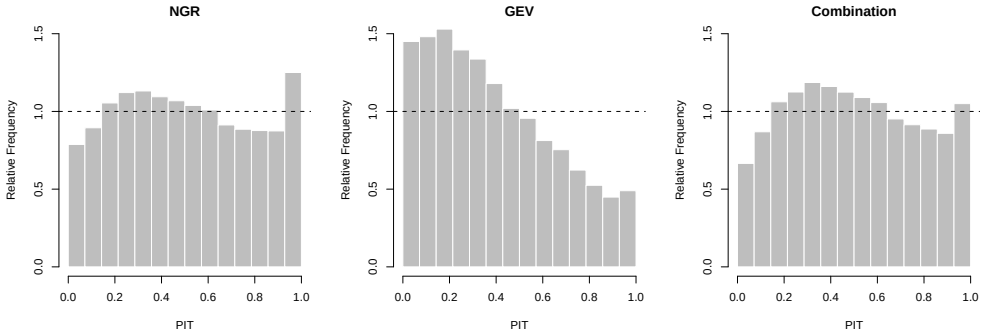

Figure 4.16.: PIT histograms for the NGR model, the GEV model and the forecasting procedure combining the NGR and the GEV3minCRPS* model.

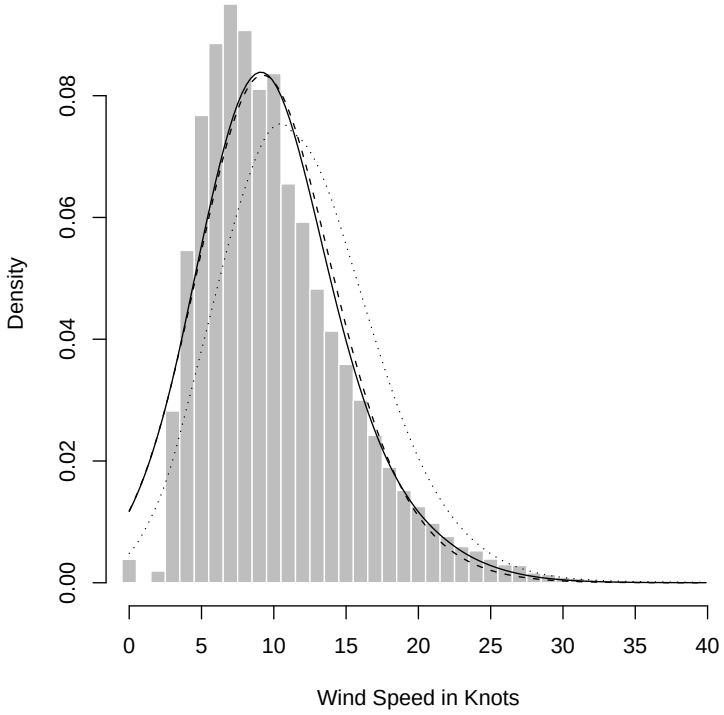

Figure 4.17.: Histogram of the observed wind speeds together with marginal predictive densities of the NGR model (dashed line), the GEV3minCRPS* model (dotted line) and the combination of those models (solid line) as defined in (4.2).

5. Connections to evaluation procedures for binary events

We discuss connections to the theory of forecast evaluation for binary predictions and events. With the advent of operational weather forecasting in the second half of the 19th century, questions about the quality of forecasts arose (Murphy, 1996). The work of Finley (1884) marks the beginning of substantial developments in the discipline of forecast verification. Since this seminal paper and the following responses, the forecast verification for binary events and predictions such as the tornado warnings analyzed by Finley (1884) focuses on performance measures based on 2×2 *contingency tables*. A contingency table C lists the number of 0's predicted as 0, the number of 0's predicted as 1 (false alarms), the number of 1's predicted as 0 (misses), and the number of 1's predicted as 1 (hits). In the literature, there exists no unique notation and ordering of the entries of contingency tables. Here, contingency tables will always be of the form shown in Table 5.1.

For the purpose of verifying binary predictions of binary events, various summary measures of contingency tables have been proposed and discussed in the meteorological literature. Prominent examples are

- the *Hit Rate*, $H = \frac{d}{b+d}$,
- the *False Alarm Rate*, $F = \frac{c}{a+c}$,
- the *Fraction Correct*, $\mathrm{FRC} = \frac{a+d}{n}$,
- the *Critical Success Index*, $\mathrm{CSI} = \frac{d}{b+c+d}$, and
- *Heidke's Skill Score*, $\mathrm{HSS} = \frac{2 \det C}{n_0(c+d)+n_1(a+b)}$.

Except for the False Alarm Rate, all given performance measures are positively oriented, larger values indicate a better predictive performance. For more examples

Table 5.1.: Contingency table for the evaluation of deterministic forecasts of binary events.

