Comparing Density Forecasts Using Threshold and Quantile Weighted Scoring Rules

Tilmann Gneiting and Roopesh Ranjan

Technical Report no. 533

Department of Statistics, University of Washington

April 2008

Abstract

We propose a method for comparing density forecasts that is based on weighted versions of the continuous ranked probability score. The weighting emphasizes regions of interest, such as the tails or the center of a variable's range, while retaining propriety, as opposed to a recently developed weighted likelihood ratio test, which can be hedged. Threshold and quantile based decompositions of the continuous ranked probability score can be illustrated graphically and prompt insights into the strengths and deficiencies of a forecasting method. We illustrate the use of the test and graphical tools in case studies on the Bank of England's density forecasts of quarterly inflation rates in the United Kingdom, and probabilistic predictions of wind resources in the Pacific Northwest.

KEY WORDS: Continuous ranked probability score; Predictive ability testing; Probabilistic forecast; Proper scoring rule; Quantile; Weighted likelihood ratio test

1 Introduction

One of the major tasks of statistical analysis is to make forecasts for the future. To realize their full potential, forecasts ought to be probabilistic in nature, taking the form of probability distributions over future quantities or events (Dawid 1984). Here we are concerned with density forecasts of a continuous variable, such as inflation rate, gross domestic product, temperature or wind speed, to name but a few examples. With the continued proliferation of probabilistic forecasts in economic, environmental and meteorological applications, among others, there is a critical need for principled techniques for the comparison and ranking of density forecasts (Timmermann 2000; Elliott and Timmermann 2008; Gneiting 2008).

Following Amisano and Giacomini (2007), we consider density forecasts in a time series context, in which a rolling window consisting of the past m observations is used to fit a density forecast for the observation that is k time steps ahead. Specifically, suppose that $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T$ is a stochastic process which can be partitioned as $\boldsymbol{Z}_t = (Y_t, \boldsymbol{X}_t)$ where Y_t is the variable of interest and $\boldsymbol{X}_t$ is a vector of predictors. Suppose that $T = m + n + k$. At times $t = m, \ldots, m + n$, density forecasts $\hat{f}_{t+k}$ and $\hat{g}_{t+k}$ for Y_{t+k} are generated, each of which depends only on $\boldsymbol{Z}_{t-m+1}, \ldots, \boldsymbol{Z}_t$. In this framework, the only requirement imposed on how the forecasts are produced is that they are measurable functions of the data in the rolling estimation window. We are interested in comparing and ranking the competing density forecasting methods.

The comparison typically uses a proper scoring rule. A scoring rule is a loss function $\mathrm{S}(f, y)$ whose arguments are the density forecast f and the realization y of the future observation Y. The density forecast is ideal if the sampling density of Y is indeed f. Diebold, Gunther and Tay (1998) argue powerfully that the ideal forecaster is preferred by any rational user, irrespectively of the cost-loss structure at hand. Hence, it is critically important that a scoring rule be proper, in the sense that

$$\begin{aligned} E_f \, \mathrm{S}(f, Y) &= \int f(y) \, \mathrm{S}(f, y) \, \mathrm{d}y \\ &\leq \int f(y) \, \mathrm{S}(g, y) \, \mathrm{d}y = E_f \, \mathrm{S}(g, Y) \end{aligned} \tag{1}$$

for all density functions f and g. A scoring rule is strictly proper if (1) holds, with equality if and only if $f = g$ almost surely. Clearly, a strictly proper scoring rule prefers the ideal forecaster over any other. Prominent examples of strictly proper scoring rules include the logarithmic, quadratic, spherical, and continuous ranked probability scores (Matheson and Winkler 1976; Winkler 1996; Gneiting and Raftery 2007). We take scoring rules to be negatively oriented penalties, so the lower, the better.

Density forecast methods are then ranked by comparing their average scores. Specifically, if

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

then we prefer f if $\overline{\mathrm{S}}_n^f < \overline{\mathrm{S}}_n^g$, and prefer g otherwise. Amisano and Giacomini (2007) consider tests of equal forecast performance based on the test statistic

$$t_n = \sqrt{n} \, \frac{\overline{\mathrm{S}}_n^f - \overline{\mathrm{S}}_n^g}{\hat{\sigma}_n}, \tag{2}$$

where

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{j=-(k-1)}^{k-1} \sum_{t=m}^{m+n-|j|} \Delta_{t,k} \Delta_{t+|j|,k} \quad \text{and} \quad \Delta_{t,k} = \mathrm{S}(\hat{f}_{t+k}, y_{t+k}) - \mathrm{S}(\hat{g}_{t+k}, y_{t+k}), \tag{3}$$

Table 1: Weighted likelihood ratio tests for density forecasts for the conditionally heteroscedastic process (5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1} = \mathcal{N}(0, \frac{1}{2}\hat{\sigma}^2_{t+1})$ uses a deliberately misspecified predictive variance. The width of the sliding training window is $m = 100$, and we consider $n = 900$ one-step-ahead density forecasts. Counterintuitive test statistics are shown in bold. See text for details.

Weight Function	Emphasis	$\overline{\mathrm{S}}^f_n$	$\overline{\mathrm{S}}^g_n$	$\hat{\sigma}_n$	t_n	P
$w_0(x) = 1$	uniform	1.312	1.490	0.862	-6.20	< 0.001
$w_1(x) = \phi(x)$	center	0.294	0.267	0.100	**7.98**	< 0.001
$w_2(x) = 1 - \phi(x)/\phi(0)$	tails	0.575	0.759	0.645	-9.69	< 0.001
$w_3(x) = \Phi(x)$	right tail	0.667	0.633	0.535	-4.73	< 0.001
$w_4(x) = 1 - \Phi(x)$	left tail	0.645	0.542	0.510	-4.34	< 0.001

as proposed by Diebold and Mariano (1995). Assuming suitable regularity conditions, the statistic t_n is asymptotically standard normal under the null hypothesis of vanishing expected score differentials. In the case of rejection, f is chosen if t_n is negative and g is chosen if t_n is positive.[1]

What scoring rule should be used? Amisano and Giacomini (2007) consider a weighted logarithmic scoring rule,

$$\mathrm{S}(f, y) = w\left(\frac{y - \mu}{\sigma}\right) \mathrm{S}_0(f, y), \tag{4}$$

where w is a fixed, nonnegative weight function, μ and σ are estimates of the unconditional mean and standard deviation of the predictand, based on the past m observations, and S_0 is the logarithmic scoring rule, $\mathrm{S}_0(f, y) = -\log f(y)$. The weight function emphasizes regions of interest, such as the tails or the center of a variable's range. With ϕ and Φ denoting the standard normal probability density and cumulative distribution function, the weight functions $w_1(x) = \phi(x)$, $w_2(x) = 1 - \phi(x)/\phi(0)$, $w_3(x) = \Phi(x)$ and $w_4(x) = 1 - \Phi(x)$ emphasize the center, the tails, the right tail and the left tail, respectively. The approach of Mitchell and Hall (2005) and Bao, Lee and Saltoğlu (2007) employs the unweighted, original logarithmic score.

The weighting approach seems appealing; however, it corresponds to the use of an improper scoring rule and incurs misguided inferences. For instance, consider the GARCH(1,1) process $Y_1, Y_2, \ldots$, where

$$Y_{t+1} = \epsilon_{t+1}, \qquad \epsilon_{t+1} \sim \mathcal{N}(0, \sigma^2_{t+1}), \qquad \sigma^2_{t+1} = \alpha\epsilon^2_t + \beta\sigma^2_t + \gamma. \tag{5}$$

Following Christoffersen and Diebold (1996), we set the GARCH parameters at $\alpha = 0.2$ and $\beta = 0.75$, which are typical of estimates reported in the literature, and we let $\gamma = 0.05$, which

[1] Amisano and Giacomini (2007) use the logarithmic score in positive orientation, so they choose f if t_n is negative and g if t_n is positive.

normalizes the unconditional process variance to 1.[2] The rolling estimation window is of size $m = 100$, and we consider $n = 900$ density forecasts at the prediction horizon $k = 1$. The density forecast $\hat{f}_{t+1}$ is Gaussian with mean zero and variance $\hat{\sigma}^2$, which is derived from a GARCH fit for (5). Except for uncertainty in parameter estimation, this is the ideal density forecast. In contrast, the density forecast $\hat{g}_{t+1}$ is Gaussian with mean zero and variance one half time times $\hat{\sigma}^2$, deliberately misspecifying the conditional variance. Results for the weighted likelihood ratio test are shown in Table 1. Using the weight functions w_0, w_2, w_3 and w_4 the test prefers f, as expected. With weight function w_1, the test prefers the misspecified density forecast g, which is a counterintuitive result.

