



**HAL**  
open science

## Probabilistic forecasts, calibration and sharpness.

Tilmann Gneiting, Fadoua Balabdaoui, Adrian Raftery

► **To cite this version:**

Tilmann Gneiting, Fadoua Balabdaoui, Adrian Raftery. Probabilistic forecasts, calibration and sharpness.. 2007. hal-00363242

**HAL Id: hal-00363242**

**<https://hal.science/hal-00363242v1>**

Preprint submitted on 25 Feb 2009

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Probabilistic Forecasts, Calibration and Sharpness

JRSSB Submission B6257 Revision 1

Tilmann Gneiting<sup>1</sup>, Fadoua Balabdaoui<sup>2</sup> and Adrian E. Raftery<sup>1</sup>

tilmann@stat.washington.edu, fadoua@math.uni-goettingen.de, raftery@stat.washington.edu

Corresponding author: Tilmann Gneiting

<sup>1</sup>Department of Statistics, University of Washington  
Seattle, Washington 98195-4322, U.S.A.

<sup>2</sup>Institut für Mathematische Stochastik, Georg-August-Universität Göttingen  
37073 Göttingen, Germany

## Abstract

Probabilistic forecasts of a continuous variable take the form of predictive densities or predictive cumulative distribution functions. We propose a diagnostic approach to the evaluation of predictive performance that is based on the paradigm of *maximizing the sharpness of the predictive distributions subject to calibration*. Calibration refers to the statistical consistency between the distributional forecasts and the observations and is a joint property of the predictions and the events that materialize. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only. A simple theoretical framework phrased in terms of a game between nature and forecaster allows us to distinguish probabilistic calibration, exceedance calibration and marginal calibration. We propose and study tools for checking calibration and sharpness, among them the probability integral transform (PIT) histogram, marginal calibration plots, the sharpness diagram and proper scoring rules. The diagnostic approach is illustrated by an assessment and ranking of probabilistic forecasts of wind speed at the Stateline wind energy center in the U.S. Pacific Northwest. In combination with cross-validation or in the time series context, our proposal provides very general, nonparametric alternatives to the use of information criteria for model diagnostics and model selection.

*Keywords:* Cross-validation; Density forecast; Ensemble prediction system; Ex post evaluation; Forecast verification; Model diagnostics; Posterior predictive assessment; Predictive distribution; Prequential principle; Probability integral transform; Proper scoring rule

## 1 Introduction

A major human desire is to make forecasts for the future. Forecasts characterize and reduce but generally do not eliminate uncertainty. Consequently, forecasts should be probabilistic in nature, taking the form of probability distributions over future events (Dawid 1984). Indeed, over the past two decades the quest for good probabilistic forecasts has become a driving force in meteorology (Gneiting and Raftery 2005). Major economic forecasts such as the quarterly Bank of England inflation report are issued in terms of predictive distributions, and the rapidly growing area of financial risk management is dedicated to probabilistic forecasts of portfolio values (Duffie and Pan 1997; Granger 2006). In the statistical literature, advances in Markov chain Monte Carlo methodology (see, for example, Besag *et al.* 1995) have led to explosive growth in the use of predictive distributions, mostly in the form of Monte Carlo samples from the posterior predictive distribution of quantities of interest.

It is often critical to assess the predictive ability of forecasters, or to compare and rank competing forecasting methods. Atmospheric scientists talk of forecast verification when they refer to this process (Jolliffe and Stephenson 2003), and much of the underlying methodology has been developed by meteorologists. There is also a relevant strand of work in the econometrics literature (Diebold and Mariano 1995; Christoffersen 1998; Diebold *et al.* 1998). Murphy and Winkler (1987) proposed a general framework for the evaluation of point forecasts that uses a diagnostic approach based on graphical displays, summary measures and scoring rules. In this paper, we consider probabilistic forecasts (as opposed to point forecasts) of continuous and mixed discrete-continuous variables, such as temperature, wind speed, precipitation, gross domestic product, inflation rates and portfolio values. In this situation, probabilistic forecasts take the form of predictive densities or predictive cumulative distribution functions (CDFs), and the diagnostic approach faces a challenge, in that the forecasts take the form of probability distributions while the observations are real-valued.