	Non-Event observed	Event observed	
Non-Event forecasted	a	b	$a+b$
Event forecasted	c	d	$c+d$
	$n_0 = a+c$	$n_1 = b+d$	n

and a detailed comparison of the performance measures, we refer to Doswell et al. (1990), Marzban (1998) and Mason (2003).

Some of the given performance measures allow for hedging. For example, a simple hedging strategy for the Hit Rate is given by always predicting the event. This obviously unskillful forecasting approach leads to the largest possible Hit Rate of 1. A similar hedging strategy can be found for the False Alarm Rate by always predicting a non-event. Gilbert (1884) demonstrates that the Fraction Correct (FRC), which was used by Finley (1884), can be hedged, Gandin and Murphy (1992) show the same for the Critical Success Index. More elaborated measures such as Heidke's Skill Score (Heidke, 1926) appear to be less prone to hedging. Heidke's Skill Score is currently used by the German National Weather Service (DWD) to evaluate storm warning systems (Deutscher Wetterdienst, 2009).

The *base rate*, i.e. the proportion $p = \frac{n_1}{n}$ of observed events, cannot be controlled by a forecasting system and should thus not affect the assessment of predictive quality (Mason, 2003). However, Marzban (1998) shows that under certain regularity conditions, most of the widely used performance measures are base-rate dependent and furthermore converge to a trivial limit as the rarity of the predicted event increases. In a nutshell, the performance measures degenerate because the entries b, c, and d of the contingency table tend to converge to zero at unequal rates (Ferro, 2007; Ferro and Stephenson, 2011). Therefore, these measures do not allow for a meaningful performance evaluation for rare binary events with low base rates. This observation was followed by the development of new base-rate independent performance measures which do not degenerate and converge to meaningful limits for rare events. Stephenson et al. (2008) proposed the *Extreme Dependency Score*, $\text{EDS} = \frac{2\log[(d+b)/n]}{\log[d/n]} - 1$, which, however, was later shown to be base-rate dependent and prone to hedging (Primo and Ghelli, 2009; Ghelli and Primo, 2009; Ferro and Stephenson, 2011). These and other shortcomings of the EDS led to the development of the *Extremal Dependence Index*,

$$\text{EDI} = \frac{\log(F) - \log(H)}{\log(F) + \log(H)},$$

by Ferro and Stephenson (2011), where H denotes the Hit Rate and F denotes the False Alarm Rate. The EDI is positively oriented and restricted to the interval $[-1, 1]$. Ferro and Stephenson (2011) show that the EDI is base-rate independent and does not allow for hedging.

Note that as the summary measures discussed before, the EDI was developed for the performance evaluation for binary predictions of binary events. However, binary predictions and events can be easily obtained from general real-valued point predictions and observations by choosing thresholds which identify (in our case extreme) events. In this process, some information about the distribution of the predictions and observations is lost.

In this thesis, we investigate the more general situation of probabilistic forecasts. Point predictions can be obtained from probabilistic forecasts as functionals of the predictive distributions. However, using these point predictions obtained from probabilistic forecasts is accompanied by a further loss of information since the

point predictions do not contain any uncertainty information, and leads to erroneous performance evaluation. This can be demonstrated using the data example of gust speed predictions discussed in Chapter 4.

The EDI suffers from the drawback that it is not able to detect differences in the predictive performance arising from a bias of the predictions and should thus only be used to evaluate calibrated forecasts (Ferro and Stephenson, 2011). Here, forecasts are said to be calibrated if the number of forecasted events $(c+d)$ equals the number of observed events $(b+d)$. If extreme events are defined as those observations larger or equal to a threshold $r \in \mathbb{R}$, the following contingency tables for the NGR forecasting procedure and the raw ensemble forecasts are obtained:

- $r = 29$ knots (corresponds to the 90th percentile of the marginal distribution of the observations):

$$C_{\text{NGR}} = \begin{pmatrix} 795 & 7234 \\ 175 & 120 \end{pmatrix} \quad \text{and} \quad C_{\text{UWME}} = \begin{pmatrix} 571 & 6880 \\ 399 & 474 \end{pmatrix}$$

- $r = 33$ knots (corresponds to the 95th percentile of the marginal distribution of the observations):

$$C_{\text{NGR}} = \begin{pmatrix} 377 & 7856 \\ 50 & 41 \end{pmatrix} \quad \text{and} \quad C_{\text{UWME}} = \begin{pmatrix} 297 & 7640 \\ 130 & 157 \end{pmatrix}$$

Here, the same thresholds were used to identify observed and predicted extreme events and the point predictions were obtained as the median of the corresponding predictive distributions. However, obviously neither the NGR model nor the ensemble are calibrated for both thresholds. In order to apply the EDI, the predictions thus have to be recalibrated. Ferro (2007) and Stephenson et al. (2008) propose a simple recalibration technique, where the thresholds for observations and predictions are chosen as upper q quantiles for the same value of q. However, this recalibration technique did not result in calibrated forecasts for our data example. Therefore, we recalibrate the forecasts using numerical optimization to minimize the absolute difference between the number of observed and predicted events to find thresholds which identity predicted extreme events for given values of r. Note that there are situations in which it might be impossible or not desirable to recalibrate the forecasts (Hogan et al., 2009).