The goal of this paper is to propose a test that adopts the weighting approach of Amisano and Giacomini (2007), avoids misguided inferences, and comes with associated graphical tools that can be used to diagnose strengths and weaknesses of a forecasting method. We retain the test statistic (2), but base our test on appropriately weighted, proper versions of the continuous ranked probability score (CRPS; Matheson and Winkler 1976; Gneiting and Raftery 2007; Laio and Tamea 2007). Any density forecast f induces a probability forecast for the binary event $\{Y \leq z\}$ via the value of the associated cumulative distribution function (CDF)

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

at the threshold $z \in \mathbb{R}$. Similarly, it induces the quantile forecast $F^{-1}(\alpha)$ at the level $\alpha \in (0, 1)$. The continuous ranked probability score is then defined as

$$\mathrm{CRPS}(f, y) = \int_{-\infty}^{\infty} \mathrm{PS}(F(z), \mathbb{I}\{y \leq z\}) \, \mathrm{d}z = \int_{0}^{1} \mathrm{QS}_\alpha(F^{-1}(\alpha), y) \, \mathrm{d}\alpha, \tag{6}$$

where

$$\mathrm{PS}(p, \mathbb{I}\{y \leq z\}) = (p - \mathbb{I}\{y \leq z\})^2$$

is the Brier probability score (Selten 1998; Gneiting and Raftery 2007) for a probability forecast p of the binary event $\{Y \leq z\}$ at the threshold $z \in \mathbb{R}$, and

$$\mathrm{QS}_\alpha(q, y) = 2\,(\mathbb{I}\{y < q\} - \alpha)\,(q - y)$$

is the quantile score (Gneiting and Raftery 2007) for a quantile forecast q at the level $\alpha \in (0, 1)$. Here and in the following, the symbol $\mathbb{I}$ stands for an indicator function.

Following Matheson and Winkler (1976) and Gneiting and Raftery (2007), it is straightforward to construct weighted versions of the continuous ranked probability score (6) that emphasize regions of interest and retain propriety. A threshold weighted version of the continuous ranked probability score is obtained as

$$\mathrm{S}(f, y) = \int_{-\infty}^{\infty} \mathrm{PS}(F(z), \mathbb{I}\{y \leq z\}) \, u(z) \, \mathrm{d}z, \tag{7}$$

[2] See Engle (1982) and Bollerslev (1986) for details on ARCH and GARCH processes. We set the initial conditional variance equal to $\sqrt{609}/7$, that is, the unconditional variance plus one standard deviation of the conditional variance, and discard the first 1,000 values.

Table 2: Weighted CRPS tests for density forecasts for the conditionally heteroscedastic process (5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1} = \mathcal{N}(0, \frac{1}{2}\hat{\sigma}^2_{t+1})$ uses a deliberately misspecified predictive variance. The width of the sliding training window is $m = 100$, and we consider $n = 900$ one-step-ahead density forecasts. In contrast to the weighted likelihood ratio test, all tests prefer f over g.

Threshold Weight	Emphasis	$\overline{\mathrm{S}}^f_n$	$\overline{\mathrm{S}}^g_n$	$\hat{\sigma}_n$	t_n	P
$u_0(y) = 1$	uniform	0.511	0.521	0.070	-3.95	< 0.001
$u_1(y) = \phi(y)$	center	0.153	0.155	0.018	-4.24	< 0.001
$u_2(y) = 1 - \phi(y)/\phi(0)$	tails	0.129	0.132	0.030	-2.88	0.004
$u_3(y) = \Phi(y)$	right tail	0.258	0.262	0.046	-2.83	0.005
$u_4(y) = 1 - \Phi(y)$	left tail	0.254	0.259	0.046	-3.24	0.001

Quantile Weight	Emphasis	$\overline{\mathrm{S}}^f_n$	$\overline{\mathrm{S}}^g_n$	$\hat{\sigma}_n$	t_n	P
$v_0(q) = 1$	uniform	0.511	0.521	0.070	-3.95	< 0.001
$v_1(q) = q(1-q)$	center	0.100	0.101	0.009	-2.79	0.005
$v_2(q) = (2q-1)^2$	tails	0.113	0.118	0.036	-4.85	< 0.001
$v_3(q) = q^2$	right tail	0.157	0.161	0.041	-2.53	0.014
$v_4(q) = (1-q)^2$	left tail	0.155	0.159	0.041	-3.00	0.003

where u is a nonnegative weight function on the real line. Similarly, a quantile weighted version is obtained as

$$\mathrm{S}(f, y) = \int_0^1 \mathrm{QS}_\alpha(F^{-1}_\alpha, y)\, v(\alpha)\, \mathrm{d}\alpha, \tag{8}$$

where v is a nonnegative weight function on the unit interval. For a constant weight function, both (7) and (8) reduce to the unweighted score (6).

Table 2 returns to the simulation study for the GARCH model (5) and reports results based on the test statistic (2) and threshold or quantile weighted versions of the continuous ranked probability score, which are proper, as opposed to the weighted logarithmic score. In contrast to the results for the weighted likelihood ratio test, all t_n values in Table 2 are negative, favoring the nearly ideal density forecast f over its deliberately misspecified competitor g.

The remainder of the paper is organized as follows. In Section 2 we show that the weighted likelihood ratio test incurs the use of an improper scoring rule, and explore ways in which the test can be hedged. In Section 3 we study threshold and quantile weighted versions of the continuous ranked probability score in further detail, and discuss conditions under which the test statistic t_n is asymptotically standard normal. We also note graphical representations of the threshold and quantile decomposition of the continuous ranked probability score, which can be used diagnostically to assess strengths and deficiencies of forecasting techniques.

Section 4 applies these methods to compare density forecasts for quarterly inflation rates in the United Kingdom and wind resources in the North American Pacific Northwest. The paper closes with a discussion in Section 5.

2 Hedging strategies for the weighted likelihood ratio test

Recall that a scoring rule $\mathrm{S}(f, y)$ for a density forecast is proper if

$$\begin{aligned} E_f\,\mathrm{S}(f, Y) &= \int f(y)\,\mathrm{S}(f, y)\,\mathrm{d}y \\ &\le \int f(y)\,\mathrm{S}(g, y)\,\mathrm{d}y = E_f\,\mathrm{S}(g, Y) \end{aligned}$$

for all density functions f and g. It is strictly proper if the above holds, with equality if and only if $f = g$ almost surely. Examples of proper scoring rules for density forecasts include the logarithmic score, $\mathrm{S}(f, y) = -\log f(y)$, the quadratic score, $\mathrm{S}(f, y) = -2f(y) + \|f\|^2$, and the spherical score $\mathrm{S}(f, y) = -f(y)/\|f\|$, where

$$\|f\| = \left(\int_{-\infty}^{\infty} f(y)^2\,\mathrm{d}y \right)^{1/2}.$$

The continuous ranked probability score and its weighted versions are also proper (Matheson and Winkler 1976; Gneiting and Raftery 2007).

The following result shows that if $\mathrm{S}_0(f, y)$ is a strictly proper scoring rule, then its product with a weight function $w(y)$ is improper, unless the weight function is constant.

Theorem 2.1. *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 weighted score*

$$\mathrm{S}(g, Y) = w(Y)\,\mathrm{S}_0(g, Y) \tag{9}$$

is minimized if we issue the density forecast

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

Proof. Let h be any density forecast. Then

$$\begin{aligned} E_f\,\mathrm{S}(g, Y) &= \int w(y) f(y)\,\mathrm{S}_0(g, y)\,\mathrm{d}y = \int w(y) f(y)\,\mathrm{d}y \int g(y)\,\mathrm{S}_0(g, y)\,\mathrm{d}y \\ &\le \int w(y) f(y)\,\mathrm{d}y \int g(y)\,\mathrm{S}_0(h, y)\,\mathrm{d}y = \int w(y) f(y)\,\mathrm{S}_0(h, y)\,\mathrm{d}y = E_f\,\mathrm{S}(h, Y), \end{aligned}$$

where the inequality reflects the propriety of S_0. □

Table 3: Weighted likelihood ratio tests for density forecasts for the conditionally heteroscedastic process (5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1}$ is deliberately misspecified as described in (10). Counterintuitive test statistics are shown in bold. See text for details.

