Focusing on regions of interest in forecast evaluation

Hajo Holzmann*
Fachbereich Mathematik und Informatik
Philipps-Universität Marburg
holzmann@mathematik.uni-marburg.de

Bernhard Klar
Institut für Stochastik
Karlsruher Institut für Technologie (KIT)
bernhard.klar@kit.edu

July 14, 2017

Abstract

Often, interest in forecast evaluation focuses on certain regions of the whole potential range of the outcome, and forecasts should mainly be ranked according to their performance within these regions. A prime example is risk management, which relies on forecasts of risk measures such as the value-at-risk or the expected shortfall and hence requires appropriate loss distribution forecasts in the tails. Further examples include weather forecasts with a focus on extreme conditions, or forecasts of environmental variables such as ozone with a focus on concentration levels with adverse health effects.

In this paper we show how weighted scoring rules can be used to this end, and in particular that they allow to rank several potentially misspecified forecasts objectively with the region of interest in mind. This is demonstrated in various simulation scenarios. We introduce desirable properties of weighted scoring rules and present general construction principles based on conditional densities or distributions and on scoring rules for probability forecasts. In our empirical application to log-return time series all forecasts seem to be slightly misspecified, as is often unavoidable in practice, and no method performs best overall. However, using weighted scoring functions the best method for predicting losses can be identified, which is hence the method of choice for the purpose of risk management.

Keywords. financial time series, predictive performance, probabilistic forecast, locally proper weighted scoring rule, misspecified forecast, rare and extreme events, risk management

1 Introduction

Generating and evaluating forecasts is a central task in many scientific disciplines such as makroeconomics and finance (Elliott and Timmermann, 2016) or climate and weather research (Casati et al.,

*Corresponding author. Prof. Dr. Hajo Holzmann, Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, Hans-Meerweinstr., 35043 Marburg, Germany

2008). While point forecasts for parameters such as the mean or a quantile are more frequently issued (Gneiting, 2011), probabilistic forecasts for the whole predictive distribution are most informative and generally preferable (Dawid, 1984). Comparisons of distinct forecasts should be based on proper scoring rules (Gneiting and Raftery, 2007), which encourage the forecaster to be honest and make careful assessments according to her true believes.

Often, interest focuses on certain regions of the whole potential range of the outcome. As a consequence, forecasts should mainly or even exclusively be ranked according to their performance within these regions, while outside they are only of minor or no interest.

A prime example is short-term risk management, which relies on forecasts of risk measures such as the value-at-risk or the expected shortfall (Nolde and Ziegel, 2017; McNeil, Frey and Embrechts, 2005) that summarize tail properties of the loss distribution. Hence forecasts of risk measures typically are preceded by forecasts of the profit-and-loss distribution, the quality of which should therefore be assessed in its lower tail. For regulatory purposes and in particular the evaluation of capital requirements, it is finally the value of the risk measure itself that matters. However, the overall quality of the forecast of the profit-and-loss distribution, in particular in its lower tail, is also of interest, and rankings of distinct forecasting schemes that rely directly on the loss distribution thus do not depend on the choice of the risk measure, be it VaR, expected shortfall or expectile (Holzmann and Klar, 2017b).

Examples for the use of weighted scoring rules in a risk-management context are De Nicolò and Lucchetta (2017), who evaluate the performance of multi-period forecasts of indicators of real and financial risks over the left tail, as well as Opschoor et al. (2017), who use focused scoring rules in the context of measuring downside risk in equity markets.

In our empirical application we consider daily log-returns of the S&P 500 index as well as the Deutsche Bank stock over the period from January 1, 2009 until December 31, 2016. During this time span, for the Deutsche Bank series about 10% of the log-returns are below -3%. Fig. 1 shows the series from 2009 - 2011, which includes the volatile period after the financial crisis with large negative, but also large positive returns. For the purpose of risk management and the computation of risk measures, accurate forecasts of the lower tail of the distribution below say -3% are paramount. Choosing the forecasting method directly based on the loss distribution with emphasis on the lower tail allows flexibility concerning the subsequent choice of the risk measure.

Another economic example is the evaluation of inflation forecasts (Gneiting and Ranjan, 2011) when taking into account inflation targets. For example, the Bank of England sets the inflation target at 2%, in case of inflation rates below 1% as well as above 3% it must write an open letter to the Chancellor of the Exchequer (Bank of England, 2017). Thus, if emphasis in inflation forecast evaluation is on missing the target by more than 1%, it would be natural to focus on the two-sided range outside the interval from 1% to 3%.

As for GDP growth, China sets a minimum of 6.5% in its 13th five-year plan from 2015 (Giesbergen, 2017), whereas most developed countries do not fix formal targets for GDP growth rates. However, nominal GDP targeting is discussed as a potential policy rule in the popular press, where sometimes rates between 2% and 4% are considered ideal for sustained growth in developed countries. Markedly lower growth rates indicate recessions, while much higher rates may indicate some kind of bubble.

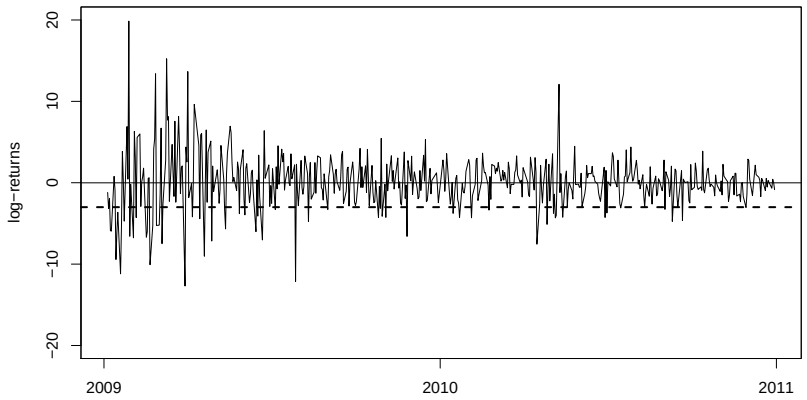

Figure 1: Log returns of Deutsche Bank shares, January 1, 2009 - December 31, 2010

Moreover, nominal GDP targeting has recently gained attention among academics, as can be seen in Garín et al. (2016) or Billi (2016). Thus, there may be reasons for evaluating GDP forecasts with focus on specific regions.

There is also an interest in evaluating weather forecasts with a focus on severe weather conditions like extreme winds, temperatures or rainfall. For example, Haiden et al. (2014) consider 10-metre wind speeds above the 98% quantile of the climatology corresponding to a threshold of 16ms^{-1}.

Further, in environmental science, Pisoni et al. (2011) state that "over-threshold event forecasting is of paramount importance in the monitoring of environmental variables, such as those related to air pollution". For example, for Ozone gas concentrations at ground level the threshold is set as the maximum daily eight-hour average, with a target value of 120 $\mu g/m^3$ by EU Directive 2008/50. Formally, the target value should not be exceeded on more than 25 days a year, but evidently the magnitude of exceedance should also be taken into account to judge the damage to human health.

To accommodate forecast comparisons with scoring rules by including regions of interest, Amisano and Giacomini (2007) introduced a weighted version of the logarithmic score $S(p, x) = -w(x) \log p(x)$, where $w(x)$ is the weight function such as $w(x) = 1\{x \in A\}$ for some set A, and $p(x)$ is the forecast density. However, as observed in Diks et al. (2011) and Gneiting and Ranjan (2011), this is not a proper scoring rule. Indeed, it favors forecasts which put more mass into the region of interest than does the true conditional distribution. As a remedy, Diks et al. (2011) proposed the conditional and the censored likelihood rules, which depend on weight functions but are proper scoring rules, while Gneiting and Ranjan (2011) developed weighted versions of the continuous ranked probability score (CRPS). Pelenis (2014) defined and discussed relevant theoretical properties of weighted scoring rules.

In this paper we propose a general construction principle for strictly locally proper weighted scoring rules based on conditional densities or distributions and on scoring rules for probability forecasts. We show how the likelihood-based weighted scoring rules from Diks et al. (2011) and Pelenis (2014) fit into this framework and how they are related. Further, our method gives rise to strictly locally proper weighted versions of the continuous ranked probability score and more general potentially multivariate energy scores.

We further argue that weighted scoring rules are mainly useful for comparing distinct misspecified forecasts. If interest focuses on a region A, a weighted scoring rule allows to ignore possible problems or advantages of the density forecast outside of A. Thus, even if a density forecast performs poorly outside of A but well on A, it is useful to us if we only focus on the region A, indeed as useful as another density forecast which performs well overall. Slightly misspecified forecasts are certainly the rule rather than the exception (Patton, 2017). For example, in our empirical study we use GARCH(1,1)-models for the log-return series with normal, t and skew-t innovation distributions, all of which seem to be slightly misspecified. In this simple situation it might be possible to achieve better fits and predictions with more sophisticated models, but in more complex settings one is typically confined to a small set of benchmark models. When focusing on regions of interest such misspecified models can still be ranked in a reasonable way for the application at hand.

On a more formal level, for hypothesis testing based in score differences (Diebold and Mariano, 1995), we argue that using a weighted scoring rule introduces a censoring mechanism, in which the form of the density is irrelevant outside the region of interest. For the resulting testing problem with composite null - and alternative hypotheses based on i.i.d. observations, the optimal test uses score differences based on the censored likelihood rule of Diks et al. (2011), see Holzmann and Klar (2016).

The paper is organised as follows. After a motivating illustration, in Section 2 we present our theoretical results on the construction of weighted scoring rules, and briefly discuss the relation to censoring and hypothesis testing. Section 3 illustrates the findings in a simulation study, while Section 4 gives an empirical application to financial time series data. Section 5 concludes. Some proofs are given to Section 7, while further technical details, examples and simulation results are deferred to the supplementary material Holzmann and Klar (2017a).

2 Weighted scoring rules

2.1 Motivation

Let us first illustrate the use of weighted scoring rules in Diebold-Mariano tests for equal forecast performance (for details, see Subsection 2.4), and how they allow to focus interest on subregions $\{x \geq r\}$, $r > 0$ some fixed threshold, of the whole potential domain of the outcome variable x. For example, observations could correspond to losses, and an investment bank or financial corporation wants to predict losses or extreme losses in their portfolio for regulatory purposes. Consider the following stylized simulation scenario, where the aim is to decide between two competing forecasts.