The thresholds for identifying predictions of extreme events strongly differ from the corresponding thresholds for the observations. For $r = 29$ knots, thresholds of 24.6 knots (NGR) and 28.4 knots (UWME) are obtained, for $r = 33$ knots, thresholds of 27.7 knots (NGR) and knots (UWME) are obtained, respectively.

The standard error of the EDI can be estimated by

$$s_{\text{EDI}} = \frac{2|\log(F) + \frac{H}{1-H}\log(H)|}{H(\log(F) + \log(H))^2} \sqrt{\frac{H(1-H)}{pn}},$$

as proposed by Ferro and Stephenson (2011). Table 5.2 summarizes values of the EDI and estimates of the corresponding standard errors for the recalibrated fore-

casts. Recall that the EDI is positively oriented and restricted to the interval $[-1, 1]$. For $r = 29$ knots, the ensemble outperforms the NGR forecasting procedure, for $r = 33$ knots, the NGR model outperforms the ensemble. Note that the estimated standard errors are relatively large compared to the difference of the values of the EDI. Thus, the EDI is not able to correctly and significantly distinguish between the predictive performance of the NGR model and the ensemble.

Table 5.2.: Values and estimated standard errors of the EDI for the NGR model and ensemble predictions after recalibration.

Threshold	Forecast	EDI	s_{EDI}
29 knots	NGR	0.51	0.02
	UWME	**0.52**	0.02
33 knots	NGR	**0.51**	0.03
	UWME	0.49	0.03

If, on the other hand, the threshold-weighted CRPS is used to evaluate the predictive performance of the probabilistic forecasts, no recalibration is necessary and the NGR model significantly outperforms the ensemble. Table 5.3 shows the values of the threshold-weighted CRPS where an indicator function $w(z) = \mathbb{1}(z \geq r)$ was chosen as weight function. Here, the NGR model outperforms the ensemble predictions for both thresholds. The observed score differences are significant. The test statistic t_n of the test of equal performance comparing the NGR model and the ensemble attains a value of -13.4 (-11.2) for $r = 29$ knots ($r = 33$ knots), which corresponds to a p-value smaller than 10^{-40} (10^{-25}) under the standard normal hypothesis. A plot of the test statistic as a function of the threshold r in terms of quantiles of the marginal distribution of the gust speed observations can be found in Figure 4.8.

Table 5.3.: Values of the threshold-weighted CRPS for the NGR model and ensemble predictions (without recalibration).

Threshold	Forecast	CRPS^{t}
29 knots	NGR	**0.41**
	UWME	0.59
33 knots	NGR	**0.19**
	UWME	0.29

To summarize, this data example demonstrates that the EDI, a state-of-the-art forecast evaluation procedure for binary predictions of binary events based on contingency tables, does not suffice to assess the predictive performance of probabilistic forecasts for extreme events. Furthermore, using the EDI requires a recalibration of the predictions which might often be impossible or not desirable and can lead to counterintuitive results. Therefore, other approaches for the verification of probabilistic forecasts for rare and extreme events are needed. In this

thesis, we demonstrated that proper scoring rules for extreme events such as the threshold-weighted CRPS are well able to correctly and significantly distinguish the predictive performance of competing forecasting procedures.

6. Summary and discussion

Based on the work of Gneiting (2010), we have developed a general framework for the evaluation of probabilistic forecasts and analyzed the observation of the forecaster's dilemma in this framework. The theoretical results discussed in Chapter 2 show that conditioning proper scoring rules on extreme observations yields improper scoring rules. However, proper scoring rules for extreme events can be obtained using the approaches of Gneiting and Ranjan (2011b) and Diks et al. (2011) by employing appropriately weighted versions of the CRPS and the logarithmic score. For the purpose of comparing the predictive performance of competing density forecasts, tests of equal performance as proposed by Diebold and Mariano (1995) prove useful due to the different magnitude of the weighted scores for different weight functions or threshold values.