Weight Function	Emphasis	$\overline{\mathrm{S}}^f_n$	$\overline{\mathrm{S}}^g_n$	$\hat{\sigma}_n$	t_n	P
$w_0(x) = 1$	uniform	1.312	1.611	0.727	−12.31	< 0.001
$w_1(x) = \phi(x)$	center	0.294	0.436	0.215	−19.84	< 0.001
$w_2(x) = 1 - \phi(x)/\phi(0)$	tails	0.575	0.518	0.331	**5.23**	< 0.001
$w_3(x) = \Phi(x)$	right tail	0.667	0.744	0.515	−4.48	< 0.001
$w_4(x) = 1 - \Phi(x)$	left tail	0.645	0.867	0.310	−21.51	< 0.001

In particular, we are now in a position to explain the failure of the weighted likelihood ratio test in the simulation example in the introduction. The weighted logarithmic score (4) is similar to the composite scoring rule (9) where S_0 is the logarithmic score, and the composite score is improper, unless the weight function is constant. Moreover, Theorem 2.1 suggests a hedging strategy if forecasters are compared by the weighted likelihood ratio test, namely to issue the density function g that is proportional to the product of the forecaster's true belief, f, and the weight function, w. For example, if both $f = \phi$ and $w = \phi$ are standard normal, the suggested hedge uses a normal density function g with mean zero and variance one half. Essentially, this is the situation in the simulation study in the introduction. The misspecified density forecast $\hat{g}_{t+1}$ halves the estimated Gaussian variance; hence, to a good degree of approximation, it is proportional to the product of the true belief $\hat{f}_{t+1}$ and the weight function $w_1 = \phi$. Not surprisingly, the weighted likelihood ratio test with weight function w_1 fails.

Before closing this section, we present another simulation study in which the weighted likelihood ratio test yields counterintuitive results. Once again, we study density forecasts for the conditionally heteroscedastic process (5) with parameter values $\alpha = 0.2$, $\beta = 0.75$ and $\gamma = 0.05$. The rolling estimation window is of size $m = 100$, and we issue $n = 900$ density forecasts at the prediction horizon $k = 1$. As previously, the density forecast $\hat{f}_{t+1}$ is Gaussian with mean zero and variance $\hat{\sigma}^2_{t+1}$, derived from a GARCH fit under the correct model specification. Except for estimation uncertainty, this is the ideal density forecast. Its competitor is the density forecast $\hat{g}_{t+1}$, which is deliberately misspecified as

$$\hat{g}_{t+1}(y) = \hat{f}_{t+1}(y) \left(\mathbb{I}\{y < -\hat{\sigma}_{t+1}\} + \frac{1}{2}\,\mathbb{I}\{|y| \le \hat{\sigma}_{t+1}\} + \frac{1}{2(1-\Phi(1))}\,\mathbb{I}\{y > \hat{\sigma}_{t+1}\} \right). \tag{10}$$

Note that $\hat{g}_{t+1}$ is identical to $\hat{f}_{t+1}$ in the left tail, underspecifies the center of the distribution, and makes this up in the right tail. Table 3 shows results for the weighted likelihood ratio test, which are misguided and inconsistent. Specifically, the test suggests that both in the left tail and in the right tail f is preferable. Looking at both tails simultaneously, the test stipulates that g is better.

3 Weighting and testing with the continuous ranked probability score

3.1 Threshold and quantile weighting for the continuous ranked probability score

Suppose that the density forecast is f and y realizes. Let F denote the CDF associated with the density f, and write $F^{-1}(\alpha)$ for the quantile at level $\alpha \in (0, 1)$. The continuous ranked probability score then can be defined in three equivalent ways, as

$$\begin{aligned} \mathrm{CRPS}(f, y) &= E_F|Y - y| - \frac{1}{2} E_F|Y - Y'| && (11) \\ &= \int_{-\infty}^{\infty} (F(z) - \mathbb{I}\{y \le z\})^2 \, \mathrm{d}z && (12) \\ &= 2 \int_0^1 (\mathbb{I}\{y < F^{-1}(\alpha)\} - \alpha) \, (F^{-1}(\alpha) - y) \, \mathrm{d}\alpha, && (13) \end{aligned}$$

where Y and Y' are independent random variables with common distribution F. Gneiting and Raftery (2007) showed the equivalence of the kernel score representation (11) and the standard form (12), to which we refer as the threshold decomposition of the continuous ranked probability score. The equivalence to (13), to which we refer as the quantile decomposition of the score, was noted by Laio and Tamea (2007). Both (11) and (13) show that the continuous ranked probability score is reported in the same unit as the observations. The score is strictly proper within the class of forecast densities that have finite first moment, and attains an infinite value otherwise. It applies to predictive distributions with discrete components and reduces to the absolute error in the case of a point forecast.

The integrand in (12) equals the quadratic or Brier probability score (Selten 1998; Gneiting and Raftery 2007)

$$\mathrm{PS}(p, \mathbb{I}\{y \le z\}) = (p - \mathbb{I}\{y \le z\})^2$$

for the probability forecast $p = F(z)$ of the binary event $\{Y \le z\}$ at the threshold $z \in \mathbb{R}$. The integrand in (13) equals the quantile score

$$\mathrm{QS}_\alpha(q, y) = 2 \, (\mathbb{I}\{y < q\} - \alpha) \, (q - y)$$

for the quantile forecast $q = F^{-1}(\alpha)$ (Cervera and Muñoz 1996; Gneiting and Raftery 2007). It has also been referred to as the tick loss function (Giacomini and Komunjer 2005) or, more traditionally, as the asymmetric linear or lin-lin loss function (Koenker and Basset 1978; Christoffersen and Diebold 1996).

Using the Brier probability score, we define threshold weighted versions of the continuous ranked probability score as

$$\mathrm{S}(f, y) = \int_{-\infty}^{\infty} \mathrm{PS}(F(z), \mathbb{I}\{y \le z\}) \, u(z) \, \mathrm{d}z, \tag{14}$$

Table 4: Proposed weight functions for threshold and quantile weighted versions of the continuous ranked probability score. The threshold weight functions are specified in terms of the probability density function $\phi_{a,b}$ and the cumulative distribution function $\Phi_{a,b}$ of the normal distribution with mean a and standard deviation b.

Emphasis	Threshold Weight Function	Quantile Weight Function
center	$u_1(y) = \phi_{a,b}(y)$	$v_1(q) = q(1-q)$
tails	$u_2(y) = 1 - \phi_{a,b}(y)/\phi_{a,b}(0)$	$v_2(q) = (2q-1)^2$
right tail	$u_3(y) = \Phi_{a,b}(y)$	$v_3(q) = q^2$
left tail	$u_4(y) = 1 - \Phi_{a,b}(y)$	$v_4(q) = (1-q)^2$

where u is a nonnegative weight function on the real line; if $u \equiv 1$, this reduces to the unweighted score (12). Table 4 lists some potential weight functions that emphasize the center or tails of a variable's range. The threshold weight functions resemble the suggestions of Amisano and Giacomini (2007); however, in our implementation, the parameters are fixed and user specified, depending on the application at hand. For instance, in the case of inflation rates we set the location parameter a at the policy target. If the weight function is integrable, such as in the case of the center weight function $\phi_{a,b}$, the threshold-weighted continuous ranked probability score (14) is finite and bounded by the integral of the weight function. Other options for integrable weight functions with center emphasis include t and Laplace densities.