We employ the following, simple theoretical framework to provide guidance in our methodological work. At times or instances  $t = 1, 2, \dots$ , nature chooses a distribution,  $G_t$ , which we think of as the true data generating process, and the forecaster picks a probabilistic forecast in the form of a predictive cumulative distribution function,  $F_t$ . The outcome,  $x_t$ , is a random number with distribution  $G_t$ . Throughout, we assume that nature is omniscient, in the sense that the forecaster's information basis is at most that of nature. Hence, if

$$F_t = G_t \quad \text{for all } t \tag{1}$$

we talk of the ideal forecaster. In practice, the true distribution,  $G_t$ , remains hypothetical, and the predictive distribution,  $F_t$ , is an expert opinion that may or may not derive from a statistical prediction algorithm. In accordance with Dawid's (1984) prequential principle, the predictive distributions need to be assessed on the basis of the forecast-observation pairs  $(F_t, x_t)$  only, irrespectively of their origins. Dawid (1984) and Diebold *et al.* (1998) proposed the use of the probability integral transform (PIT) value,

$$p_t = F_t(x_t), \tag{2}$$

for doing this. If the forecasts are ideal and  $F_t$  is continuous, then  $p_t$  has a uniform distribution. Hence, the uniformity of the probability integral transform is a necessary condition for the forecaster to be ideal, and checks for its uniformity have formed a cornerstone of forecast evaluation. In the classical time series framework, each  $F_t$  corresponds to a one-step ahead forecast, and checks for the uniformity of the PIT values have been supplemented by tests for independence (Frühwirth-Schnatter 1996; Diebold *et al.* 1998).

Hamill (2001) gave a thought-provoking example of a forecaster for whom the histogram of the PIT values is essentially uniform, even though every single probabilistic forecast is biased. His example aimed to show that the uniformity of the PIT values is a necessary but not a sufficient condition for the forecaster to be ideal. To fix the idea, we consider a simulation study based on the scenario described in Table 1. At times or instances  $t = 1, 2, \dots$ , nature draws a standard normal random number  $\mu_t$  and selects the data generating distribution,  $G_t = \mathcal{N}(\mu_t, 1)$ . In the context of weather forecasts, we might think of  $\mu_t$  as an accurate description of the latest observable state of the atmosphere, summarizing all information that a forecaster might possibly have access to. The ideal forecaster is an expert meteorologist who conditions on the current state,  $\mu_t$ , and issues an ideal probabilistic forecast,  $F_t = G_t$ . The climatological forecaster takes the unconditional distribution,  $F_t = \mathcal{N}(0, 2)$ , as probabilistic forecast. The unfocused forecaster observes the current state,  $\mu_t$ , but adds a mixture component to the forecast, which can be interpreted as distributional

Table 1: Scenario for the simulation study. At times  $t = 1, 2, \dots$ , nature chooses a distribution,  $G_t$ , and forecaster chooses a probabilistic forecast,  $F_t$ . The observations are independent random numbers  $x_t$  with distribution  $G_t$ . We write  $\mathcal{N}(\mu, \sigma^2)$  for the normal distribution with mean  $\mu$  and variance  $\sigma^2$ , respectively. The sequences  $(\mu_t)_{t=1,2,\dots}$ ,  $(\tau_t)_{t=1,2,\dots}$  and  $(\delta_t, \sigma_t^2)_{t=1,2,\dots}$  are independent identically distributed, and independent of each other and the observations.

Nature	$G_t = \mathcal{N}(\mu_t, 1)$ where $\mu_t \sim \mathcal{N}(0, 1)$
Ideal forecaster	$F_t = \mathcal{N}(\mu_t, 1)$
Climatological forecaster	$F_t = \mathcal{N}(0, 2)$
Unfocused forecaster	$F_t = \frac{1}{2} (\mathcal{N}(\mu_t, 1) + \mathcal{N}(\mu_t + \tau_t, 1))$ where $\tau_t = \pm 1$ with probability $\frac{1}{2}$ each
Hamill's forecaster	$F_t = \mathcal{N}(\mu_t + \delta_t, \sigma_t^2)$ where $(\delta_t, \sigma_t^2) = (\frac{1}{2}, 1), (-\frac{1}{2}, 1)$ or $(0, \frac{169}{100})$ with probability $\frac{1}{3}$ each

bias. A similar comment applies to Hamill's forecaster. Clearly, our forecasters are caricatures; yet, climatological reference forecasts and conditional biases are frequently observed in practice. The observation,  $x_t$ , is a random draw from  $G_t$ , and we repeat the prediction experiment 10000 times. Figure 1 shows that the PIT histograms for the four forecasters are essentially uniform.