Scenario A: Forecast 1: F_{hlt} vs. Forecast 2: F_{hrt}.

Here, F_{hlt} denotes a piecewise defined distribution with continuous density and heavy left tail, consisting of a scaled t_4-distribution on $(-\infty, 0]$ and a standard normal distribution on $(0, \infty)$. Conversely, F_{hrt} denotes a piecewise defined distribution with continuous density and heavy right tail, consisting of a standard normal distribution on $(-\infty, 0]$ and a scaled t_4-distribution on $(0, \infty)$. The data-generating process is given by independent standard normally distributed observations with sample size n.

We apply the two-sided Diebold-Mariano test of equal predictive performance, nominal level $\alpha = 0.05$, based on the following scoring rules. If p denotes the predictive density, we employ first the standard

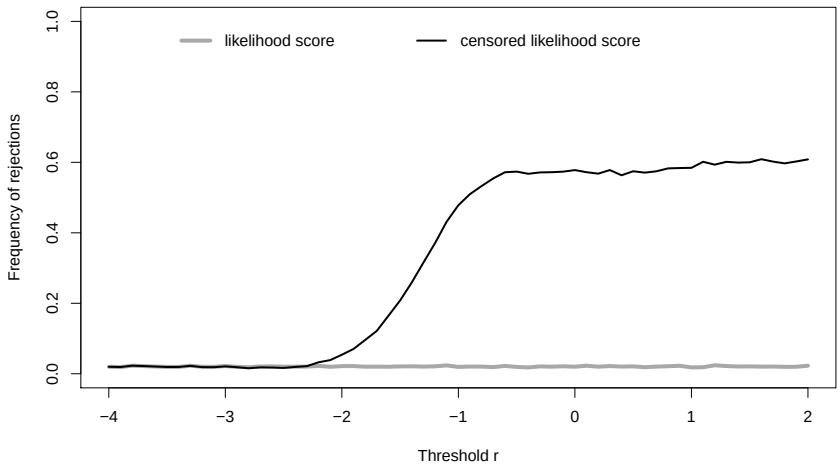

Figure 2: Scenario A. The null hypothesis of equal predictive performance of F_{hlt} and F_{hrt} is tested under a standard normal population. The plot shows the frequency of rejections in two-sided Diebold-Mariano test in favor of F_{hlt} for the likelihood and the censored likelihood scoring rule for sample size $n = 100$.

logarithmic score $S_l(p, x) = -\log p(x)$, which of course does not depend on any threshold. Second we use the censored likelihood rule from Diks et al. (2011) at threshold r, that is, with weight function $w(x) = 1\{x \geq r\}$, defined by

$$S^{CSL}(p, x; r) = \begin{cases} -\log p(x), & \text{if} \quad x \geq r, \\ -\log\left(1 - \int_{-\infty}^{r} p(z)\mathrm{d}z\right), & \text{if} \quad x < r. \end{cases}$$

Note that $S^{CSL}(p, x; r)$ takes into account the form of the density p only for observations $x \geq r$ above the threshold, below it relies on the total mass $\int_{-\infty}^{r} p(z)\mathrm{d}z$.

Fig. 2 shows the proportion of rejections of the null hypothesis of equal predictive performance in favor of F_{hlt} as a function of the threshold value r. Without restricting attention to a subregion of interest, i.e. for $r = -\infty$, by symmetry both forecasts equally strongly deviate from the (true) standard normal distribution, and neither of them should be rejected in favor of the other. However, for $r > 0$, Forecast 1 coincides with the standard normal distribution, and Forecast 2 should be rejected. The rejection frequencies in favor of F_{hlt} of the logarithmic scoring rule is around 0.025, as this is the proportion of tests that reject the null hypothesis at the 5% level and that additionally have a test statistic that indicates that F_{hlt} is better. Clearly, the rejection frequency of 0.025 for the logarithmic scoring rule does not depend on r.

In contrast, the rejection frequencies of the censored likelihood rule increase for threshold values larger than -2.5, and reach a plateau for values larger than -0.5 at a rejection level of about 0.6. The censored likelihood rule thus allows to focus on the region of interest $\{x \geq r\}$, where for r close to or greater than zero the forecast F_{hlt} is evidently preferable over F_{hrt}. We will take a closer look on this example in section 3.

2.2 Weighted scoring rules for density forecasts

In economic applications, point forecasts such as mean or quantiles are most prominent, but if the target is a complete forecast distribution, it is issued in the form of a density forecast (Elliott and Timmermann, 2016).

Thus, in this section we investigate weighted scoring rules for density forecasts, and consider the general case in the next subsection. We shall work over an abstract measurable space $(\mathcal{X}, \mathcal{F})$, endowed with a σ-finite measure μ and consider a family $\mathcal{P}$ of probability densities w.r.t. μ on $(\mathcal{X}, \mathcal{F})$. Continuous observations are the main example, where $\mathcal{X}$ corresponds to the real numbers or to $\mathbb{R}^d$, and where μ is the Lebesgue measure. However, an at most countable set $\mathcal{X}$ formally endowed with counting measure also fits into this framework.

In terms of density forecasts, a *scoring rule* is a map $S : \mathcal{P} \times \mathcal{X} \to \overline{\mathbb{R}}$, where we denote $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, for which for every $p \in \mathcal{P}$ the map $x \mapsto S(p, x)$ is quasi-integrable for every $q \in \mathcal{P}$, and for which

$$S(p, q) = \int_{\mathcal{X}} S(p, x)\, q(x)\, \mathrm{d}\mu(x) > -\infty \quad \text{and} \quad S(q, q) \in \mathbb{R}$$

for every $p, q \in \mathcal{P}$. A scoring rule is called *proper* if

$$S(p, q) \geq S(q, q), \quad q, p \in \mathcal{P}, \tag{1}$$

and it is called *strictly proper* if it is proper and if there is equality in (1) if and only if $p = q$ μ-almost everywhere. Note the normalization: $S(p, x)$ denotes the loss, and we aim to minimize the expected loss.

We shall consider scoring rules which depend on weight functions, i.e. measurable functions $w : \mathcal{X} \to [0, 1]$, and use notation and terminology which is closely related to that of Pelenis (2014). Write $S(p, x; w)$, so that a *weighted scoring rule* is a map $S : \mathcal{P} \times \mathcal{X} \times \mathcal{W} \to \overline{\mathbb{R}}$ such that $S(\cdot, \cdot; w)$ is a scoring rule for each $w \in \mathcal{W}$, where $\mathcal{W}$ is a set of weight functions. The weighted scoring rule is called *localizing* if

$$S(h, x; w) = S(p, x; w) \quad \text{for } \mu - \text{a.e. } x \in \mathcal{X} \quad \text{if} \quad p = h \ \mu - \text{a.e. on } \{w > 0\},\ p, h \in \mathcal{P}, \tag{2}$$

where we use the notation $\{w > 0\} = \{x \in \mathcal{X} :\ w(x) > 0\}$. Thus, a localizing weighted scoring rule only depends on the values of the forecast densities on the set $\{w > 0\}$ for each $w \in \mathcal{W}$. Integrating (2) we find that

$$S(h, q; w) = S(p, q; w) \quad \text{if} \quad p = h \ \mu - \text{a.e. on } \{w > 0\},\ p, q, h \in \mathcal{P}.$$

In particular, if the localizing weighted scoring rule S is also proper, i.e. $S(\cdot, \cdot; w)$ is proper for each $w \in \mathcal{W}$, it is called a *localizing proper* weighted scoring rule, implying

$$S(h, q; w) \geq S(q, q; w) = S(p, q; w) \quad \text{if} \quad p = q \ \mu - \text{a.e. on } \{w > 0\},\ p, q, h \in \mathcal{P}. \tag{3}$$

Further, a localizing proper weighted scoring rule is called *strictly locally proper* if $S(p, q; w) = S(q, q; w)$ implies $p = q$ on $\{w > 0\}$ μ-a.e., $p, q \in \mathcal{P}$, and it is called *proportionally locally proper* if $S(p, q; w) = S(q, q; w)$ if and only if $p = c\, q$ on $\{w > 0\}$ μ-a.e., for some constant $c > 0$ which

depends on $p, q \in \mathcal{P}$. Let us stress that *strictly locally proper* is not a special case of *proportionally locally proper*, these properties for a localizing proper weighted scoring rule are mutually exclusive. Note that Pelenis (2014) does not use the pointwise concept of a localizing weighted scoring rule as in (2), but rather takes (3) as starting point. Our requirement (2) is natural, however, and is indeed satisfied for the rules discussed below.

Next we shall construct weighted scoring rules which satisfy the properties defined above. To this end, assume that the class of densities $\mathcal{P}$ and the class of weight functions $w \in \mathcal{W}$ are such that

$$\int_{\mathcal{X}} p(x)\, w(x)\, \mathrm{d}\,\mu(x) =: \int pw > 0.$$

For $p \in \mathcal{P}$, $w \in \mathcal{W}$ we let

$$p_w(x) = \frac{w(x)\, p(x)}{\int w\, p}$$

denote the renormalized density of p w.r.t. w. For formulating the next result, let $\widetilde{\mathcal{P}}$ be another class of densities such that $p_w \in \widetilde{\mathcal{P}}$ for every $w \in \mathcal{W}$, $p \in \mathcal{P}$. We show how to construct proportionally locally proper weighted scoring rules from strictly proper scoring rules. Gneiting (2011), Theorem 5, has a version of this result for scoring functions for evaluating forecasts of certain functionals. This connection is further discussed in Example 2.

Theorem 1. *Let $\widetilde{S} : \widetilde{\mathcal{P}} \times \mathcal{X} \to \overline{\mathbb{R}}$ be a proper scoring rule. Then*

$$S : \mathcal{P} \times \mathcal{X} \times \mathcal{W} \to \overline{\mathbb{R}}, \quad S(p, x; w) = w(x)\, \widetilde{S}(p_w, x)$$

is a localizing proper weighted scoring rule. Further, if $\widetilde{S}$ is strictly proper, then S is proportionally locally proper.

The reason why the scoring rule S can be at most proportionally locally proper is that the weighted density p_w depends on p only up to proportionality on the set $\{w > 0\}$.