Similarly, we define quantile weighted versions of the continuous ranked probability score as

$$\mathrm{S}(f, y) = \int_0^1 \mathrm{QS}_\alpha(F^{-1}(\alpha), y)\, v(\alpha)\, \mathrm{d}\alpha, \tag{15}$$

where v is a nonnegative weight function on the unit interval. If $v \equiv 1$, we recover the unweighted score (13). Table 4 suggests weight functions with center or tail emphasis. The two weighting approaches can be traced back at least to Matheson and Winkler (1976); they retain propriety, because convex sums and limits of proper scoring rules remain proper. The threshold weighting idea is also employed by Corradi and Swanson (2006a, pp. 194–195), though their emphases and terminology differ from ours.

Closed form expressions for the evaluation of (14) or (15) may or may not be available; however, the computation of a suitably discretized approximate version is always feasible, to any degree of accuracy. In the case of threshold weighting, we approximate (14) by

$$\mathrm{S}(f, y) = \frac{y_u - y_l}{I - 1} \sum_{i=1}^{I} w(y_i)\, \mathrm{PS}(F(y_i), \mathbb{I}\{y \le y_i\}) \quad \text{where} \quad y_i = y_l + i\, \frac{y_u - y_l}{I} \tag{16}$$

and (y_l, y_u) is the range of interest. In the case of the quantile weighted score, we approximate the integral in (15) by a discrete version,

$$\mathrm{S}(f,y) = \frac{1}{J-1} \sum_{j=1}^{J-1} v(\alpha_j)\,\mathrm{QS}_{\alpha_j}(F^{-1}(\alpha_j),y) \quad \text{where} \quad \alpha_j = \frac{j}{J}. \tag{17}$$

Note that the discrete versions themselves are proper scoring rules, that arise as special cases in (14) and (15) if the integral is taken with respect to a discrete Stieltjes measures rather than a weight function.

3.2 Asymptotic normality of the test statistic

Following Amisano and Giacomini (2007), we consider tests of equal forecast performance based on the test statistic

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

where

$$\overline{\mathrm{S}}_n^f = \frac{1}{n-k+1} \sum_{t=m}^{m+n-k} \mathrm{S}(\hat{f}_{t+k}, y_{t+k}) \quad \text{and} \quad \overline{\mathrm{S}}_n^g = \frac{1}{n-k+1} \sum_{t=m}^{m+n-k} S(\hat{g}_{t+k}, y_{t+k}) \tag{18}$$

and $\hat{\sigma}_n^2$ is defined in (3). Under general conditions, t_n is asymptotically standard normal under the null hypothesis of vanishing expected score differentials, and the test will reject with probability tending to 1 under a fixed alternative. When S is a weighted logarithmic rule, Amisano and Giacomini (2007) prove these claims under regularity assumptions[3], which include a mixing condition on the process $\{\boldsymbol{Z}_t\}$ defined in the introduction, boundedness of the weight function, consistency of $\hat{\sigma}_n^2$ as an estimate of

$$\sigma_n^2 = \mathrm{var}(\sqrt{n}\,(\overline{\mathrm{S}}_n^f - \overline{\mathrm{S}}_n^g)) > 0,$$

and moment conditions. In our case, in which S is a weighted version of the continuous ranked probability score, the same result holds, except for the moment condition, which now requires that

$$E_{\hat{f}_{t+k}}|X|, \quad E_{\hat{g}_{t+k}}|X| \quad \text{and} \quad E_{t+k}|Y_{t+k}|^{2r} \quad \text{are finite for all} \quad t, \tag{19}$$

where the power $r \geq 2$ depends on the mixing condition. In the case of threshold weighting with an integrable (rather than just bounded) weight function, the moment condition can be

[3] Amisano and Giacomini consider the case $k = 1$ only. The extension to a general prediction horizon $k \geq 1$ is straightforward. We wish to emphasize that our aforementioned concerns are not with the asymptotic arguments in Amisano and Giacomini (2007) nor with the weighting idea, which is appealing indeed. However, we disagree with the particular choice of a weighted logarithmic scoring rule for the test, which can lead to rejection in favor of an inferior forecast.

dropped. In analogy to the arguments of Amisano and Giacomini (2007), these results can be proved by verifying the assumptions of Theorem 4 of Giacomini and White (2006). The only novel argument is in the derivation of the moment condition (19), which is presented in an appendix.

In practical applications, the full set of assumptions cannot be verified; yet, the assumptions are plausible as approximations. Recall that the continuous ranked probability score attains an infinite value if the forecast density has infinite first moment. In this light, the first two conditions in (19) assure that each individual score is finite. The third condition stipulates that the true data generating density has a finite moment of order $2r$, where typically one can take $r = 2$. Hence, as a rule of thumb, the normal approximation for t_n is appropriate, unless the forecast densities have infinite moments of low order. In the case of threshold weighting with an integrable weight function, the moment condition can just be ignored.

Table 5 summarizes results for weighted CRPS tests in the simulation example of Section 2. The density forecasts $\hat{f}_{t+1}$ and $\hat{g}_{t+1}$ and the true data generating density have Gaussian tails, so the normal approximation for t_n is justified. In contrast to the respective results for weighted likelihood ratio tests, all t_n values are strongly negative, favoring f over its deliberately misspecified competitor g.

3.3 Forecast diagnostics via threshold and quantile decomposition

The threshold and quantile decompositions of the continuous ranked probability score carry over to mean scores, and in the latter form they can be used diagnostically, to assess strengths and deficiencies of density forecasting techniques.

Consider a mean score of the form (18). The threshold decomposition (12) applies to the mean score, in that

$$\overline{\mathrm{CRPS}}_n^f = \int_{-\infty}^{\infty} \overline{\mathrm{PS}}_n^f(z)\,\mathrm{d}z \tag{20}$$

where

$$\overline{\mathrm{PS}}_n^f(z) = \frac{1}{n-k+1} \sum_{t=m}^{m+n-k} \mathrm{PS}(\hat{F}_{t+k}(z), y_{t+k}) \tag{21}$$

denotes the mean Brier probability score for the probability forecast of the binary event $\{Y_{t+k} \leq z\}$ at the threshold $z \in \mathbb{R}$. Schumacher, Graf and Gerds (2003) and Gneiting, Balabdaoui and Raftery (2007) proposed a plot of the mean Brier score (21) versus z as a diagnostic tool and coined the terms prediction error curve and Brier score plot, respectively. The representation (20) shows that the plot illustrates the threshold decomposition of the continuous ranked probability score.

Similarly, the quantile decomposition (13) suggests the representation

$$\overline{\mathrm{CRPS}}_n^f = \int_{0}^{1} \overline{\mathrm{QS}}_n^f(\alpha)\,\mathrm{d}z, \tag{22}$$

where

$$\overline{\mathrm{QS}}^{f}_{n}(\alpha) = \frac{1}{n-k+1} \sum_{t=m}^{m+n-k} \mathrm{QS}_{\alpha}(\hat{F}^{-1}_{t+k}(\alpha), y_{t+k}). \tag{23}$$

Laio and Tamea (2007) proposed a plot of the mean quantile score (23) versus α as a diagnostic tool in the assessment of density forecasts. We adopt their suggestion and note that it illustrates the quantile decomposition (22) of the continuous ranked probability score.

Figure 1 applies the threshold decomposition (20) and quantile decomposition (22) to the density forecasting techniques f and g in the simulation study described in Sections 2 and 3.2. It is apparent that f and g are on equal footing in the lower tail, but f is superior in the center, which is in accordance with (10). As shown in Table 5, the mean continuous ranked probability score is 0.483 for f and 0.592 for g; this equals the integral under the respective curves. The weighted scores in the table correspond to weighted integrals.

4 Case studies

4.1 Bank of England projections of quarterly inflation rates

The Bank of England's Monetary Policy Committee (MPS) has issued probabilistic forecasts of inflation rates and gross domestic product every quarter since February 1996 and November 1997, respectively, using fan charts to visualize the deciles of the predictive distributions (Wallis 2003, 2004; Clements 2004; Elder, Kapetanios, Taylor and Yates 2005; Mitchell and Hall 2005).[4]

We compare the Bank of England's density forecasts of inflation rates (RPIX) to those derived from a simplistic autoregressive time series model. The Bank of England employs potentially asymmetric two-piece normal distributions with parameters $\mu \in \mathbb{R}$ and $\sigma_1, \sigma_2 > 0$ and forecast density

$$f(y) = \begin{cases} \left(\frac{\pi}{2}\right)^{-1/2} (\sigma_1 + \sigma_2)^{-1} \exp\left(-\frac{(y-\mu)^2}{2\sigma_1^2}\right) & \text{if} \quad y \le \mu, \\ \left(\frac{\pi}{2}\right)^{-1/2} (\sigma_1 + \sigma_2)^{-1} \exp\left(-\frac{(y-\mu)^2}{2\sigma_2^2}\right) & \text{if} \quad y \ge \mu. \end{cases}$$

The simplistic competitor is a Gaussian autoregression of order one that uses a rolling estimation window of length $m = 6$ quarters. This method results in Gaussian density forecasts.