In view of the reliance on the probability integral transform in the literature, this is a disconcerting result. As Diebold *et al.* (1998) pointed out, the ideal forecaster is preferred by all users, regardless of the respective loss function. Nevertheless, the probability integral transform is unable to distinguish between the ideal forecaster and her competitors. To address these limitations, we propose a diagnostic approach to the evaluation of predictive performance that is based on the paradigm of *maximizing the sharpness of the predictive distributions subject to calibration*. Calibration refers to the statistical consistency between the distributional forecasts and the observations, and is a joint property of the predictions and the observed values. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only. The more concentrated the predictive distributions, the sharper the forecasts, and the sharper the better, subject to calibration.

The remainder of the paper is organized as follows. Section 2 develops our theoretical framework for the assessment of predictive performance. We introduce the notions of probabilistic, exceedance and marginal calibration, give examples and counterexamples, and discuss a conjectured sharpness principle. In Section 3, we propose diagnostic tools such as marginal calibration plots and sharpness diagrams that complement the PIT histogram. Proper scoring rules address calibration as well as sharpness and allow to rank competing forecast procedures. Section 4 turns to a case study on probabilistic forecasts at the Stateline wind energy center in the U.S. Pacific Northwest. The diagnostic approach yields a clear-cut ranking of statistical algorithms for forecasts of wind speed, and suggests improvements that can be addressed in future research. Similar approaches hold considerable promise as very general, nonparametric tools for statistical model selection and model

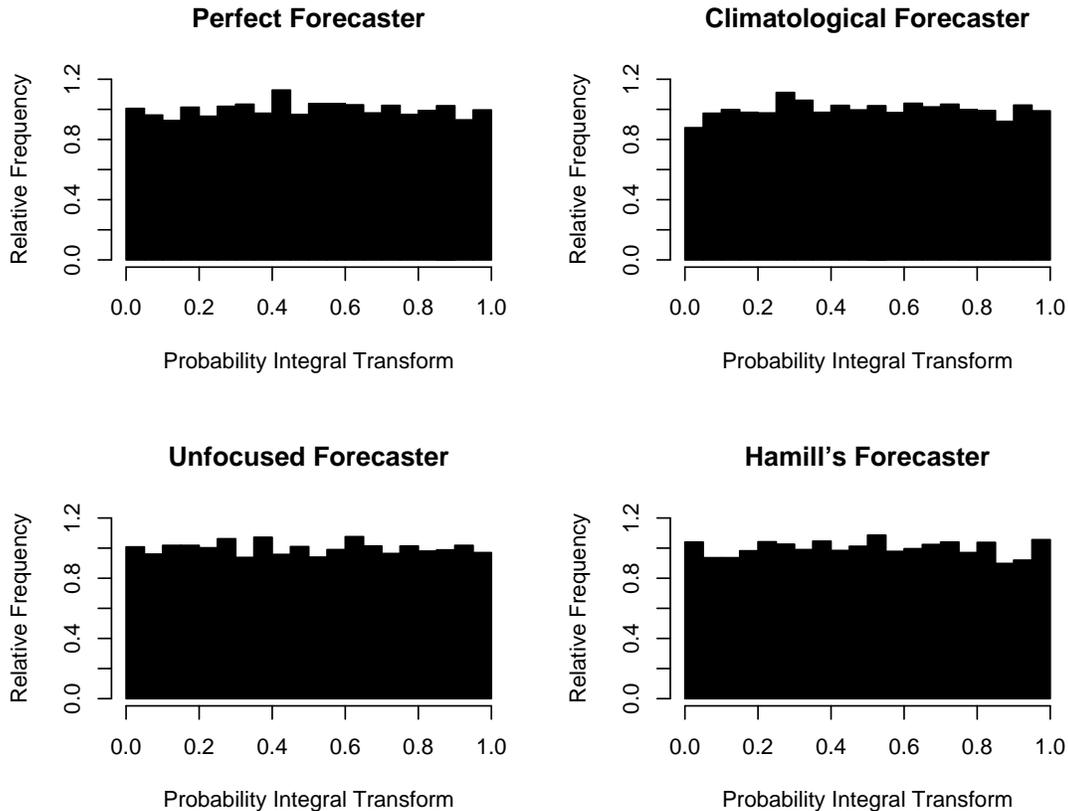


Figure 1: Probability integral transform (PIT) histograms.

diagnostics. The paper closes with a discussion in Section 5 that emphasizes the need for routine assessments of sharpness in the evaluation of predictive performance.