In Section 2.2.2, we discussed adequate combinations of proper scoring rules $S_1(P, \omega)$ and $S_2(P, \omega)$ of the form $S(P, \omega) = h(S_1(P, \omega), S_2(P, \omega))$ which result in proper scoring rules. Theorem 2.19 might be of use to develop proper scoring rules for extreme events which combine the advantages of the threshold- or quantile-weighted CRPS and the conditional or the censored likelihood scoring rules by combining them in an appropriate way. However, it might be difficult to find combinations which are readily interpretable and due to the large class of admissible weight functions, both the threshold- or quantile-weighted versions of the CRPS and the CL or CSL scoring rule appear to be sufficiently flexible. Therefore, we have focused on assessing the performance of the approaches of Gneiting and Ranjan (2011b) and Diks et al. (2011).

The simulation study conducted in Chapter 3 empirically confirms the theoretical results of the preceding chapter. If proper scoring rules are applied to subsets of extreme events, they prefer the biased forecaster over the ideal forecaster who predicts the true unconditional distribution of the observations. The scoring rules proposed by Gneiting and Ranjan (2011b) and Diks et al. (2011), on the other hand, are able to correctly and significantly distinguish the predictive performance. A comparison of the weighted versions of the CRPS and the logarithmic score suggests that these approaches also work for small sample sizes and different choices of weight functions. An important aspect is the choice of an estimator of the asymptotic variance of the score difference. The variance estimation procedure proposed by Diks et al. (2011) appeared to overestimate the asymptotic variance by taking into account autocorrelation up to a large lag which resulted in less significant score differences. Gneiting and Ranjan (2011b) follow the suggestions of Diebold and Mariano (1995) and only take into account auto-correlation up to a lag of at most $(k-1)$ if k-step-ahead forecasts are compared. For the situation of our simulation study, the forecast errors are 0-dependent by construction and the variance estimation procedure proposed by Gneiting and Ranjan (2011b) appears

to be the more suitable choice and produces more significant score differences. In general, the choice of an estimator of the asymptotic variance should be based on an empirical assessment of the range of dependence of the forecast errors at hand.

In Chapter 4, we investigated the forecaster's dilemma and the weighted proper scoring rules for extreme events in a real-world example using an application to wind gust forecasting. A deficiency of the CL and CSL scoring rules proposed by Diks et al. (2011) becomes obvious at this point. Both scoring rules require the predictive densities to attain strictly positive values at all observations. Since this did not hold for all models considered in our case study, we focused on the threshold-weighted version of the CRPS which does not suffer from this deficiency. Unlike restricted versions of the CRPS, the threshold-weighted CRPS is well able to correctly distinguish the predictive performance of various forecasting models. Again, the variance estimation procedure proposed by Gneiting and Ranjan (2011b) appeared to be more suitable than the variance estimator proposed by Diks et al. (2011).

Whenever ensemble predictions are available, it might be of interest to assess the predictive performance of competing forecasting procedures in the situation of extreme ensemble predictions. Within the framework developed in Chapter 2, we demonstrated that unlike conditioning on extreme events, conditioning on extreme ensemble predictions does not result in the use of improper evaluation procedures. For our example of gust speed forecasts, the NGR forecasting procedure clearly outperforms any competing forecasting procedure in the situation of extreme median ensemble predictions and still produces relatively calibrated and sharp forecasts except for very large thresholds. Based on this observations, we develope a simple regime-switching forecasting procedure for wind speed which combines the NGR model with GEV models based on results from extreme value theory. This novel approach to wind speed forecasting was able to significantly improve the NGR approach of Thorarinsdottir and Gneiting (2010), a state-of-the-art ensemble postprocessing technique.

The results presented in this thesis can be related to strands of work from various scientific disciplines such as the theory of forecast evaluation for binary events, economics and social psychology. In Chapter 5, we discussed connections to the forecast verification for binary predictions of extreme events and concluded that state-of-the-art summary measures for contingency tables do not suffice to correctly assess the predictive performance of probabilistic forecasts for extreme events. The forecast evaluation based on summary measures for contingency tables suffers from a further drawback. By allowing a large class of admissible weight functions, the weighted proper scoring rules can be easily extended to more complex situations. If, for example, very small and very large values of the observations are of interest, admissible weight functions emphasizing the tails of the distributions can easily be found. By contrast, generalizations of summary measures based on contingency tables require the development of summary measures for multidimensional contingency tables (Gandin and Murphy, 1992; Agresti, 2002; Livezey, 2003). Similar problems associated with the generalization of the region of interest arise for other approaches to forecast verification for binary predictions and events such as the use of the Relative Operating Characteristic (ROC) which is based on signal detec-

tion theory and is widely used in medical diagnostics and experimental psychology (Mason, 2003; Livezey, 2003). A condensed summary of the ROC approach can be found in Swets (1988).