[4] The quarterly Bank of England inflation report is available online at `http://www.bankofengland.co.uk/publications/inflationreport/`. Archived forecasts can be downloaded at `http://www.bankofengland.co.uk/publications/inflationreport/irprobab.htm`. Observed RPIX inflation rates are available at `http://www.statistics.gov.uk/StatBase/tsdataset.asp?vlnk=7173&More=Y` under Office of National Statistics code CDKQ. The rates are percentage changes over 12 months. The first quarter ranges from March to May, the second from June to August, and so on.

Table 5: Threshold and quantile weighted CRPS tests for density forecasts for the conditionally heteroscedastic process (5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1}$ is deliberately misspecified as described in (10).

Threshold Weight	Emphasis	$\overline{\mathrm{S}}_n^f$	$\overline{\mathrm{S}}_n^g$	$\hat{\sigma}_n$	t_n	P
$u_0(y) = 1$	uniform	0.511	0.625	0.317	-10.72	< 0.001
$u_1(y) = \phi(y)$	center	0.153	0.184	0.095	-10.01	< 0.001
$u_2(y) = 1 - \phi(y)/\phi(0)$	tails	0.129	0.163	0.097	-10.39	< 0.001
$u_3(y) = \Phi(y)$	right tail	0.258	0.343	0.227	-11.33	< 0.001
$u_4(y) = 1 - \Phi(y)$	left tail	0.254	0.281	0.098	-8.39	< 0.001

Quantile Weight	Emphasis	$\overline{\mathrm{S}}_n^f$	$\overline{\mathrm{S}}_n^g$	$\hat{\sigma}_n$	t_n	P
$v_0(q) = 1$	uniform	0.511	0.625	0.317	-10.72	< 0.001
$v_1(q) = q(1-q)$	center	0.100	0.125	0.069	-10.99	< 0.001
$v_2(q) = (2q-1)^2$	tails	0.113	0.125	0.045	-7.98	< 0.001
$v_3(q) = q^2$	right tail	0.157	0.198	0.116	-10.44	< 0.001
$v_4(q) = (1-q)^2$	left tail	0.155	0.177	0.069	-9.60	< 0.001

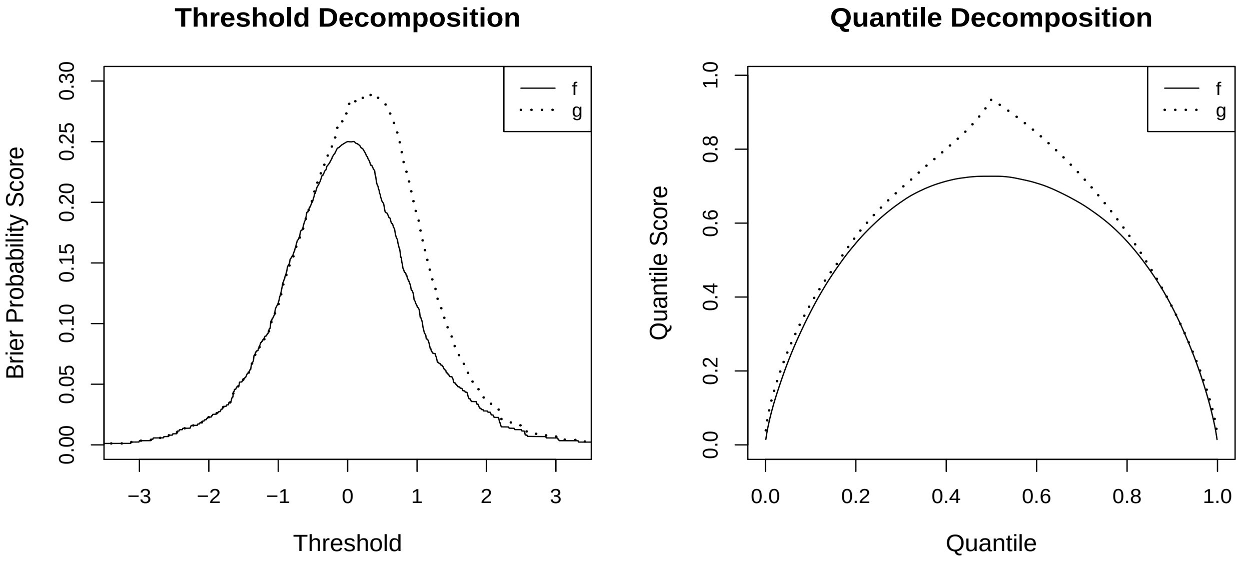

Figure 1: Threshold and quantile decomposition of the mean continuous ranked probability score for density forecasts for the conditionally heteroscedastic process (5). The density forecast $\hat{f}_{t+1} = \mathcal{N}(0, \hat{\sigma}^2_{t+1})$ is estimated under the correct model assumption. Its competitor $\hat{g}_{t+1}$ is deliberately misspecified as described in (10).

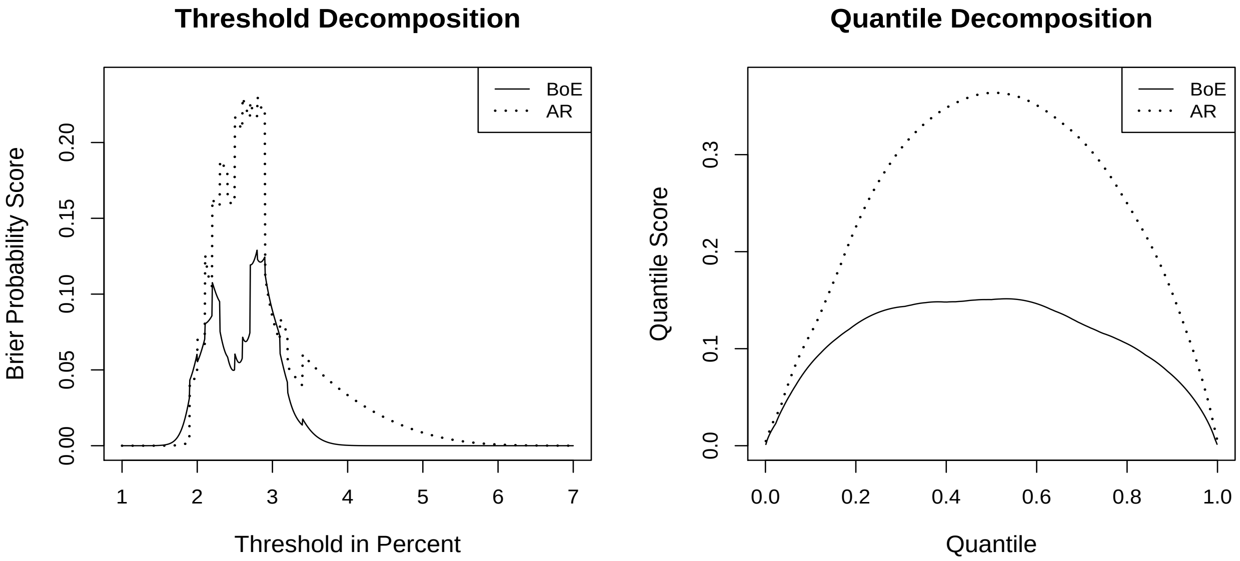

Figure 2: Threshold and quantile decomposition of the mean continuous ranked probability score for Bank of England (BoE) and autoregressive (AR(1)) density forecasts of inflation rates, at a prediction horizon of one quarter.

Table 6: Threshold and quantile weighted CRPS tests for density forecasts of inflation rates, at a prediction horizon of one quarter, in percent. The Bank of England forecast takes the role of f and the autoregressive benchmark the role of g.