## 2 Modes of calibration

In our theoretical framework, we consider probabilistic forecasting as a game played between nature and forecaster. At times or instances  $t = 1, 2, \dots$ , nature picks a probability distribution,  $G_t$ , and forecaster chooses a probabilistic forecast in the form of a predictive distribution,  $F_t$ . The observation,  $x_t$ , is a random draw from nature's proposal distribution,  $G_t$ . Throughout, we assume that nature is omniscient, in the sense that the information basis of the forecaster is at most that of nature. For simplicity, we assume that  $F_t$  and  $G_t$  are continuous and strictly increasing on  $\mathbb{R}$ .

Some critical comments are in order. Evidently,  $G_t$  is not observed in practice and any operational evaluation must be performed on the basis of the forecasts,  $F_t$ , and the outcomes,  $x_t$ , only. The game theoretic framework of Shafer and Vovk (2001) and Vovk and Shafer (2005) also views forecasting as a game, with three players: forecaster, sceptic, and reality or nature. Forecaster and sceptic have opposite goals, and one of them wins, while the other loses. There is no goal assigned to nature, who directly chooses and reveals the outcome,  $x_t$ , without recourse to any underlying data generating distribution. The key question in this deep strand of literature, which culminates

in Theorem 3 of Vovk and Shafer (2005), is that for the existence of certain types of strategies for forecaster. Shafer and Vovk consider probability forecasts for dichotomous events, rather than distributional forecasts of real-valued quantities, and they do not consider the problem tackled here, namely the comparative evaluation of competing forecasters, for which they hint at future work (Shafer and Vovk 2001, p. 50).

In comparing forecasters, we take the pragmatic standpoint of a user, who is to rank and choose among a number of competitors, as exemplified in Section 4. In this type of situation, it is absolute performance that matters, rather than relative performance that may result from the use of possibly distinct information bases.

Our approach seems slightly broader than Dawid's (1984) prequential framework, in that we think of  $(F_t)_{t=1,2,\dots}$  as a general countable sequence of forecasts, with the index referring to time, space or subjects, depending on the prediction problem at hand. The forecasts need not be sequential, and when  $F_{t+1}$  is issued,  $x_t$  may or may not be available yet.

## 2.1 Probabilistic calibration, exceedance calibration and marginal calibration

Henceforth,  $(F_t)_{t=1,2,\dots}$  and  $(G_t)_{t=1,2,\dots}$  denote sequences of continuous and strictly increasing CDFs, possibly depending on stochastic parameters. We think of  $(G_t)_{t=1,2,\dots}$  as the true data generating process and of  $(F_t)_{t=1,2,\dots}$  as the associated sequence of probabilistic forecasts. The following definition refers to the asymptotic compatibility between the data generating process and the predictive distributions in terms of three major modes of calibration. Given that  $(F_t)_{t=1,2,\dots}$  and  $(G_t)_{t=1,2,\dots}$  might depend on stochastic parameters, convergence is understood as almost sure convergence and is denoted by an arrow. For now, these notions are of theoretical interest only; in Section 3 below, they lend support to our methodological proposals.

### Definition 1 (modes of calibration)

- (a) The sequence  $(F_t)_{t=1,2,\dots}$  is *probabilistically calibrated* relative to the sequence  $(G_t)_{t=1,2,\dots}$  if

$$\frac{1}{T} \sum_{t=1}^T G_t \circ F_t^{-1}(p) \longrightarrow p \quad \text{for all } p \in (0, 1). \quad (3)$$

- (b) The sequence  $(F_t)_{t=1,2,\dots}$  is *exceedance calibrated* relative to  $(G_t)_{t=1,2,\dots}$  if

$$\frac{1}{T} \sum_{t=1}^T G_t^{-1} \circ F_t(x) \longrightarrow x \quad \text{for all } x \in \mathbb{R}. \quad (4)$$

- (c) The sequence  $(F_t)_{t=1,2,\dots}$  is *marginally calibrated* relative to  $(G_t)_{t=1,2,\dots}$  if the limits  $\bar{G}(x) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T G_t(x)$  and  $\bar{F}(x) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T F_t(x)$  exist and equal each other for all  $x \in \mathbb{R}$ , and if the common limit distribution places all mass on finite values.