We now turn to connections to work from other scientific disciplines. In economics, Denrell and Fang (2010) discuss an observation similar to the forecaster's dilemma. Performance evaluation in economics often focuses on extreme events. Managers and entrepreneurs are assessed by their ability to judge the success of new products, and those economists who were able to predict that something was to become "the next big thing" are seen as better forecasters. Furthermore, studying business success in the economic literature and the press is mostly based on case studies of mainly extreme events by focusing on entrepreneurs who became successful by predicting new trends. The fact that a product becomes successful can be seen as an extreme event since this is only accomplished by a small number of new products.

Exploiting the basic idea that due to the rarity of extreme events, managers who take all available information into account are less likely to predict extreme events, Denrell and Fang (2010) argue that accurately forecasting a rare and extreme event actually is a sign of poor judgment. They illustrate this observation using data from two lab experiments and from the Wall Street Journal Survey of Economic Forecasts. For the lab experiments, the mean squared error (MSE) of the participants is modeled as a function of the distance and quadratic distance between the predictions and the observations using a simple linear regression model. For the Wall Street Journal data, the absolute percentage deviation was used because of the different scales of the forecasted economic variables (gross domestic product, unemployment rate, consumer price index, Treasury bill rates and exchange rates). The absolute percentage deviation between the prediction p_t and the actual outcome y_t is given by $|p_t - y_t|/y_t$. The average absolute percentage deviation is used as a performance measure and modeled as a function of the percentage deviation from the actual value, $Dev_t = (p_t - y_t)/y_t$ and the same value squared, Dev_t^2, using a simple linear regression model. Extreme events are identified as threshold-exceedances of the known underlying model (lab experiments) and as observations which are at least 20% larger than the average value of the corresponding predictions (Wall Street Journal Survey data), respectively.

If only these extreme outcomes are considered, accurate predictions (low values of the distances and the percentage deviations, respectively) are associated with high values of the negatively oriented performance measures (MSE and average absolute percentage deviation, respectively) and therefore interpreted as a sign of poor (general) forecasting ability. The authors conclude that forecasting ability should be determined based on all observations, not on subsets of extreme events. Potential issues with this approach are that the general assumptions of linear regression are not fulfilled and that the verification procedures are not in line with the theoretical foundations described in Gneiting (2011).

The work discussed in this thesis generalizes the work of Denrell and Fang (2010) in two ways. Here, more general probabilistic forecasts are investigated and performance measures for extreme events are presented. These proper scoring rules allow for assessing the predictive ability if the interest lies in extreme events in a

mathematically sound way. Furthermore, we discussed approaches to the forecast evaluation in case of extreme ensemble predictions and demonstrated that the observation of the forecaster's dilemma is not only restricted to economics, but also often occurs when forecast evaluation in the media or public takes place.

Consider, next, connections to social psychology and political forecasts. Tetlock (2005) analyzes the forecast quality of probabilistic political forecasts over the last three decades. The author develops measures of forecast quality tailored to the specific format of these forecasts, which is particularly difficult due to the problem of determining what actually happened. In contrast to meteorology or economics, the outcome cannot be readily observed and political discussions are shaped by subjective interpretations of historical events. Tetlock (2005) finds that human experts are hardly able to outperform simple statistical extrapolation algorithms and perform slightly better than the simple approach of assigning equal probabilities to all possible outcomes. The forecast quality appears to be mainly independent of the political world views of the human competitors, but is determined by the way how forecasters think. Based on an essay by Isaiah Berlin published in 1953 (Berlin, 2009), Tetlock (2005) distinguishes between two types of forecasters. Dependent on their tendency to state extreme predictions, the experts are classified as 'hedgehogs' and 'foxes'. While the 'hedgehogs' who "know one big thing" (Tetlock, 2005, page 2) tend to state more extreme predictions, the 'foxes' who "know many little things" (Tetlock, 2005, page 2) tend to state more careful predictions.

The findings of Denrell and Fang (2010) are consistent with those of Tetlock (2005) in that the 'foxes' significantly outperform the 'hedgehogs'. Furthermore, Tetlock (2005) finds an inverse relationship between the media attention received by the human experts and the accuracy of their predictions. In addition to our observation that the media attention is focused on the performance evaluation for extreme events (see Chapter 1), it is also focused on forecasters with strong convictions stating extreme predictions. Tetlock (2005) offers psychological explanations for the attractiveness of extreme prediction for forecast consumers, the tendency of many forecasters to state extreme predictions, and the tendency of the media attention to restrict the performance evaluation to subsets of extreme observations.

Note that both Denrell and Fang (2010) and Tetlock (2005) focus on general predictive performance while the work discussed in this thesis focuses on performance evaluation if the interest lies in extreme events. The conditioning on extreme predictions in the public and media attention observed by Tetlock (2005) differs from the conditioning on extreme ensemble predictions discussed in Chapter 4. The ensemble predictions and forecasting procedures can be seen as 'neutral' physical models without any psychological tendency towards extreme predictions.