Example 1. If applied to the logarithmic score $S_l(p, x) = -\log p(x)$, Theorem 1 yields

$$\begin{aligned} S_l(p, x; w) &= -w(x)\, \log p_w(x) \\ &= -w(x)\, \log p(x) + w(x)\, \log\Big(\int p\, w\Big) - w(x)\, \log w(x) \\ &= S^{CL}(p, x; w) - w(x)\, \log w(x), \end{aligned} \tag{4}$$

the conditional likelihood rule suggested by Diks et al. (2011) up to a normalizing term which does not depend on the forecast density p. Here, we set $0 \log(0) = 0 \log(\infty) = 0$. It is remarkable that even though evaluation of the conditional likelihood rule S^{CL} requires evaluation of the integral $\int p\, w$, which in case of $w(x) = 1\{x \in A\}$ amounts to the probability $P(A)$ under p, this scoring rule is only proportionally locally proper and thus insensitive to this probability. Theorem 1 can also be applied to the Hyvärinen score from Hyvärinen (2005), see the supplementary material Holzmann and Klar (2017a) for further details. ◇

For probabilistic forecasts, localizing proper weighted scoring rules provide a practical way of addressing the forecasters dilemma, which refers to instances in which forecasts are only evaluated when

extreme events occur (Lerch et al., 2016). These rules should be constructed to be not only proportionally locally proper as in Theorem 1, but rather strictly locally proper as in Theorem 2 below. Lerch et al. (2016, p. 3) further state that for point forecasts *there is no obvious way to abate the forecaster's dilemma by adapting existing forecast evaluation methods appropriately*. In the following example we discuss this issue, and start by relating Theorem 1 to the evaluation of point forecasts.

Example 2 (*Regions of interest of functionals*)**.** Suppose that the aim is to predict a functional $T : \mathcal{P} \to \mathbb{R}$, such as the mean. A scoring function $\mathcal{S}(t, x)$ is consistent for T if

$$\mathcal{S}\big(T(p), p\big) \leq \mathcal{S}\big(t, p\big) \quad \forall\, t \in T(\mathcal{P}),\ p \in \mathcal{P},$$

and it is strictly consistent if it is consistent and there is equality only if $t = T(p)$. Gneiting (2011), Theorem 3, points out that if $\mathcal{S}$ is consistent for T, then $S_T(p, x) = \mathcal{S}\big(T(p), x\big)$ is a proper scoring rule for the density forecast p. In this fashion, Theorem 1 can also be applied to scoring functions, the formal result being Theorem 5 in Gneiting (2011).
One could think about this as applying the original functional T not to p but to the weighted density p_w. For example, if T is the mean and $w(x) = 1\{x \geq r\}$ we would focus interest still on the mean, but of the conditional distribution above the threshold r. However, this approach is not feasible in practice since the forecaster specifies $T(p)$ only rather than p itself, and so in general the evaluator is unable to find p_w. ⋄

Proportionally locally proper weighted scoring rules do not evaluate the normalization constant $\int pw$. However, they can be turned into strictly locally proper weighted scoring rules by adding a weighted scoring rule based on probability forecasts, as shown in the following theorem.

Theorem 2. *Let* $\mathbf{s}(\alpha, z)$ *be a strictly proper scoring rule for the success probability* $\alpha \in (0, 1)$ *of a binary outcome variable* $z \in \{0, 1\}$*. Then*

$$S_{\mathbf{s}}(p, x; w) = w(x)\, \mathbf{s}\Big(\int pw, 1\Big) + \big(1 - w(x)\big)\, \mathbf{s}\Big(\int pw, 0\Big) \tag{5}$$

is a localizing proper weighted scoring rule for the density forecast p*. Further, if* $S(p, x; w)$ *is a proportionally locally proper weighted scoring rule, then*

$$\widehat{S}(p, x; w) = S_{\mathbf{s}}(p, x; w) + S(p, x; w)$$

is strictly locally proper.

Selecting different scoring rules $\mathbf{s}(\alpha, z)$ in Theorem 2 yields various ways to turn a proportionally locally proper weighted scoring rule such as the conditional likelihood rule S^{CL} into a strictly locally proper weighted scoring rule. Let us illustrate the choices used in the literature to modify S^{CL}.

Example 3. The scoring rule for a binary outcome defined by

$$\bar{\mathbf{s}}(\alpha, z) = -z\,\big(\log \alpha + 1\big) + \alpha, \quad \alpha \in (0, 1), \tag{6}$$

is strictly proper. To see this, let

$$\bar{\mathbf{s}}(\alpha, \beta) = \beta\, \bar{\mathbf{s}}(\alpha, 1) + (1 - \beta)\, \bar{\mathbf{s}}\big(\alpha, 0\big), \quad \beta \in (0, 1).$$

Then, we have that

$$\bar{\mathbf{s}}(\alpha,\beta)-\bar{\mathbf{s}}(\beta,\beta)=\beta\left(\frac{\alpha}{\beta}-1-\log(\alpha/\beta)\right)\geq 0,$$

since $\log x \leq x-1$, with equality if and only if $x=1$, that is $\alpha=\beta$. Moreover,

$$S_{\bar{\mathbf{s}}}(p,x;w)=-w(x)\big(\log\int wp\big)-w(x)+\int w\,p,$$

and a simple computation shows that

$$S_{\bar{\mathbf{s}}}(p,x;w)+S^{CL}(p,x;w)=S^{PWL}(p,x;w)$$

where

$$S^{PWL}(p,x;w)=-w(x)\,\log p(x)-w(x)+\int p\,w,$$

the penalized weighted likelihood rule by Pelenis (2014). It has the attractive property of being linear in the weight function. Hence, if density forecasts p,q are compared and p is preferred over q in terms of the PWL score for both weight functions w_1 and w_2, then it is also preferred for the weight function w_1+w_2. This is coined *preference preserving* by Pelenis (2014). ⋄

Example 4. For the logarithmic scoring rule $\mathbf{s}_l(\alpha,z)=-z\log\alpha-(1-z)\,\log(1-\alpha)$ for a binary outcome we have that

$$\mathbf{s}_l(\alpha,z)=\bar{\mathbf{s}}(\alpha,z)+\bar{\mathbf{s}}(1-\alpha,1-z),$$

where $\bar{s}(\alpha,z)$ is defined in (6), and one obtains the censored likelihood rule of Diks et al. (2011),

$$\begin{aligned}S^{CL}(p,x;w)+S_{\mathbf{s}_l}(p,x;w)&=S^{PWL}(p,x;w)+S_{\bar{\mathbf{s}}}(p,x;1-w)\\&=-w(x)\,\log p(x)-\big(1-w(x)\big)\,\log\big(1-\int wp\big)\\&=S^{CSL}(p,x;w).\end{aligned}\tag{7}$$

The penalized likelihood rule by Pelenis (2014) is "between" the conditional and the censored likelihood rules in terms of average score differences, as follows. Let $p,q,h\in\mathcal{P}$, and assume that $p=q\ \mu$ a.e. on $\{w>0\}$. Then

$$\begin{aligned}S^{CSL}(h,q;w)-S^{CSL}(p,q;w)\geq&S^{PWL}(h,q;w)-S^{PWL}(p,q;w)\\\geq&S^{CL}(h,q;w)-S^{CL}(p,q;w),\end{aligned}$$

where both inequalities are strict if and only if $\int pw\neq\int hw$. We shall further compare their behaviour in the simulation section. ⋄

2.3 Weighted scoring rules: The general case and applications to the continuous ranked probability score

The continuous ranked probability score (CRPS) is a strictly proper scoring rule which uses distribution function forecasts rather then density forecasts. It has become a widely used tool in climatological

and weather forecasting, e.g. for statistical postprocessing of forecast ensembles (Gneiting et al., 2005; Thorarinsdottir and Gneiting, 2010). See also Casati et al. (2008) for an overview.

In order to develop and discuss weighted versions of the CRPS and more general energy scores (Gneiting and Raftery, 2007), we introduce a framework which considers forecast distributions rather than just density forecasts.

To this end, let $\mathcal{M}$ be a family of distributions, e.g. probabilities on the observational space $(\mathcal{X}, \mathcal{F})$. Call a *scoring rule* a map $S : \mathcal{M} \times \mathcal{X} \to \overline{\mathbb{R}}$, for which for every $P \in \mathcal{M}$ the map $x \mapsto S(P, x)$ is quasi-integrable for every $Q \in \mathcal{M}$, and for which

$$S(P, Q) = \int_{\mathcal{X}} S(P, x) \, \mathrm{d}Q(x) > -\infty \quad \text{and} \quad S(Q, Q) \in \mathbb{R}$$

for every $P, Q \in \mathcal{M}$. It is *proper* if $S(P, Q) \geq S(Q, Q)$, $P, Q \in \mathcal{M}$, and *strictly proper* if there is equality if and only if $P = Q$. Once again, for a family of weight functions $\mathcal{W}$, a *weighted scoring rule* is a map $S : \mathcal{M} \times \mathcal{X} \times \mathcal{W} \to \overline{\mathbb{R}}$ such that $S(\cdot, \cdot; w)$ is a scoring rule for each $w \in \mathcal{W}$. In this context, and slightly deviating from (2), we call S *localizing* if for any $P, Q \in \mathcal{M}$,

$$\forall\, F \in \mathcal{F} : \; P\big(\{w > 0\} \cap F\big) = Q\big(\{w > 0\} \cap F\big) \implies S(P, x; w) = S(Q, x; w) \text{ for all } x \in \mathcal{X}, \tag{8}$$

the condition meaning that the restrictions of P and Q to $\{w > 0\}$ coincide.

A localizing proper weighted scoring rule is called *strictly locally proper* if $S(P, Q; w) = S(Q, Q; w)$ already implies that the restrictions of P and Q to $\{w > 0\}$ coincide, $P, Q \in \mathcal{M}$, and it is called *proportionally locally proper* if $S(P, Q; w) = S(Q, Q; w)$ if and only if for all $F \in \mathcal{F}$, $P\big(\{w > 0\} \cap F\big) = c\, Q\big(\{w > 0\} \cap F\big)$ for some constant $c > 0$ which depends on $P, Q \in \mathcal{M}$.

In this more general framework, the statements of Theorem 1 and its application to functionals as in Example 2, as well as the second part of Theorem 2 remain valid. To formulate the result, assume that for all $w \in \mathcal{W}$ and $P \in \mathcal{M}$ we have that $\int w \, \mathrm{d}P > 0$, and set

$$\mathrm{d}P_w(x) = \frac{w(x) \, \mathrm{d}P(x)}{\int w \, \mathrm{d}P},$$

the probability distribution with density proportional to w w.r.t. P, which is assumed to belong to the family $\widetilde{\mathcal{M}}$.