Threshold Weight	Emphasis	$\overline{\mathrm{S}}_n^{\mathrm{BoE}}$	$\overline{\mathrm{S}}_n^{\mathrm{AR}}$	$\hat{\sigma}_n$	t_n	P
$u_0(y) = 1$	uniform	0.112	0.246	0.248	-3.62	< 0.001
$u_1(y) = \phi_{2.5,1}(y)$	center	0.041	0.081	0.064	-4.16	< 0.001
$u_2(y) = 1 - \phi_{2.5,1}(y)/\phi_{2.5,1}(2)$	tails	0.010	0.044	0.137	-1.69	0.090
$u_3(y) = \Phi_{2.5,1}(y)$	right tail	0.061	0.152	0.200	-3.07	0.002
$u_4(y) = 1 - \Phi_{2.5,1}(y)$	left tail	0.051	0.094	0.076	-3.75	< 0.001

Quantile Weight	Emphasis	$\overline{\mathrm{S}}_n^{\mathrm{BoE}}$	$\overline{\mathrm{S}}_n^{\mathrm{AR}}$	$\hat{\sigma}_n$	t_n	P
$v_0(q) = 1$	uniform	0.112	0.246	0.248	-3.62	< 0.001
$v_1(q) = q(1-q)$	center	0.022	0.049	0.050	-3.67	< 0.001
$v_2(q) = (2q-1)^2$	tails	0.026	0.050	0.049	-3.35	< 0.001
$v_3(q) = q^2$	right tail	0.033	0.077	0.078	-3.78	< 0.001
$v_4(q) = (1-q)^2$	left tail	0.036	0.071	0.076	-3.12	0.002

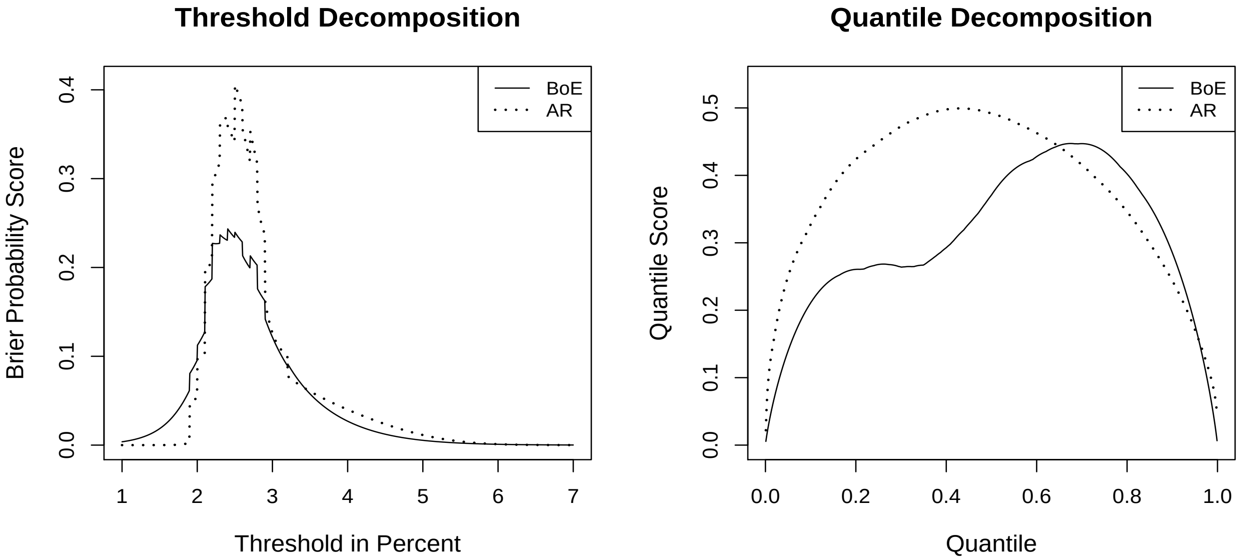

Figure 3: Threshold and quantile decomposition of the mean continuous ranked probability score for Bank of England (BoE) and autoregressive (AR) density forecasts of inflation rates, at a prediction horizon of seven quarters.

Table 7: Threshold and quantile weighted CRPS tests for density forecasts of inflation rates, at a prediction horizon of seven quarters. The Bank of England forecast takes the role of f and the autoregressive benchmark the role of g.

Threshold Weight	Emphasis	$\overline{\mathrm{S}}_n^{\mathrm{BoE}}$	$\overline{\mathrm{S}}_n^{\mathrm{AR}}$	$\hat{\sigma}_n$	t_n	P
$u_0(y) = 1$	uniform	0.304	0.381	0.437	-1.19	0.235
$u_1(y) = \phi_{2.5,1}(y)$	center	0.102	0.129	0.131	-1.43	0.152
$u_2(y) = 1 - \phi_{2.5,1}(y)/\phi_{2.5,1}(2)$	tails	0.049	0.057	0.166	-0.30	0.761
$u_3(y) = \Phi_{2.5,1}(y)$	right tail	0.170	0.226	0.324	-1.15	0.251
$u_4(y) = 1 - \Phi_{2.5,1}(y)$	left tail	0.134	0.155	0.148	-0.99	0.321

Quantile Weight	Emphasis	$\overline{\mathrm{S}}_n^{\mathrm{BoE}}$	$\overline{\mathrm{S}}_n^{\mathrm{AR}}$	$\hat{\sigma}_n$	t_n	P
$v_0(q) = 1$	uniform	0.304	0.381	0.437	-1.19	0.235
$v_1(q) = q(1-q)$	center	0.057	0.072	0.081	-1.23	0.217
$v_2(q) = (2q-1)^2$	tails	0.077	0.095	0.124	-0.96	0.338
$v_3(q) = q^2$	right tail	0.108	0.111	0.080	-0.24	0.813
$v_4(q) = (1-q)^2$	left tail	0.083	0.127	0.263	-1.14	0.255

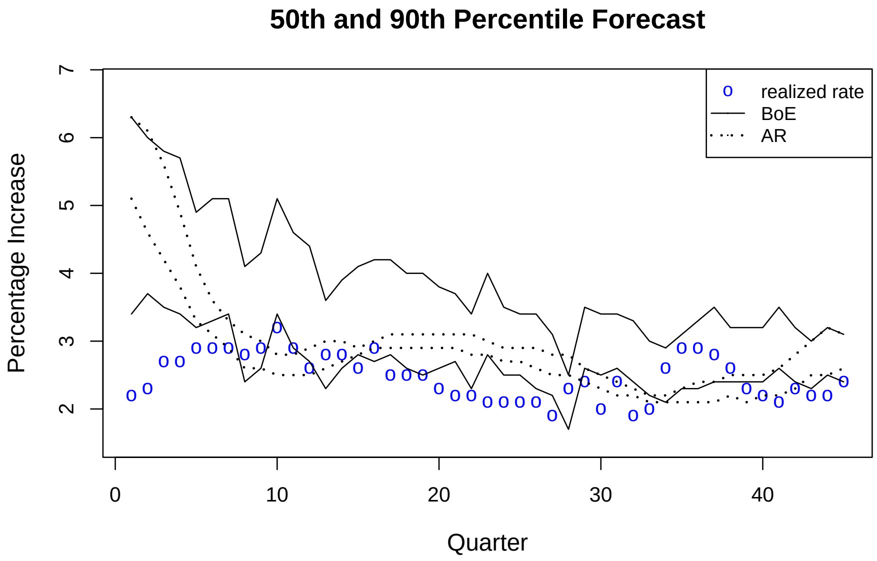

Figure 4: Bank of England (BoE) and autoregressive (AR) forecasts of inflation rates, at a prediction horizon of seven quarters ahead, for the third quarter of 1994 through the third quarter of 2005. The plot shows the 50th and 90th percentiles of the density forecasts for the two methods along with the observed rates.

Figure 2 and Table 6 compare the two methods at a prediction horizon of $k = 1$ quarters ahead, for a test period ranging from the first quarter of 1993 to the first quarter of 2004, for a total of $n = 45$ density forecast cases. Figure 2 shows the threshold and quantile decompositions (20) and (22) of the continuous ranked probability score for the two techniques. The Bank of England forecast has a clear edge at almost all thresholds and quantiles, with a mean continuous ranked probability score of 0.112%, as opposed to 0.246% for the autoregressive forecast. The integrals under the respective curves in Figure 2 equal these values. The superiority of the Bank of England forecast is corroborated by Table 6, which reports the results of weighted CRPS tests, using the weight functions of Table 4, where $a = 2.5\%$ equals the MPC's 1997–2003 policy target and $b = 1.0\%$ reflects the relative constancy of the inflation rate during the evaluation period.