The thesis at hand suggests various starting points for further research. Thus far, we have only regarded forecasts for a single variable at a single location and a single look-ahead time. In high-impact weather situations, spatial, temporal and inter-variable coherence is of critical importance. Therefore, mathematically justifiable forecast evaluation procedures and calibration checks for these situations have to be developed. The pre-rank approach of Gneiting et al. (2008) is a first step in this direction. The proper scoring rules for extreme events discussed in this thesis might prove helpful to develop verification procedures that retain the

critical property of maximizing the sharpness subject to calibration.

Furthermore, the results of this thesis might help to develop new approaches to the forecast verification for deterministic predictions and observations of extreme events. Using the proper scoring rules for extreme events discussed in this thesis, new insights into the forecast evaluation based on contingency tables or the quality of economic forecasts provided by the Wall Street Journal Survey of Economic Forecasts might be gained.

The simple regime-switching combination of the NGR model and forecasting procedures based on EVT discussed in Chapter 4 might lead to further improvements of state-of-the-art ensemble postprocessing techniques. For a larger coherence between the ensemble predictions and the observations, even more significant improvements can be expected. The questions of how to find optimal thresholds and parameter estimates for the models based on EVT, and how to generalize this approach to other variables as temperature or pressure remain open.

To conclude this thesis, we point to the important task of communicating the findings of Gneiting and Ranjan (2011b), Diks et al. (2011) and this thesis to the scientific community and all forecast users. We call for a shift in the current forecast verification mechanisms in the "media-driven marketplace of ideas" (Tetlock, 2005, page 232). The observation that forecast verification of probabilistic forecasts for rare and extreme events should not be carried out by restricting the attention to subsets of extreme observations is relevant for all consumers of forecasts. In this thesis, we demonstrated that there are proper scoring rules for extreme events providing verification procedures that are well suited to this task.

List of symbols and abbreviations

Symbols

$(\Omega, \mathcal{A}, \mathcal{Q})$	Global space consisting of a set Ω, a σ-algebra $\mathcal{A}$ on Ω and a probability measure $\mathcal{Q}$ on the measurable space $(\Omega, \mathcal{A})$
$\mathcal{P}$	Class of probability measures
$\mathcal{L}(Y\|\mathcal{A})$	Conditional distribution of the random variable Y given the σ-Algebra $\mathcal{A}$
$\mathcal{B}$	Borel-σ-algebra on $\mathbb{R}$
Φ_{μ,σ^2}	Cumulative distribution function of a normal distribution with mean value μ and variance σ^2
$\mathbb{1}(C)$	Indicator function with condition C
$S(P, \omega)$	Value of a scoring rule S for the probabilistic forecast P if ω is observed
$\boldsymbol{S}(P, Q)$	Expected score of the probabilistic forecast P under the true distribution Q
$\mathbb{E}_Q Y$	Expected value of the random variable Y under distribution $\mathcal{Q}$, i.e. $\int X \mathrm{d}Q$
φ_{μ,σ^2}	Density function of a normal distribution with mean value μ and variance σ^2
$\mathcal{N}_{[l,u)}(\mu, \sigma^2)$	Truncated normal distribution with mean value μ and variance σ^2 restricted to the interval $[l, u)$
$\triangle_\alpha(x)$	Triangular quantile weight function which has a peak of height 1 at α and decays to 0 at $x = 0$ and $x = 1$
t_n	Test statistic of the Diebold-Mariano-type test of equal performance
σ_n^2	Estimator of the asymptotic variance of the score difference for the Diebold-Marino-type test of equal performance

Abbreviations

ASOS	North American Automated Surface Observation System
BMA	Bayesian Model Averaging
CDF	Cumulative Distribution Function
CL	Conditional Likelihood
CRPS	Continuous Ranked Probability Score
CRPS^{q}	Quantile-weighted verison of the CRPS
CRPS^{t}	Threshold-weighted version of the CRPS
CSL	Censored Likelihood
EDI	Extremal Dependence Index
EVT	Extreme Value Theory
F	False Alarm Rate
GEV	Generalized Extreme Value (distribution)
GP	Generalized Pareto (distribution)
H	Hit Rate
KLIC	Kullback-Leibler Information Criterion
kt	Knots
LinS	Linear Score
LogS	Logarithmic Score
MAE	Mean Absolute Error
ML	Maximum Likelihood
MSE	Mean Squared Error
NGR	Nonhomogeneous Gaussian Regression
NWP	Numerical Weather Prediction
PIT	Probability Integral Transform
PseudoS	Pseudospherical Score
QS_α	Quantile Score
QuadrS	Quadratic Score

Bibliography

Agresti, A. (2002), *Categorical Data Analysis*, Hoboken: John Wiley and Sons, 2nd ed.

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.