Theorem 3. *(i) Let $\widetilde{S} : \widetilde{\mathcal{M}} \times \mathcal{X} \to \overline{\mathbb{R}}$ be a proper scoring rule. Then*

$$S : \mathcal{M} \times \mathcal{X} \times \mathcal{W} \to \overline{\mathbb{R}}, \quad S(P, x; w) = w(x) \, \widetilde{S}(P_w, x)$$

is a localizing proper weighted scoring rule. Further, if $\widetilde{S}$ is strictly proper, then S is proportionally locally proper.

(ii) If $S_{\mathbf{s}}(P, x; w)$ is as in (5) with $\int pw$ replaced by $\int w \, \mathrm{d}P$, and if $S(P, x; w)$ is a proportionally locally proper weighted scoring rule, then

$$\widehat{S}(P, x; w) = S_{\mathbf{s}}(P, x; w) + S(P, x; w)$$

is strictly locally proper.

The proof of this theorem is deferred to the supplementary material Holzmann and Klar (2017a).

Example 5. On the real line we identify probability distributions P with the associated distribution functions $F(x) = P\big((-\infty, x]\big)$, $x \in \mathbb{R}$. Now, for a family of distribution functions $\mathcal{P}$ on the real line with finite first moment, the continuous ranked probability score (CRPS) is given by

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} \big(F(z) - 1\{x \le z\}\big)^2 \,\mathrm{d}z, \quad F \in \mathcal{P}, \tag{9}$$

see Matheson and Winkler (1976), and it can be evaluated as

$$\mathrm{CRPS}(F, x) = \mathrm{E}_F\big|x - X\big| - \frac{1}{2}\,\mathrm{E}_F\big|X' - X\big|, \tag{10}$$

where X, X' are independent copies distributed according to F (Gneiting and Raftery, 2007). There is an obvious interest in evaluating weather forecasts with a focus on severe weather conditions like extreme winds or temperatures. For example, the weighted version of the CRPS as introduced in Gneiting and Ranjan (2011) is mentioned as a possible means to do so by Haiden et al. (2014) in the newsletter of the European Centre for Medium - Range Weather Forecasts.
Thus, let us discuss weighted versions of the CRPS. Given $r \in \mathbb{R}$, for the weight function $w(x) = 1\{x > r\}$ the weighted CRPS from Theorem 3, (i), is

$$\begin{aligned} \mathrm{wCRPS}(F, x; r) &= 1\{x > r\} \int_r^{\infty} \Big(\frac{F(z) - F(r)}{1 - F(r)} - 1\{x \le z\}\Big)^2 \,\mathrm{d}z \\ &= \frac{1\{x > r\}}{1 - F(r)} \Big(\mathrm{E}_F\big(\big|x - X\big|\, 1\{X > r\}\big) - \frac{1}{2(1 - F(r))}\,\mathrm{E}_F\big(\big|X' - X\big|\, 1\{\min(X', X) > r\}\big)\Big). \end{aligned} \tag{11}$$

If we complement it as described in Theorem 3, (ii), with the Brier Score, then we obtain the strictly locally proper version

$$\begin{aligned} &\mathrm{wsCRPS}(F, x; r) \\ &= 1\{x > r\} \Big[F(r)^2 + \int_r^{\infty} \Big(\frac{F(z) - F(r)}{1 - F(r)} - 1\{x \le z\}\Big)^2 \,\mathrm{d}z\Big] + 1\{x \le r\}\,\big(1 - F(r)\big)^2. \end{aligned} \tag{12}$$

Explicit forms in terms of the original distribution function are also possible when taking other indicators of intervals as weight functions.
Theorem 3 also allows to obtain weighted versions of general, possibly multivariate energy scores, which are defined in analogy to (10), but for which a representation in terms of Brier scores (9) does not exist. Details can be found in the supplementary material Holzmann and Klar (2017a).
For the CRPS, Gneiting and Ranjan (2011) introduced different weighted versions than those proposed in (12). Motivated by its representation in (9) as an integral over Brier scores, they proposed

$$\mathrm{twCRPS}(F, x; w) = \int_{-\infty}^{\infty} \big(F(z) - 1\{x \le z\}\big)^2 \, w(z) \,\mathrm{d}z \tag{13}$$

for a measurable weight function $0 \le w(z) \le 1$. This scoring rule remains proper for every w. Pelenis (2014) shows that it is not a localizing weighted scoring rule if the class of weight functions contains indicators of compact intervals $w(x) = 1\{a \le x \le b\}$, $a < b$.
However, we have the following result, which is hinted at in Pelenis (2014, p.16).

Theorem 4. *For the class of one-sided weight functions*

$$\mathcal{W}_{os} = \big\{w(x) = 1\{x > r\},\ r \in \mathbb{R}\big\} \cup \big\{w(x) = 1\{x < r\},\ r \in \mathbb{R}\big\},$$

the weighted CRPS in (13) is a localizing and strictly locally proper scoring rule.

Note that the weight functions $w(x) = 1\{x \geq r\}$ and $w(x) = 1\{x > r\}$ yield the same weighted scoring rule in (13), but not necessarily in (12) (e.g. Theorem 3). Let us also point out that Pelenis (2014) proposes a variant of the weighted CRPS in (13) called incremental CRPS. When well-defined, it is localizing and actually strictly locally proper, but the defining integral is infinite for one-sided weight functions $w(x) = 1\{x \geq r\}$.

Finally, let us mention that for continuous distribution functions F, the CRPS can also be written in terms of quantile forecasts as

$$CRPS(F, x) = \int_0^1 QS_\alpha\big(F^{-1}(\alpha), x\big)\, \mathrm{d}\alpha, \quad QS_\alpha(q, x) = 2\,\big(1_{x<q} - \alpha\big)(q - x),$$

and where F^{-1} is the quantile function of F. For a weight function $v : (0,1) \to [0,1]$, Gneiting and Ranjan (2011) define the quantile-weighted version of the CRPS as

$$QCRPS(F, x; v) = \int_0^1 QS_\alpha\big(F^{-1}(\alpha), x\big)\, v(\alpha)\, \mathrm{d}\alpha.$$

This is not a weighted scoring rule and hence cannot be localizing in the sense of this paper, since the weight function is not defined on the sample space $\mathbb{R}$ but rather on $(0,1)$. However, it satisfies another interesting property. Assume that the distribution functions are strictly increasing on their support, so that quantiles are unique and the quantile curve is continuous. If we choose $v(\alpha) = 1_{[r,1)}(\alpha)$, $r \in (0,1)$, then $QCRPS(G, F; v) = QCRPS(F, F; v)$ if and only if $F^{-1}(\alpha) = G^{-1}(\alpha)$ for all $\alpha \in [r, 1)$. Equivalently, $F^{-1}(r) = G^{-1}(r)$, and the probability distributions associated with F and G coincide on $\big[F^{-1}(r), \infty)$. Thus, the quantile-weighted CRPS evaluates the forecast F on a forecast-dependent region of interest. $\diamond$

2.4 Weighted scoring rules and tests of equal forecast performance

Formal comparisons of forecast equality are based on the so-called Diebold-Mariano test (Diebold and Mariano, 1995; Giacomini and White, 2006; Diebold, 2015), which uses normalized score differences as test statistic. Typically, in a time series setting, one has probabilistic forecasts F_t and G_t for an observation x_{t+k} that lies k time steps ahead. Given a scoring rule S, the test statistic is given by

$$T = \frac{\sqrt{n}\,\big(\bar{S}_F - \bar{S}_G\big)}{\hat{\sigma}}, \tag{14}$$

where $\bar{S}_F = 1/n \sum_{t=1}^n S(F_t, x_{t+k}), \bar{S}_G = 1/n \sum_{t=1}^n S(G_t, x_{t+k})$ and $\hat{\sigma}^2$ is a suitable estimator of the asymptotic variance of the score difference. One possible choice for $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = \begin{cases} \hat{\gamma}_0 & \text{if } k = 1, \\ \hat{\gamma}_0 + 2\sum_{j=1}^{k-1} \hat{\gamma}_j & \text{if } k \geq 2, \end{cases} \tag{15}$$

where $\hat{\gamma}_j$ denotes the lag j sample autocovariance of the sequence of score differences (Gneiting and Ranjan, 2011; Lerch et al., 2016). Under the null hypothesis of a vanishing expected score difference and standard regularity conditions, the test statistic T in (14) is asymptotically standard normal. When the null hypothesis is rejected in a two-sided test, F is preferred if the test statistic T is negative, and G is preferred if T is positive.

Following Lerch et al. (2016) we interpret the Diebold-Mariano test when using weighted scoring rules, and cast it into a framework in which two distributions are compared for the special case of independent observations. Lerch et al. (2016) argue that if one density is the true data-generating distribution, the optimal test is given by the Neyman-Pearson test. In terms of score differences, this corresponds to the ordinary logarithmic score, which therefore is optimal in this sense. Improvement by using weighted scoring rules can hence only be expected when comparing two misspecified densities. However, in their simulations Lerch et al. (2016) find no such systematic improvement.

Here we argue that for weight functions $w(x) = 1\{x \in A\}$, the aim is to ignore possible problems or advantages of the forecast outside the region of interest A. Thus, even if a forecast distribution P performs poorly outside of A but well on A, it is useful to us, indeed as useful as another forecast which performs well overall. Further, if the focus is on the region A, such a forecast P is to be preferred to a forecast Q which performs well ouside of A but poorly on A. This use of weighted scoring rules is not brought to light in the simulations of Lerch et al. (2016): In their setting, interest focuses on the right tail but all density forecasts compared are correctly specified in the left tail, and ignoring that region does not result in an increased power.

Let P_0, P_1 be two competing forecast distributions with densities p_0, p_1 w.r.t. μ, and assume that $0 < P_0(A), P_1(A) < 1$. The property (8) of localizing weighted scoring rules implies that the forecasts are only relevant through their values on A. Thus testing using score differences with weight function $w(x) = 1\{x \in A\}$ amounts to testing

$$H_0 : p = p_0 \quad \mu - \text{a.e. on } A \quad \text{vs.} \quad H_1 : p = p_1 \quad \mu - \text{a.e. on } A \tag{16}$$

for the unknown true density p. Hence, we have composite null and alternative hypotheses arising from a censoring of the forecasting distributions. The density forecast p is only relevant for the hypotheses through observations $x \in A$, for $x \notin A$ only the total probability $1 - P(A)$ matters.