Figure 3 and Table 7 show results at a prediction horizon of $k = 7$ quarters ahead, for the third quarter of 1994 (September through November) to the third quarter of 2005. Perhaps surprisingly, the dominance of the Bank of England forecast is much less pronounced. In Figure 3, the simplistic autoregressive forecast seems competitive at moderately large thresholds and quantiles. The mean continuous ranked probability score is 0.304% for the Bank of England forecast, as opposed to 0.382% for the autoregressive forecast. None of the tests in Table 7 rejects the null hypothesis of vanishing expected score differentials.

To explain this we point at Figure 4, which shows quantiles of the two density forecasts at a prediction horizon of seven quarters along with the realized inflation rates. The 90th percentile of the Bank of England forecast was much too conservative, resulting in unnecessarily wide prediction intervals that are penalized by the scores.

4.2 Probabilistic forecasts of wind resources at the Stateline wind energy center

With the proliferation of wind power, probabilistic short-term forecasts of wind resources at wind energy sites are becoming a critical requirement. Gneiting, Larson, Westrick, Genton and Aldrich (2006) introduced the regime-switching space-time (RST) technique that merges meteorological and statistical expertise to obtain accurate and calibrated, fully probabilistic forecasts of wind speed and wind power. Briefly, the RST method identifies forecast regimes at the wind energy site and fits a conditionally heteroscedastic predictive model for each regime. Geographically dispersed meteorological observations in the vicinity of the wind farm are used as predictor variables. The forecast densities are truncated normal.

Gneiting et al. (2006) applied the RST technique to obtain probabilistic forecasts of hourly average wind speed near the Stateline wind energy center in the states of Oregon and Washington, at a prediction horizon of $k = 2$ hours. In what follows, we compare the RST density forecasts to probabilistic forecasts derived from autoregressive time series models, as proposed by Brown, Katz and Murphy (1984) and widely implemented since. Both methods employ a rolling estimation window of 45 days or $1,080$ hours. The evaluation period ranges from 1 May through 30 November 2003, for a total of $n = 5,136$ density forecast cases. See Gneiting et al. (2006) for details.[5]

Figure 5 shows the threshold and quantile decomposition of the continuous ranked probability score for the two probabilistic forecasting methods. The RST technique is superior at all thresholds and quantiles, with a mean continuous ranked probability score of 0.961 meters per second, as opposed to 1.115 meters per second for the autoregressive benchmark. Table 8 shows the results of weighted CRPS tests with the weight functions in Table 4, where $a = 10$ meters per second and $b = 5$ meters per second, a choice that is motivated by the marginal climatological distribution of wind speeds (Gneiting et al. 2006). All tests are overwhelmingly in favor of the RST technique.

5 Discussion

We have proposed a method for comparing density forecasts that is based on threshold and quantile weighted versions of the continuous ranked probability score. R code is available from the authors upon request.

[5] Gneiting et al. (2006) refer to the methods considered here as the RST-D-CH and AR-D-CH techniques. The autoregressive method assumes Gaussian forecast densities that assign small but positive probability mass to the negative halfaxis, which we reassign to wind speed zero. The continuous ranked probability score handles the point mass naturally.

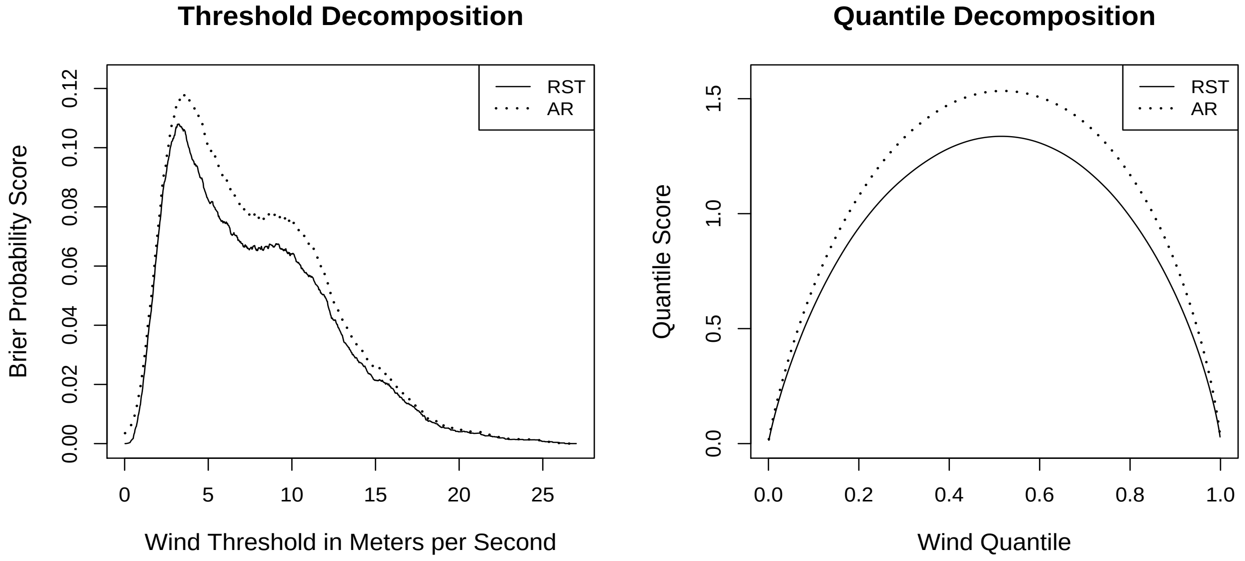

Figure 5: Threshold and quantile decomposition of the mean continuous ranked probability score for regime-switching space-time (RST) and autoregressive (AR) probabilistic forecasts of hourly average wind speed at the Stateline wind energy center, at a prediction horizon of two hours.

Table 8: Threshold and quantile weighted CRPS tests in the wind example. The regime-switching space-time (RST) forecast takes the role of f and the autoregressive benchmark the role of g.

Threshold Weight	Emphasis	$\overline{\mathrm{S}}_n^{\mathrm{RST}}$	$\overline{\mathrm{S}}_n^{\mathrm{AR}}$	$\hat{\sigma}_n$	t_n	P
$u_0(y) = 1$	uniform	0.961	1.115	0.838	-13.16	< 0.001
$u_1(y) = \phi_{10,5}(y)$	center	0.257	0.300	0.251	-12.28	< 0.001
$u_2(y) = 1 - \phi_{10,5}(y)/\phi_{10,5}(0)$	tails	0.318	0.364	0.293	-11.25	< 0.001
$u_3(y) = \Phi_{10,5}(y)$	right tail	0.342	0.398	0.386	-10.36	< 0.001
$u_4(y) = 1 - \Phi_{10,5}(y)$	left tail	0.619	0.718	0.552	-12.73	< 0.001

Quantile Weight	Emphasis	$\overline{\mathrm{S}}_n^{\mathrm{RST}}$	$\overline{\mathrm{S}}_n^{\mathrm{AR}}$	$\hat{\sigma}_n$	t_n	P
$v_0(q) = 1$	uniform	0.961	1.115	0.838	-13.16	< 0.001
$v_1(q) = q(1 - q)$	center	0.187	0.216	0.162	-12.93	< 0.001
$v_2(q) = (2q - 1)^2$	tails	0.213	0.250	0.201	-13.07	< 0.001
$v_3(q) = q^2$	right tail	0.299	0.351	0.302	-12.34	< 0.001
$v_4(q) = (1 - q)^2$	left tail	0.288	0.331	0.252	-12.30	< 0.001

Our approach is similar in spirit to the weighted likelihood ratio test of Amisano and Giacomini (2007); however, it is based on proper scoring rules, and therefore avoids misguided inferences. In the case of threshold weighting, it is formally equivalent to the approach of Corradi and Swanson (2006), who provide a wealth of relevant theoretical results under rolling and recursive estimation schemes. The threshold and quantile decompositions of the continuous ranked probability score can be illustrated graphically, to provide diagnostic tools that prompt insights into the strengths and deficiencies of forecasting methods, as we have illustrated in the case studies.