Balkema, A. A. and de Haan, L. (1974), "Residual life time at great age," *The Annals of Probability*, 2, 792–804.

Baringhaus, L. and Franz, C. (2004), "On a new multivariate two-sample test," *Journal of Multivariate Analysis*, 88, 190–206.

Bauer, H. (1992), *Maß- und Integrationstheorie*, Berlin: Walter de Gruyter, 2nd ed.

— (2002), *Wahrscheinlichkeitstheorie*, Berlin: Walter de Gruyter, 5th ed.

Beniston, M., Stephenson, D. B., Christensen, O. B., Ferro, C. A. T., Frei, C., Goyette, S., Halsnaes, K., Holt, T., Jylhä, K., Koffi, B., et al. (2007), "Future extreme events in European climate: An exploration of regional climate model projections," *Climatic Change*, 81, 71–95.

Berlin, I. (2009), *The Hedgehog and the Fox: An essay on Tolstoy's view of history*, London: Phoenix.

Bertsekas, D. P. (1995), *Nonlinear programming*, Belmont: Athena Scientific, 1st ed.

Boyd, S. P. and Vandenberghe, L. (2004), *Convex optimization*, New York: Cambridge University Press.

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.

Brockwell, A. E. (2007), "Universal residuals: A multivariate transformation," *Statistics and Probability Letters*, 77, 1473–1478.

Chang, E. K. M. and Fu, Y. (2002), "Interdecadal variations in Northern Hemisphere winter storm track intensity," *Journal of Climate*, 15, 642–658.

Coelho, C. A. S., Ferro, C. A. T., Stephenson, D. B., and Steinskog, D. J. (2008), "Methods for exploring spatial and temporal variability of extreme events in climate data," *Journal of Climate*, 21, 2072–2092.

Coles, S. (2001), *An introduction to statistical modeling of extreme values*, London: Springer Verlag.

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

— (1998), "Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design," Research Report 139, University College London, Dept. of Statistical Science.

— (2007), "The geometry of proper scoring rules," *Annals of the Institute of Statistical Mathematics*, 59, 77–93.

Denrell, J. and Fang, C. (2010), "Predicting the next big thing: Success as a signal of poor judgment," *Management Science*, 56, 1653–1667.

Deutscher Wetterdienst (2009), "Wie gut sind Wettervorhersagen? Qualitätsprüfung beim DWD," Pressestelle des DWD, available online at `http://www.dwd.de/bvbw/generator/DWDWWW/Content/Presse/Broschueren/Verifikation__PDF,templateId=raw,property=publicationFile.pdf/Verifikation_PDF.pdf`.

Diaz, H. F. and Murnane, R. J. (2008), "The significance of weather and climate extremes to society: An introduction," in *Climate Extremes and Society*, eds. Diaz, H. F. and Murnane, R. J., Cambridge University Press, pp. 1–8.

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.

Dlugolecki, A. (2008), "An overview of the impact of climate change on the insurance industry," in *Climate Extremes and Society*, eds. Diaz, H. F. and Murnane, R. J., Cambridge University Press, pp. 248–278.

Doswell, C., 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.

Eckel, F. A. and Mass, C. F. (2005), "Aspects of effective mesoscale, short-range ensemble forecasting," *Weather and Forecasting*, 20, 328–350.

Ekeland, I. and Temam, R. (1976), *Convex Analysis and Variational Problems*, Amsterdam: North-Holland Publ. Comp.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997), *Modelling extremal events for insurance and finance*, Berlin: Springer Verlag, 2nd ed.

Ferro, C. A. T. (2007), "A probability model for verifying deterministic forecasts of extreme events," *Weather and Forecasting*, 22, 1089–1100.

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.

Finley, J. P. (1884), "Tornado predictions," *American Meteorological Journal*, 1, 85–88.

Fisher, R. A. and Tippett, L. H. C. (1928), "Limiting forms of the frequency distribution of the largest or smallest member of a sample," in *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, vol. 24, pp. 180–190.

Friederichs, P., Gober, M., Bentzien, S., Lenz, A., and Krampitz, R. (2009), "A probabilistic analysis of wind gusts using extreme value statistics," *Meteorologische Zeitschrift*, 18, 615–629.

Friederichs, P. and Thorarinsdottir, T. L. (2012), "Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction," *arXiv:1204.1022*.

Galiatsatou, P., Prinos, P., and Sanchez-Arcilla, A. (2008), "Estimation of extremes: Conventional versus Bayesian techniques," *Journal of Hydraulic Research*, 46, 211–223.

Gandin, L. S. and Murphy, A. H. (1992), "Equitable skill scores for categorical forecasts," *Monthly Weather Review*, 120, 361–370.

Ghelli, A. and Primo, C. (2009), "On the use of the extreme dependency score to investigate the performance of an NWP model for rare events," *Meteorological Applications*, 16, 537–544.