Such hypotheses can be tested by score differences based on a localizing weighted scoring rule with weight function $w(x) = 1\{x \in A\}$, and it can be shown that the weighted scoring rule leading to optimal power properties in this framework is the censored likelihood rule of Diks et al. (2011), see Holzmann and Klar (2016) for formal statements and proofs. The testing performance will be further investigated in the subsequent simulation section.

3 Simulations

In this section, we consider simulation settings similar to those in Diks et al. (2011) and Lerch et al. (2016). Suppose that at time $t = 1, \ldots, n$, the observations x_t are independent standard normally distributed. We apply the two-sided Diebold-Mariano test of equal predictive performance, nominal

	scoring rule	proper	strictly proper	localizing	strictly locally proper	proportionally locally proper
unweighted	CRPS	yes	yes	no	-	-
	LogS	yes	yes	no	-	-
weighted	CSL	yes	-	yes	yes	no
	PWL	yes	-	yes	yes	no
	CL	yes	-	yes	no	yes
	twCRPS	yes	-	no (yes)	no (yes)	no
	wCRPS	yes	-	yes	no	yes
	wsCRPS	yes	-	yes	yes	no

Table 1: Summary of properties of unweighted and weighted scoring rules. The entry *no (yes)* for twCRPS indicates that it is localizing and strictly locally proper for the one-sided weight functions used in the simulations, but not in general.

level $\alpha = 0.05$, using the variance estimate in (15) with $k = 1$. As nonparametric alternative, we also apply the two-sided Wilcoxon signed-rank test, nominal level $\alpha = 0.05$, but defer those results to the supplementary material Holzmann and Klar (2017a), to save space and also since the Wilcoxon test is not easily transferred to dependent data. All results in this section are based on 10 000 replications.
We use the logarithmic score (LogS) and the continuous ranked probability score (CRPS) as typical examples of unweighted scoring rules. As weighted scoring rules, we apply three likelihood based scoring rules, namely, the censored likelihood rule (CSL), the penalized weighted likelihood rule (PWL), and the conditional likelihood rule (CL). Further, we use the following CRPS based weighted scoring rules: the threshold weighted continuous ranked probability score (twCRPS) defined in (13), wCRPS defined in (11) and wsCRPS defined in (12). Table 1 gives a summary of the properties of these scoring rules.
Suppose that we are only interested in the forecast quality on a subset of the support of the underlying distribution. For example, interest may center on the positive real numbers or on the right tail of the distribution. Hence, the tests under the weighted scoring rules are based on the indicator weight function $w(x) = \mathbf{1}\{x \geq r\}$ in all simulations. Furthermore, we use sample size $n = 100$ throughout all simulations.

Scenario A: As first example, we reconsider the scenario introduced in Section 2.1. In this scenario, Forecast 1 is a piecewise defined distribution F_{hlt} with heavy left tail, whereas Forecast 2 is a piecewise defined distribution F_{hrt} with heavy right tail.
Fig. 3 shows the proportion of rejections of the null hypothesis of equal predictive performance in two-sided Diebold-Mariano tests as a function of the threshold value r in the weight function. The upper (lower) panels show rejections in favor of F_{hlt} (in favor of F_{hrt}).
In these and the following plots, the left panels show rejections for likelihood based scoring rules, whereas the right panels show rejections for CRPS based rules.

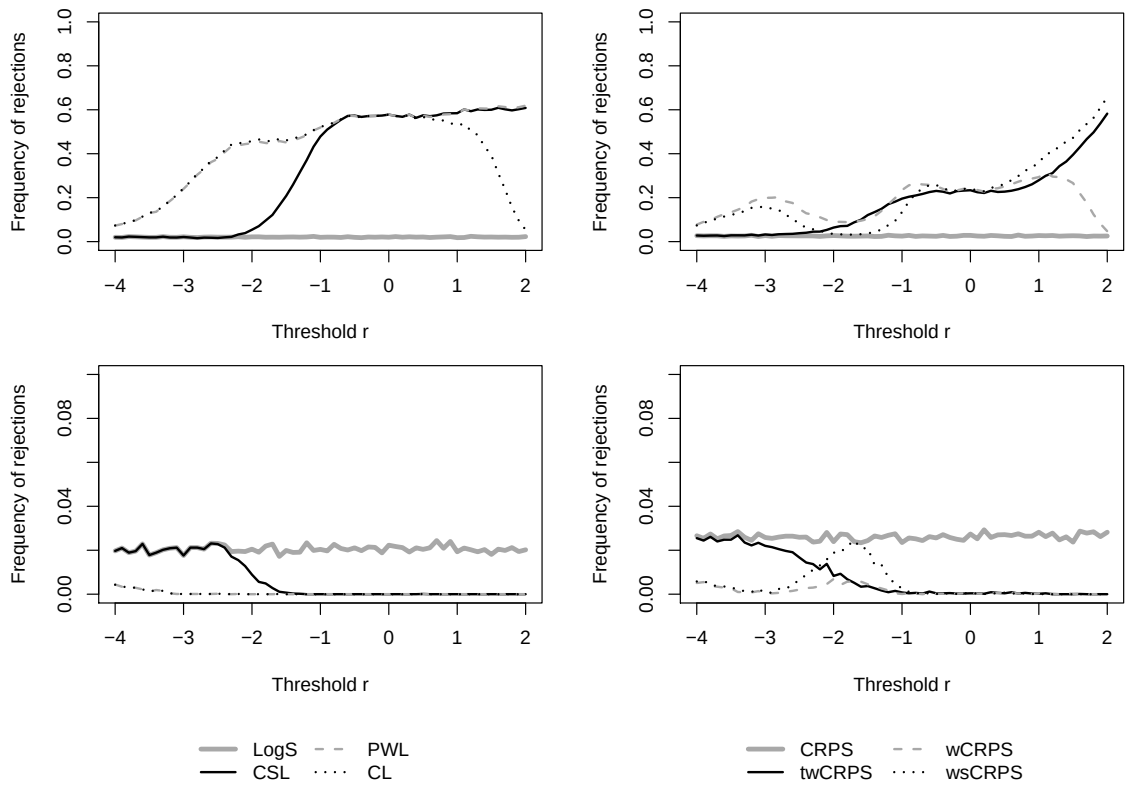

Figure 3: Scenario A. The null hypothesis of equal predictive performance of F_{hlt} and F_{hrt} is tested under a standard normal population. The panels show the frequency of rejections in two-sided Diebold-Mariano tests for likelihood based (left) and CRPS based (right) scoring rules. The upper (lower) panels show rejections in favor of F_{hlt} (in favor of F_{hrt}).

For $r = -\infty$, both forecasts have the same distance from the (true) standard normal distribution, and neither of them should be rejected in favor of the other. However, for $r > 0$, Forecast 1 coincides with Φ, and Forecast 2 should be rejected.

As one would expect, the rejection frequencies in favor of F_{hlt} and in favor of F_{hrt} for the two non-weighted scoring rules are around 0.025 (be aware of the different scaling of the lower panels!). Under the likelihood based weighted scoring rules, CL and PWL have a very similar behavior for negative values of r. They show a faster increase of the rejection frequencies in favor of F_{hlt} compared to CLS. However, CL decreases to zero for large positive values of r. This is due to the fact that the effective sample size, i.e. the number of observations exceeding r becomes very small with increasing threshold. Concerning the CRPS based weighted rules, wCRPS and wsCRPS behave quite similarly for negative and moderately positive values of r. Their rejection frequencies in favor of F_{hlt} have a first modal value around $r = -3$, decrease until -2, and increase again. However, like CL, wCRPS decreases to zero for large positive values of r.
In this scenario, generally speaking the likelihood-based rules have higher power than the CRPS-based

rules. At least for $r \geq 0$ this is certainly a desirable property.

Scenario B: A potential objection against Scenario A may be that Forecast 1 coincides exactly with the data generating distribution for positive values of r which is rather unrealistic in applications. Hence, we also consider the following modification, a smoothed version of Scenario A: denote the cdf of a normal distribution with mean μ and standard deviation σ by $\Phi_{\mu,\sigma}$. Let

$$G(x) = \Phi_{0,1/2}(x)\,\Phi(x) + \big(1 - \Phi_{0,1/2}(x)\big)\,F_4(x),$$
$$H(x) = \big(1 - \Phi_{0,1/2}(x)\big)\,\Phi(x) + \Phi_{0,1/2}(x)\,F_4(x),$$

where F_4 denotes the distribution function of the t-distribution with 4 degrees of freedom.

In Scenario B, we consider Forecast 1: G vs. Forecast 2: H.
In this scenario, both forecasts are different from the (true) standard normal distribution on each observation window $[r, \infty)$. As in Scenario A, both forecasts have the same overall distance from the standard normal for $r = -\infty$, and neither of them should be rejected in favor of the other. However, if one is only interested in the region $[r, \infty)$ for larger positive values of r, forecast G is close to Φ; hence, H should be rejected.
Qualitatively, the results of all simulations for this scenario parallel the findings for Scenario A. Hence, details are deferred to the supplementary material Holzmann and Klar (2017a).

Scenario C: Forecast 1: Φ vs. Forecast 2: F_{hlt}.
Here, Φ denotes the cumulative distribution function (cdf) of the standard normal distribution, and F_{hlt} is defined as in Scenario A. Clearly, for positive values of r, Φ and F_{hlt} coincide.

Fig. 4 shows the proportion of rejections of the null hypothesis of equal predictive performance in two-sided Diebold-Mariano tests as a function of the threshold value r in the weight function. The upper (lower) panels show rejections in favor of Φ (in favor of F_{hlt}).
For $r < 0$, rejections in favor of the standard normal distribution represent true power, but if one is interested in the region $[r, \infty)$ for positive r, both forecasts are identical, and neither of them should be rejected.
Let us first look at the non-weighted scoring rules. They have rather different rejection frequencies in favor of Φ when using the Diebold-Mariano test, with LogS well above CRPS.
Clearly, for large negative values of r, the rejection frequencies in favor of Φ of CSL, PWL and CL coincide with those of LogS, but those of CL and PWL, which are nearly identical, decrease faster to zero than for CSL. Similarly, for large negative values of r, the rejection frequencies in favor of Φ of twCRPS, wCRPS and wsCRPS coincide with those of CRPS, but those of wCRPS and wsCRPS decrease faster to zero than for twCRPS.
The rejection frequencies in favor of F_{hlt} have a peculiar and undesirable peak to the left of zero for all likelihood based weighted scoring rules This is not the case for the CRPS based weighted rules.