Gneiting, Balabdaoui and Raftery (2007) contend that the goal of probabilistic forecasting is to maximize the sharpness of the forecast densities subject to calibration. Calibration refers to the statistical consistency between the forecast densities and the observations, and is a joint property of the forecasts and the values that materialize. Sharpness refers to the concentration of the forecast densities: The sharper the densities, the less the uncertainty, and the sharper, the better, subject to calibration.

The probability integral transform (PIT) histogram is the primary diagnostic tool for calibration checks (Diebold, Gunther and Tay 1998; Corradi and Swanson 2006b; Gneiting, Balabdaoui and Raftery 2007; Laio and Tamea 2007). The PIT is simply the value that the predictive CDF attains at the observation (Dawid 1984). If the observation is drawn from the forecast density, the PIT has a uniform distribution. Hence, to assess the calibration of a density forecasting method, we find the PIT, repeat over a sizable number of forecast cases, and check the PIT histogram for uniformity. This does not take the sharpness of the density forecasts into account, as opposed to proper scoring rules, which provide a combined assessment of calibration and sharpness (Gneiting, Balabdaoui and Raftery 2007).

A possible limitation of our method is that the unweighted continuous ranked probability score is infinite if the forecast density has infinite first moment, such as in the case of a Cauchy density. Even then, the mean scores (21) and (23) can be plotted versus the threshold z and the quantile α, and the resulting plots can be interpreted diagnostically. Furthermore, the threshold-weighted continuous ranked probability score (14) is finite if the weight function is integrable, and in this latter form the weighted CRPS test continues to apply.

Appendix: Moment conditions

We supply the remaining nontrivial arguments in Section 3.2. To verify the assumptions of Theorem 4 of Giacomini and White (2006), we need to show that the moment condition (19) implies

$$E\,|\mathrm{S}(\hat{f}_{t+k}, Y_{t+k}) - \mathrm{S}(\hat{g}_{t+k}, Y_{t+k})|^{2r} \tag{24}$$

to be finite, where S is the threshold-weighted continuous ranked probability score (14) or the quantile-weighted score (15), and the weight function is bounded. For ease of notation, we substitute f, g and Y for $\hat{f}_{t+k}$, $\hat{g}_{t+k}$ and Y_{t+k}, respectively. If the weight function is bounded above by the constant $M > 0$, then

$$E|\mathrm{S}(f,Y) - \mathrm{S}(g,Y)|^{2r} \leq (2M)^{2r}\,(E\,\mathrm{CRPS}(f,Y)^{2r} + E\,\mathrm{CRPS}(g,Y)^{2r}).$$

We proceed to show that under (19) both $E\,\mathrm{CRPS}(f,Y)^{2r}$ and $E\,\mathrm{CRPS}(g,Y)^{2r}$ are finite. If X and X' are independent random variables with density f that are independent of Y, then

$$\mathrm{CRPS}(f,Y) = E|X - Y| - \frac{1}{2}E_f|X - X'| \leq 2E_f|X| + |Y|$$

by the triangle inequality, and therefore

$$E\,\mathrm{CRPS}(f,Y)^{2r} \leq\; 2^{2r}\,((2E_f|X|)^{2r} + E|Y|^{2r}).$$

A similar result holds for $E\,\mathrm{CRPS}(g,Y)^{2r}$; hence, (19) is a sufficient condition for the expectation (24) to be finite.

Finally, if the threshold weight function u in (14) is integrable, the score differential in (24) is bounded and its moments of order $r \geq 2$ are finite.

Acknowledgements

This research was supported by the National Science Foundation under Awards ATM-0724721 and DMS-0706745.

References

Amisano, G., and R. Giacomini (2007): “Comparing Density Forecasts via Weighted Likelihood Ratio Tests,” *Journal of Business and Economic Statistics*, 25, 177–190.

Bao, Y., T.-H. Lee, and B. Saltoğlu (2007): “Comparing Density Forecast Models,” *Journal of Forecasting*, 26, 203–225.

Bollerslev, T. (1986): “Generalized Autoregressive Conditional Heteroscedasticity,” *Journal of Econometrics*, 31, 307–327.

Brown, B. G., R. W. Katz, and A. H. Murphy (1984): “Time Series Models to Simulate and Forecast Wind Speed and Wind Power,” *Journal of Climate and Applied Meteorology*, 23, 1184–1195.

Cervera, J. L, and J. Muñoz (1996): “Proper Scoring Rules for Fractiles,” in *Bayesian Statistics 5*, ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, 513–519.

Christoffersen, P. F., and F. X. Diebold (1996): “Further Results on Forecasting and Model Selection Under Asymmetric Loss,” *Journal of Applied Econometrics*, 11, 561–571.

Clements, M. P. (2004): "Evaluating the Bank of England Density Forecasts of Inflation," *Economic Journal*, 114, 844–866.

Corradi V., and N. R. Swanson (2006a): "Predictive Density and Conditional Confidence Interval Accuracy Tests," *Journal of Econometrics*, 135, 187–228.

——— (2006b): "Predictive Density Evaluation," in *Handbook of Economic Forecasting*, ed. by C. W. J. Granger, G. Elliott and A. Timmermann, Amsterdam, North Holland, 197–286.

Dawid, A. P. (1984): "Statistical Theory: The Prequential Approach," *Journal of the Royal Statistical Society Series A*, 147, 278–292.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253–263.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998): "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39, 863–883.

Elder, R., G. Kapetanios, T. Taylor and T. Yates (2005): "Assessing the MPC's Fan Charts," *Bank of England Quarterly Bulletin*, Autumn, 326–348.

Elliott, G., and A. Timmermann (2008): "Economic Forecasting," *Journal of Economic Literature*, 46, 1–53.

Engle, R. F. (1982): "Autoregressive Conditional Heteroscedasticity With Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 45, 987–1007.

Giacomini, R., and I. Komunjer (2005): "Evaluation and Combination of Conditional Quantile Forecasts," *Journal of Business and Economic Statistics*, 23, 416–431.

Giacomini, R., and H. White (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578.

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

Gneiting, T., and A. E. Raftery (2007): "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007): "Probabilistic Forecasts, Calibration and Sharpness," *Journal of the Royal Statistical Society Series B*, 67, 243–268.

Gneiting, T., K. Larson, K. Westrick, M. G. Genton, and E. Aldrich (2006): "Calibrated Probabilistic Forecasting at the Stateline Wind Energy Center: The Regime-Switching Space-Time Method," *Journal of the American Statistical Association*, 101, 968–979.

Koenker, R., and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Laio, F., and S. Tamea (2007): "Verification Tools for Probabilistic Forecasts of Continuous Hydrological Variables," *Hydrology and Earth System Sciences*, 11, 1267–1277.

Matheson, J. E., and R. L. Winkler (1976): "Scoring Rules for Continuous Probability

Distributions," *Management Science*, 22, 1087–1096.

Mitchell, J., and S. G. Hall (2005): "Evaluating, Comparing and Combining Density Forecasts Using the KLIC With an Application to the Bank of England and NIESR 'Fan' Charts of Inflation," *Oxford Bulletin of Economics and Statistics*, 67S, 995–1033.

Schumacher, M., E. Graf, and T. Gerds (2003): "How to Assess Prognostic Models for Survival Data: A Case Study in Oncology," *Methods of Information in Medicine*, 42, 564–571.

Selten, R. (1998): "Axiomatic Characterization of the Quadratic Scoring Rule," *Experimental Economics*, 1, 43–62.

Timmermann, A. (2000): "Density Forecasting in Economics and Finance," *Journal of Forecasting*, 19, 231–234.

Wallis, K. F. (2003): "Chi-Squared Tests of Interval and Density Forecasts, and the Bank of England's Fan Charts," *International Journal of Forecasting*, 19, 165–175.

——— (2004): "An Assessment of Bank of England and National Institute Inflation Forecast Uncertainties," *National Institute Economic Review*, 189, 64–71.

Winkler, R. L. (1996): "Scoring Rules and the Evaluation of Probabilities," *Test*, 5, 1–60.