Gilbert, G. K. (1884), "Finley's tornado predictions," *American Meteorological Journal*, 166–172.

Gneiting, T. (2008), "Editorial: Probabilistic Forecasting," *Journal of the Royal Statistical Society, Series A*, 171, 319–321.

— (2010), "Statistische Entscheidungs- und Vorhersagetheorie," Lecture, University of Heidelberg, unpublished.

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

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

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

— (2007), "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, 102, 359–378.

Gneiting, T. and Ranjan, R. (2011a), "Combining predictive distributions," *arXiv:1106.1638*.

— (2011b), "Comparing density forecasts using threshold-and quantile-weighted scoring rules," *Journal of Business and Economic Statistics*, 29, 411–422.

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.

Hall, S. S. (2011), "Scientists on trial: At fault?" *Nature*, 477, 264–269.

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.

Heidke, P. (1926), "Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst," *Geografiska annaler*, 8, 301–349.

Hendrickson, A. D. and Buehler, R. J. (1971), "Proper scores for probability forecasters," *The Annals of Mathematical Statistics*, 42, 1916–1921.

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

Hoff, P. D. (2009), *A first course in Bayesian statistical methods*, New York: Springer.

Hogan, R. J., O'Connor, E. J., and Illingworth, A. J. (2009), "Verification of cloud-fraction forecasts," *Quarterly Journal of the Royal Meteorological Society*, 135, 1494–1511.

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*, 34, 301–310.

Johnson, B. W. (1991), "On the admissibility of improper Bayes inferences in fair Bayes decision problems," Ph.D. thesis, University of Minnesota.

Jolliffe, I. T. and Stephenson, D. B. (eds.) (2003), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Hoboken: John Wiley and Sons.

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

Livezey, R. E. (2003), "Categorical events," in *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, eds. Jolliffe, I. T. and Stephenson, D. B., John Wiley and Sons, pp. 77–96.

Marzban, C. (1998), "Scalar measures of performance in rare-event situations," *Weather and Forecasting*, 13, 753–763.

Mason, I. B. (2003), "Binary events," in *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, eds. Jolliffe, I. T. and Stephenson, D. B., John Wiley and Sons, pp. 37–76.

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

Milly, P. C. D., Wetherald, R. T., Dunne, K. A., and Delworth, T. L. (2002), "Increasing risk of great floods in a changing climate," *Nature*, 415, 514–517.

Murphy, A. H. (1993), "What is a good forecast? An essay on the nature of goodness in weather forecasting," *Weather and Forecasting*, 8, 281–293.

— (1996), "The Finley affair: A signal event in the history of forecast verification," *Weather and Forecasting*, 11, 3–20.

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

Pickands, J. (1975), "Statistical inference using extreme order statistics," *The Annals of Statistics*, 3, 119–131.

Primo, C. and Ghelli, A. (2009), "The affect of the base rate on the extreme dependency score," *Meteorological Applications*, 16, 533–535.

R Development Core Team (2010), "R: A language and environment for statistical computing," *R Foundation for Statistical Computing, Vienna, Austria.*

Reiss, R.-D. and Thomas, M. (2007), *Statistical Analysis of Extreme Values*, Basel: Birkhäuser Verlag AG, 3rd ed.

Rosenblatt, M. (1952), "Remarks on a multivariate transformation," *The Annals of Mathematical Statistics*, 23, 470–472.

Schär, C., Vidale, P. L., Lüthi, D., Frei, C., Häberli, C., Liniger, M. A., and Appenzeller, C. (2004), "The role of increasing temperature variability in European summer heatwaves," *Nature*, 427, 332–336.

Sloughter, J. M. L., 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.

Smith, R. L. (1985), "Maximum likelihood estimation in a class of nonregular cases," *Biometrika*, 72, 67–90.

Stephenson, A. and Tawn, J. (2004), "Bayesian inference for extremes: Accounting for the three extremal types," *Extremes*, 7, 291–307.

Stephenson, D. B. (2008), "Definition, diagnosis, and origin of extreme weather and climate events," in *Climate Extremes and Society*, eds. Diaz, H. F. and Murnane, R. J., Cambridge University Press, pp. 11–23.

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.

Swets, J. A. (1988), "Measuring the accuracy of diagnostic systems," *Science*, 240, 1285–1293.

Tetlock, P. E. (2005), *Expert political judgment: How good is it? How can we know?*, Princeton: Princeton University Press.

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*, 173, 371–388.

Thorarinsdottir, T. L. and Johnson, M. S. (2012), "Probabilistic wind gust forecasting using non-homogeneous Gaussian regression," *Monthly Weather Review*, 140, 889–897.

Erklärung

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

Datum Unterschrift