Scenario D: Forecast 1: Φ vs. Forecast 2: G.
This scenario is a smoothed version of Scenario C. Here, G, defined in Scenario B, nowhere equals Φ exactly, but is more similar to Φ for positive values of the threshold r than for negative ones.

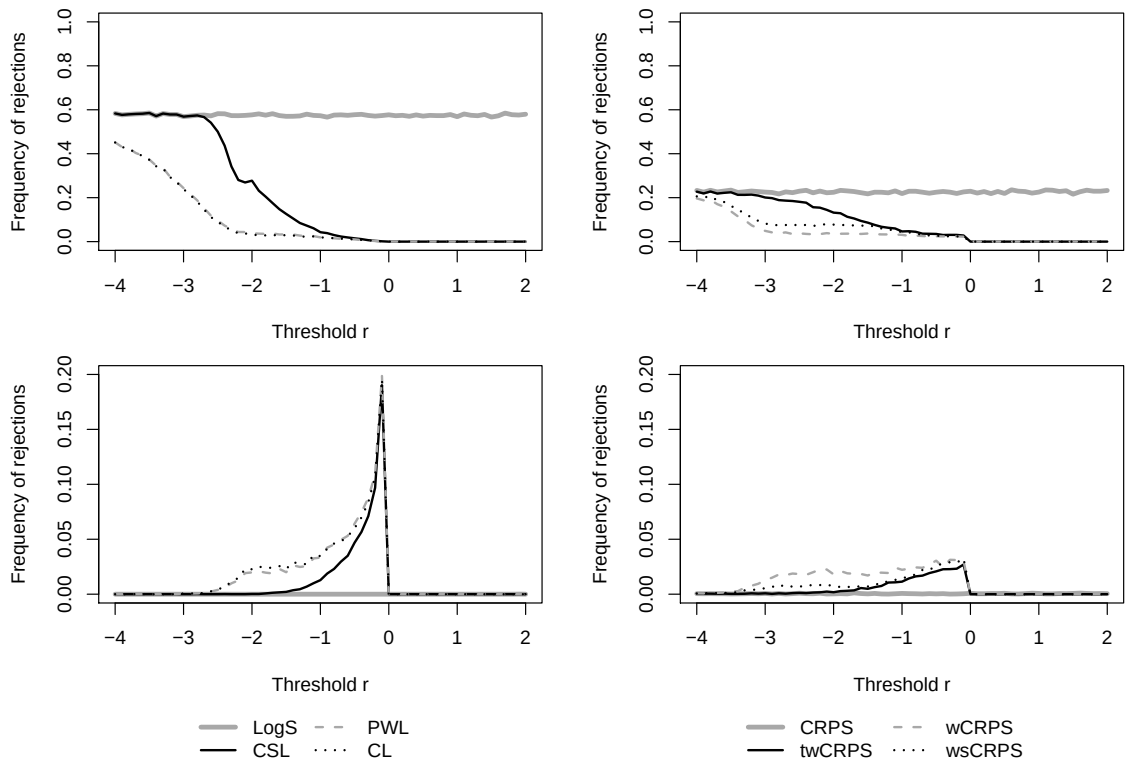

Figure 4: Scenario C. The null hypothesis of equal predictive performance of Φ and F_{hlt} is tested under a standard normal population. The panels show the frequency of rejections in two-sided Diebold-Mariano tests for likelihood based (left) and CRPS based (right) scoring rules. The upper (lower) panels show rejections in favor of Φ (in favor of F_{hlt}).

Fig. 5 shows the proportion of rejections of the null hypothesis of equal predictive performance in two-sided Diebold-Mariano tests as a function of the threshold value r in the weight function. The upper (lower) panels show rejections in favor of Φ (in favor of G).
Formally, rejections in favor of the standard normal distribution represent true power, but if one is interested in the region $[r, \infty)$ for positive r, both forecasts are quite similar. Qualitatively, most results for this scenario parallel the findings for Scenario C, but the likelihood-based rules do no longer have a much higher undesirable peak in the rejection frequency in favour of G for small, negative values of r than the CRPS-based rules.

As general conclusion, we can state that the overall power of the likelihood based rules is higher than that of the CRPS based rules in all scenarios. The faster increase in power of PWL and CL compared to CLS in Scenarios A and B occurs for values of r for which nearly no observation is below the threshold. Hence, this increase seems to be rather an artefact due to differences of the distribution functions at the threshold. Furthermore, the undesirable behaviour of the CLS in Scenario C concerning the rejections in favour of F_{hlt} vanishes under the more realistic Scenario D. Hence, the CLS is the overall preferable

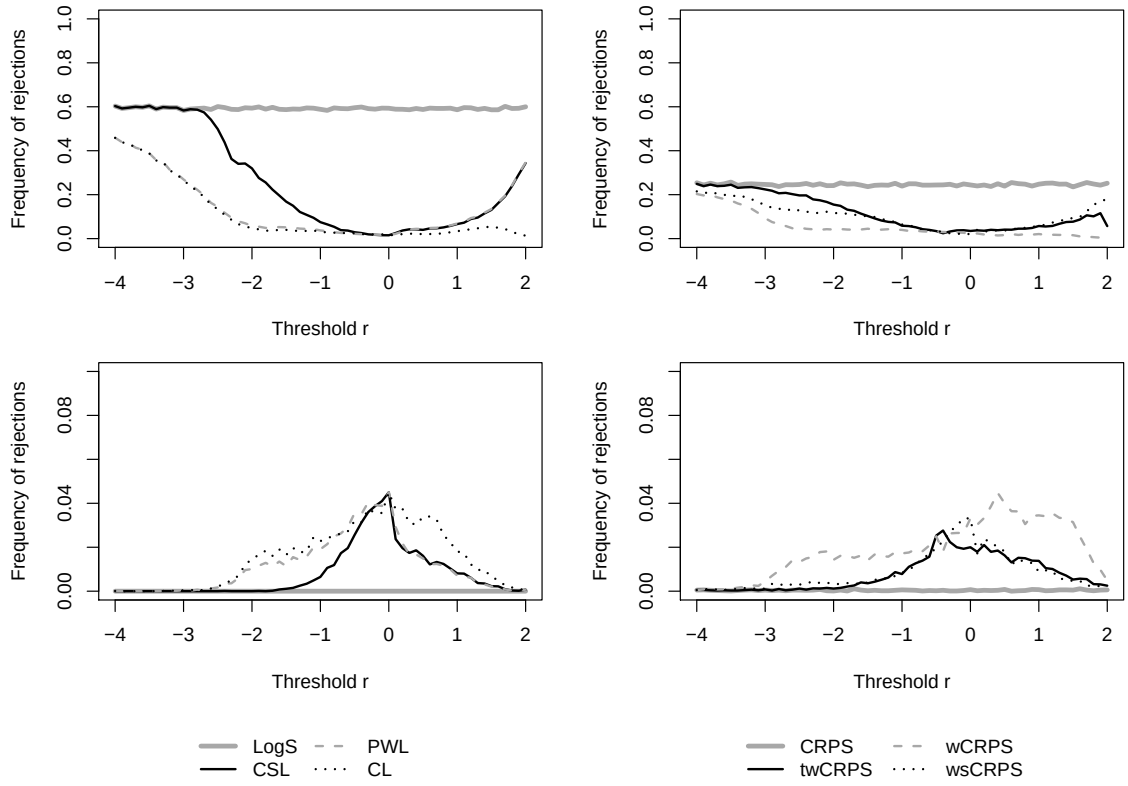

Figure 5: Scenario D. The null hypothesis of equal predictive performance of Φ and F_{hlt} is tested under a standard normal population. The panels show the frequency of rejections in two-sided Diebold-Mariano tests for likelihood based (left) and CRPS based (right) scoring rules. The upper (lower) panels show rejections in favor of Φ (in favor of G).

scoring rule under the Diebold-Mariano test. Using the Wilcoxon signed-rank test the results are overall comparable, but differ in some details. Except for the twCRPS, the CRPS-based rules show some erratic behaviour in Scenarios A and B. The twCRPS has the undesirable spike for small negative values of r in Scenario C, but not in D. In terms of power, the twCRPS is now competitive with the likelihood-based rules, and has best overall performance under the Wilcoxon signed-rank test. However, the Wilcoxon signed-rank test does not seem to be generally recommendable for testing for equal forecast quality based on score differences. First, it may severely fail under temporal dependence (Diebold and Mariano, 1995), second it sometimes reacts strongly to certain effects. For example, there are sometimes larges spikes around zero due to the fact that F and G coincide in zero.

4 Empirical illustration

We apply the proposed forecasting rules to two time series of daily log returns $x_t = \ln(P_t/P_{t-1})$, where P_t is the closing price on day t, adjusted for dividends and splits. We consider S&P 500 and Deutsche

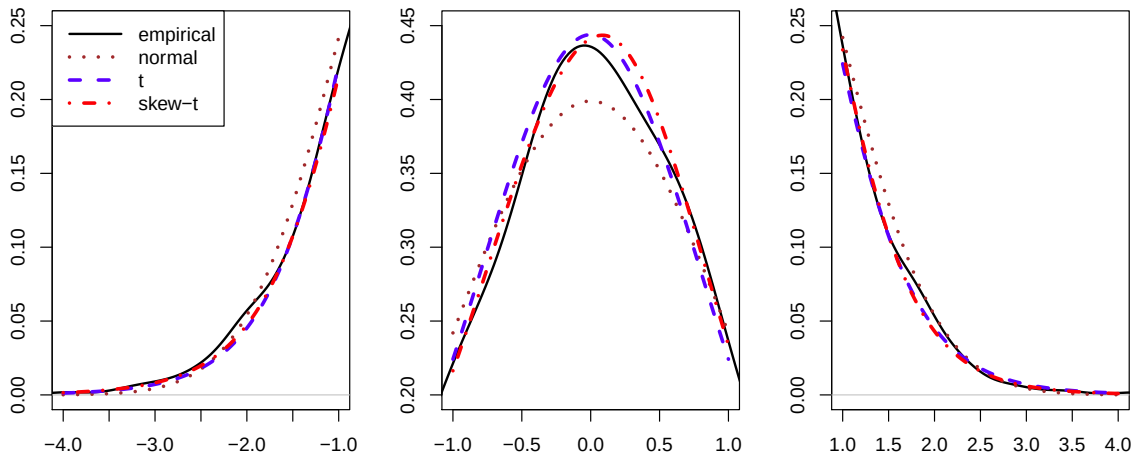

Figure 6: Empirical and theoretical density functions of the residuals of a GARCH(1,1)-model fitted to the Deutsche Bank return series. For better visibility, left tail, center and right tail of the distribution are displayed in separate panels.

Bank AG log-returns for a sample period running from January 1, 2009 until December 31, 2016, giving a total of 2013 and 2033 observations. The data is publicly available and has been downloaded from `http://finance.yahoo.com`. Since Yahoo finance data for Deutsche Bank partially includes holidays, we removed all days with zero trading volume.
We define three forecast methods based on the following GARCH(1,1) model,

$$x_t = \mu + \sigma_t z_t, \qquad \sigma_t^2 = \omega + \alpha_1 (x_{t-1} - \mu)^2 + \beta_1 \sigma_{t-1}^2, \tag{17}$$

using normal, t and skew-t distributions for the innovations to account for leptokurtosis and/or skewness. Since a typical finding in empirical applications of GARCH models is that a normal distribution for z_t does not fully account for the kurtosis observed in stock returns, we may expect that the forecast with t-distributed innovations gives better density forecasts.

To illustrate that all three methods are slightly misspecified, we start with a goodness-of-fit type residual analysis on the full time series of Deutsche Bank log-returns. The GARCH residuals are given by $e_t = (x_t - \hat{\mu})/\hat{\sigma}_t$, where $\hat{\mu}$ is the estimated mean, and $\hat{\sigma}_t$ denotes the fitted volatility process. Since the estimates for μ, ω, α_1 and β_1 are very similar for the three models, the resulting empirical distributions of the residuals are visually nearly indistinguishable.
Hence, Fig. 6 only shows the kernel density estimate (created by the R function `density`) of the residuals, when the GARCH parameters and hence the conditional standard deviations $\hat{\sigma}_t$ are estimated under normality assumption. Additionally, Fig. 6 shows the densities of a standard normal distribution, the t-distribution with shape parameter 8.4 as obtained in the estimation process, and the fitted skew-t-distribution with shape and skewness parameter 8.5 and 0.94, respectively.
At first sight, the empirical density looks fairly symmetric, and all three distributions seem to fit the tails quite well, whereas the normal density is not sufficiently peaked in the center. Looking more

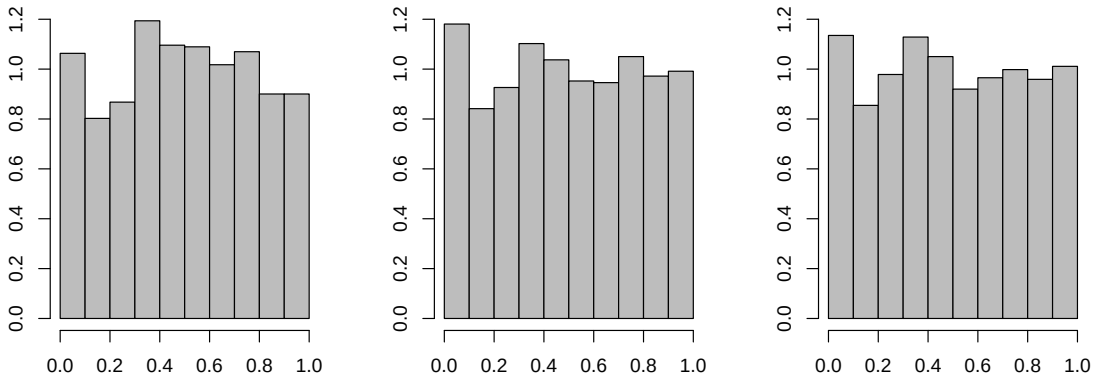

Figure 7: Histograms of the probability integral transforms for the three models applied to the Deutsche Bank returns.

closely, one actually finds regions in the right tail where the normal distribution fits better than t and skew-t; thus, the advantage of the latter diminishes. In the center, the skew-t seems to yield a better fit than the t distribution for values smaller than zero, and vice versa for positive values.

To evaluate the forecasting performance of the three methods, we use one-step-ahead density forecasts with a rolling window scheme for parameter estimation done by maximum likelihood method using R and the R package rugarch (Ghalanos (2014), R Core Team (2016)). The length of the estimation window is set to be 500 observations, so that the number of out-of-sample observations equals 1513 and 1533. The histograms of the probability integral transforms for the Deutsche Bank returns as shown Fig. 7 also indicate that all three forecasting mechanisms are somewhat misspecified.
For comparing the density forecasts' accuracy we apply the Diebold-Mariano test based on several weighted and unweighted scoring rules. Localizing weighted scoring rules are particularly suitable for comparing forecasts which are misspecified to a varying degree in distinct regions of interest. We use the threshold weight function $w(x) = 1\{x \leq r\}, r = -1, 0$, and $w(x) = 1\{x \geq r\}, r = 0, 1$, and additionally $r = 3$ and $r = -3$ for the Deutsche Bank returns. Hence we concentrate either on losses or on gains when using the weighted scoring rules. The score difference is computed by subtracting the score of the normal GARCH density forecast from the score of the t-GARCH density forecast, so that positive values indicate better predictive ability of the forecast method based on Student-t innovations, and similarly for normal vs. skew-t and t vs. skew-t innovations. The results for the S&P 500 and Deutsche Bank AG can be found in Tables 2 and 3, respectively.
On the whole, forecasts for the S&P 500 returns using a t or skew-t GARCH model are superior to a normal GARCH model; using weighted scoring rules, we see that this holds especially for losses, but only to a lesser extent for gains. In particular, the t GARCH model seems to be inferior to the normal GARCH for the threshold weight function $1\{x \geq 1\}$. As can be seen in the lower panel of Table 2, results are less clear cut between t and skew-t GARCH density forecasts depending on the weight function, with an overall advantage for the skew-t GARCH model.
For the Deutsche Bank returns, t and skew-t GARCH density forecasts are generally superior to a

	$w(x)$	$1\{x \leq -1\}$	$1\{x \leq 0\}$	$1\{x \geq 0\}$	$1\{x \geq 1\}$
	proportion	0.12	0.46	0.54	0.14
normal garch vs. t-garch	LogS	3.06	3.06	3.06	3.06
	CRPS	1.07	1.07	1.07	1.07
	CSL	2.13	2.72	0.51	-1.55
	PWL	2.15	2.85	1.00	-1.82
	twCRPS	-0.08	0.02	2.52	0.06
	wsCRPS	2.12	-0.62	1.58	-1.26
normal garch vs. skew-t-garch	LogS	3.25	3.25	3.25	3.25
	CRPS	1.60	1.60	1.60	1.60
	CSL	2.58	2.99	0.30	0.52
	PWL	2.57	3.15	0.85	0.42
	twCRPS	1.26	0.88	1.67	0.42
	wsCRPS	2.19	-0.05	0.84	0.31
t-garch vs. skew-t-garch	LogS	1.06	1.06	1.06	1.06
	CRPS	0.87	0.87	0.87	0.87
	CSL	1.82	1.27	-0.22	2.65
	PWL	1.84	1.40	0.04	2.79
	twCRPS	1.59	1.39	-0.62	0.78
	wsCRPS	1.32	0.69	-0.99	1.53

Table 2: t-statistics for Diebold-Mariano test for equal predictive accuracy for S&P 500. Positive values indicate superiority of forecasts from the second method, while negative values indicate superiority of forecasts from the first method.

normal GARCH model for all (weighted and unweighted) scoring functions, but again this holds to a lesser extent for gains, as can be seen in Table 3. The lower panel shows that there is no significant overall difference between t and skew-t GARCH density forecasts; however, the skew-t GARCH model is significantly better for predicting losses whereas the t GARCH model is clearly superior for predicting gains. As discussed in the introduction, the skew-t model is the model of choice for risk management applications, independent of the specific risk measure.

In Holzmann and Klar (2017b) we further illustrate benefits of a semiparametric, extreme-value based modeling of the distribution of the GARCH innovations, and also include rankings based on the QCRPS from Section 2.3 as well as on quantile scores for various levels.

5 Discussion and conclusions

Lerch et al. (2016) discuss the so-called forecasters dilemma, in that forecasts are often only evaluated in case that extreme events actually occur. They point out that such a restriction of forecast evaluation to subsets of the available observations has highly unwanted effects, and it discredits even the best possible forecast, that is the true conditional distribution.

Weighted scoring rules which remain proper are a valid decision-theoretic tool for emphasizing regions

		w(x) $= 1\{x \leq r\}$			w(x) $= 1\{x \geq r\}$		
		$r = -3$	$r = -1$	$r = 0$	$r = 0$	$r = 1$	$r = 3$
	proportion	0.096	0.30	0.50	0.50	0.32	0.092
normal garch vs. t-garch	LogS	2.43	2.43	2.43	2.43	2.43	2.43
	CRPS	1.51	1.51	1.51	1.51	1.51	1.51
	CSL	1.89	1.71	1.96	1.63	1.73	0.95
	PWL	1.85	1.69	1.99	1.66	1.78	0.94
	twCRPS	1.34	0.89	0.65	1.08	1.21	0.83
	wsCRPS	1.91	0.38	0.51	1.32	1.89	0.70
normal garch vs. skew-t-garch	LogS	2.18	2.18	2.18	2.18	2.18	2.18
	CRPS	1.22	1.22	1.22	1.22	1.22	1.22
	CSL	2.01	1.97	2.06	0.74	1.12	0.23
	PWL	1.96	1.94	2.13	0.83	1.18	0.24
	twCRPS	1.55	1.25	0.84	0.48	0.66	0.25
	wsCRPS	1.67	1.26	0.63	0.44	0.80	-0.25
t-garch vs. skew-t-garch	LogS	-0.61	-0.61	-0.61	-0.61	-0.61	-0.61
	CRPS	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70
	CSL	1.65	2.30	1.31	-2.10	-1.49	-1.76
	PWL	1.66	2.20	1.60	-2.03	-1.46	-1.72
	twCRPS	0.25	1.22	1.11	-1.88	-1.49	-1.11
	wsCRPS	0.53	1.45	0.07	-1.79	-0.96	-0.91

Table 3: t-statistics for Diebold-Mariano test for equal predictive accuracy for Deutsche Bank AG. Positive values indicate superiority of forecasts from the second method, while negative values indicate superiority of forecasts from the first method.

of interest. We give a general construction method for such rules, and apply it in particular to the continuous-ranked probability score, thus obtaining a novel weighted version of this popular scoring rule. Further, we show how regions of interest can also be adopted for functionals of the forecast distribution, thus clarifying an issue raised in Lerch et al. (2016).

Weighted scoring rules are particularly useful for ranking misspecified forecasts. Indeed, if a forecast, although misspecified, works well on the region of interest A (but potentially very poorly outside of A), it will be found superior to another forecast with poor performance on A (but potentially very good performance outside of A). These considerations are confirmed for basically all the proper weighted scoring rules that we use in our simulations.

Concerning the specific choice of the weighted scoring rule, the censored likelihood rule from Diks et al. (2011) is preferable in terms of power properties. If stability is also an issue, or if forecast distributions are given in terms of Monte Carlo output (Krüger et al., 2017), the twCRPS from Gneiting and Ranjan (2011) as well as the wsCRPS proposed in this paper could also be recommended.

In our empirical illustration all forecasts are slightly misspecified, as is often unavoidable in practice. While it is inferior to normal and t distributions for gains, the skew-t distribution works best for predicting losses, and hence is the method of choice for the purpose of risk management.

6 Acknowledgements

The authors are grateful to the Editor Tilmann Gneiting for his constructive guidance and important suggestions as well as to two referees for helpful comments that improved the presentation of the material in the paper.

References

Amisano, G. and Giacomini, R. (2007). *Comparing density forecasts via weighted likelihood ratio tests.* Journal of Business and Economic Statistics 25, 177-190.

Bank of England (2017). *Monetary Policy Framework.* Retrieved from `http://www.bankofengland.co.uk/monetarypolicy/Pages/framework/framework.aspx`

Billi, R.M. (2016). *A note on nominal GDP targeting and the zero lower bound.* Macroeconomic Dynamics, doi:10.1017/S136510051500111X

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown, B. G. and Mason, S. (2008). *Forecast verification: current status and future directions.* Meteorol. Appl. 15, 3 - 18.

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

De Nicolò, G. and Lucchetta, M. (2017). *Forecasting tail risk.* Journal of Applied Econometrics 32, 159-170.

Diebold, F. X. (2015). *Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests.* Journal of Business and Economic Statistics 33, 1-24.

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.

Ehm, W. and Gneiting, T. (2012). *Local proper scoring rules of order two.* Annals of Statistics 40, 609-637.

Elliott, G. and Timmermann, A. (2016). *Forecasting in Economics and Finance.* Annual Review of Economics 8, 81-110.

Garín, J., Lester, R., Sims, E. (2016). *On the desirability of nominal GDP targeting.* Journal of Economic Dynamics and Control 69, 21-44.

Ghalanos, A. (2014). *rugarch: Univariate GARCH models.* R package version 1.3-5.

Giacomini, R. and White, H. (2006). *Tests of conditional predictive ability.* Econometrica 74, 1545-1578.

Giesbergen, B. (2017). *China: how realistic is the government's growth target?* Economic Report, Rabobank. Retrieved at `https://economics.rabobank.com/publications/2017/march/china-how-realistic-is-the-governments-growth-target`

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

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

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

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

Haiden, T., Magnussen, L. and Richardson, D. (2014). *Statistical evaluation of ECMWF extreme wind forecasts.* In: European Centre for Medium - Range Weather Forecasts Newsletter, Spring 2014.

Holzmann, H. and Klar, B. (2017). *Supplementary material to: Focusing on regions of interest in forecast evaluation.*

Holzmann, H. and Klar, B. (2017). *Discussion of "Elicitability and backtesting: Perspectives for banking regulation".* to appear: *Annals of Applied Statistics.*

Holzmann, H. and Klar, B. (2016). *Weighted scoring rules and hypothesis testing.* Arxiv preprint arXiv:1611.07345v2

Hyvärinen, A. (2005). *Estimation of non-normalized statistical models by score matching.* Journal of Machine Learning Research 6, 695-709.

Krüger, F., Lerch, S., Thorarinsdottir, T. L., Gneiting, T. (2017). *Probabilistic Forecasting and Comparative Model Assessment Based on Markov Chain Monte Carlo Output.* Preprint, available in arXiv:1608.06802.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., Gneiting, T. (2016). *Forecaster's Dilemma: Extreme Events and Forecast Evaluation.* Statistical Science, to appear.

McNeil, A. J., Frey, R. and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools.* Princeton Series in Finance, Princeton University Press

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

Nolde, N. and Ziegel, J.F. (2017). *Elicitability and backtesting: Perspectives for banking regulation.* Annals of Applied Statistics, to appear.

Opschoor, A., van Dijk D. and van der Wel M. (2017). *Combining density forecasts using focused scoring Rules.* J. Appl. Econ. 2017;0:1-16. https://doi.org/10.1002/jae.2575

Pelenis, J. (2014). *Weighted scoring rules for comparison of density forecasts on subsets of interest.* Preprint. Retrieved from `https://sites.google.com/site/jpelenis/` in July, 2017.

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

Patton, A. (2017). *Evaluating and comparing possibly misspecified forecasts.* Working paper, Duke University.

Pisoni, E., Farina, M., Pagani, G. and Piroddi, L. (2011). *Environmental Over-Threshold Event Forecasting using NARX Models.* Preprints of the 18th IFAC World Congress Milano (Italy).

R Core Team (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

7 Proofs

Proof of Theorem 1. By Definition, $S(p, x; w)$ depends on p only through p_w, and hence on p only on $\{w > 0\}$, thus, $S(p, x; w)$ is localizing. Further, we have

$$\begin{aligned} S(p, q; w) &= \int q_w(x)\, \widetilde{S}(p_w, x)\, \mathrm{d}\mu(x) \int qw \\ &= \widetilde{S}(p_w, q_w) \int qw. \end{aligned}$$

Since $\widetilde{S}$ is proper, $S(p, q; w)$ is minimal in p for given q if $p_w = q_w$, which is implied by $p = q$. Hence $S(p, x; w)$ is proper. Further, $S(p, q; w) = S(q, q; w)$ implies that $\widetilde{S}(p_w, q_w) = \widetilde{S}(q_w, q_w)$. Hence, if $\widetilde{S}$ is strictly proper, this implies that $p_w = q_w$ μ-a.e. But this holds if and only if the densities p and q are proportional on $\{w > 0\}$ μ-a.e. This concludes the proof. □

Proof of Theorem 2. The rule $S_{\mathbf{s}}$ is localizing w.r.t. p since it depends only on $\int pw$. Further, it is proper since

$$S_{\mathbf{s}}(p, q; w) - S_{\mathbf{s}}(q, q; w) = \mathbf{s}\Big(\int pw, \int qw\Big) - \mathbf{s}\Big(\int qw, \int qw\Big) \geq 0, \tag{18}$$

where we used the notation

$$\mathbf{s}(\alpha, \beta) = \beta \mathbf{s}(\alpha, 1) + (1 - \beta)\mathbf{s}\big(\alpha, 0\big).$$

Now, as a sum of two locally proper scoring rules the rule $\widehat{S}$ is also a locally proper scoring rule. Further, if

$$\widehat{S}(q, q; w) = \widehat{S}(p, q; w),$$

then necessarily $S(q,q;w) = S(p,q;w)$ and $S_{\mathbf{s}}(q,q;w) = S_{\mathbf{s}}(p,q;w)$ since both rules $S(\cdot,\cdot;w)$ and $S_{\mathbf{s}}(\cdot,\cdot;w)$ are proper. By assumption on $S(\cdot,\cdot;w)$, $S(q,q;w) = S(p,q;w)$ implies that $p = c\,q$ on $w > 0$. From $S_{\mathbf{s}}(q,q;w) = S_{\mathbf{s}}(p,q;w)$, (18) and the fact that $\mathbf{s}$ is strictly proper we get that $\int pw = \int qw$. Since we assume $\int qw \neq 0$ and $\int pw \neq 0$, we get for the proportionality constant that $c = 1$ and hence $p = q$ μ-a.e. on $w > 0$, so that $\widehat{S}$ is strictly locally proper. □

Proof of Theorem 4. Consider a weight function $w(x) = 1\{x > r\}$, the other case is similar. Given two distribution functions $F, G \in \mathcal{M}$ we denote by μ_F and μ_G the corresponding probability measures. The restriction $\tilde{\mu}_F$ of μ_F to (r,∞), a sub-probability measure, has the sub-distribution function $\tilde{F}(x) = F(x) - F(r)$, $x \geq r$, and $\tilde{F}(x) = 0$ otherwise, which uniquely determines this restriction. As $x \to \infty$, we recover $F(r)$ and hence $F(x)$, $x \geq r$ from $\tilde{\mu}_F$. On the other hand, $\tilde{F}$ and hence $F(x)$ for $x \geq r$ uniquely determine $\tilde{\mu}_F$.
Thus if the restrictions of μ_F and μ_G to (r,∞) are equal, $F(x) = G(x)$ for all $x \geq r$, so that for all x,

$$\mathrm{twCRPS}(F,x;w) = \int_r^\infty \big(F(z) - 1\{x \leq z\}\big)^2 \,\mathrm{d}z = \int_r^\infty \big(G(z) - 1\{x \leq z\}\big)^2 \,\mathrm{d}z = \mathrm{twCRPS}(G,x;w),$$

and the weighted CRPS is localizing.
A computation shows that

$$\mathrm{twCRPS}(F,G;w) - \mathrm{twCRPS}(F,F;w) = \int_r^\infty \big(F(z) - G(z)\big)^2 \,dz.$$

Thus if $\mathrm{twCRPS}(F,G;w) = \mathrm{twCRPS}(F,F;w)$, $F(x) = G(x)$ for Lebesgue-almost all $x \geq r$, and by right continuity of F and G, the equality holds for all $x \geq r$. From the discussion above, this implies that the restrictions of μ_F and μ_G to (r,∞) are equal. □