- (d) The sequence  $(F_t)_{t=1,2,\dots}$  is *strongly calibrated* relative to  $(G_t)_{t=1,2,\dots}$  if it is probabilistically calibrated, exceedance calibrated and marginally calibrated.

If each subsequence of  $(F_t)_{t=1,2,\dots}$  is probabilistically calibrated relative to the associated subsequence of  $(G_t)_{t=1,2,\dots}$ , we talk of complete probabilistic calibration. Similarly, we define completeness for exceedance, marginal and strong calibration. Probabilistic calibration is essentially

equivalent to the uniformity of the PIT values. Exceedance calibration is defined in terms of thresholds, and marginal calibration requires that the limit distributions  $\bar{G}$  and  $\bar{F}$  exist and equal each other. The existence of  $\bar{G}$  is a natural assumption in meteorological problems and corresponds to the existence of a stable climate. Hence, marginal calibration can be interpreted in terms of the equality of observed and forecast climatology.

Various authors have studied calibration in the context of probability forecasts for sequences of binary events (De Groot and Fienberg 1982; Dawid 1982, 1985a, 1985b; Oakes 1985; Schervish 1985, 1989; Dawid and Vovk 1999; Shafer and Vovk 2001; Sandroni *et al.* 2003). The progress is impressive and culminates in the elegant game theoretic approach of Vovk and Shafer (2005). Krzysztofowicz (1999) discussed calibration in the context of Bayesian forecasting systems, and Krzysztofowicz and Sigrest (1999) studied calibration for quantile forecasts of quantitative precipitation. We are unaware of any prior discussion of notions of calibration for probabilistic forecasts of continuous variables.

## 2.2 Examples

The examples in this section illustrate the aforementioned modes of calibration and discuss some of the forecasters in our initial simulation study. Throughout,  $(\mu_t)_{t=1,2,\dots}$ ,  $(\sigma_t)_{t=1,2,\dots}$  and  $(\tau_t)_{t=1,2,\dots}$  denote independent sequences of independent identically distributed random variables. We write  $\mathcal{N}(\mu, \sigma^2)$  for the normal distribution with mean  $\mu$  and variance  $\sigma^2$ , identify distributions and CDFs, and let  $\Phi$  denote the standard normal cumulative. In each example, nature draws a standard normal random number,  $\mu_t$ , that corresponds to the information basis at time  $t$  and picks the data generating distribution  $G_t = \mathcal{N}(\mu_t, 1)$ .

**Example 1 (ideal forecaster)** The predictive distribution of the ideal forecaster equals nature's proposal distribution, that is,  $F_t = G_t = \mathcal{N}(\mu_t, 1)$  for all  $t$ . This forecaster is strongly calibrated.

**Example 2 (climatological forecaster)** The climatological forecaster issues the distributional forecast  $F_t = \mathcal{N}(0, 2)$ , irrespectively of  $t$ . This forecaster is probabilistically calibrated and marginally calibrated. However,

$$\frac{1}{T} \sum_{t=1}^T G_t^{-1} \circ F_t(x) = \frac{1}{T} \sum_{t=1}^T \left[ \Phi^{-1} \left( \Phi \left( \frac{x}{\sqrt{2}} \right) \right) + \mu_t \right] \longrightarrow \frac{x}{\sqrt{2}}$$

for  $x \in \mathbb{R}$ , in violation of exceedance calibration.

The characteristic property in Example 2 is that each predictive distribution,  $F_t$ , equals nature's limiting distribution,  $\bar{G}$ . We call any forecaster with this property a climatological forecaster. For climatological forecasts, probabilistic calibration is essentially equivalent to marginal calibration. Indeed, if  $\bar{G}$  is continuous and strictly increasing, then putting  $p = F_t(x) = \bar{G}(x)$  in (3) recovers the marginal calibration condition. In practice, climatological forecasts are constructed from historical records of observations, and they are often used as reference forecasts.

**Example 3 (unfocused forecaster)** The predictive distribution of the unfocused forecaster is the mixture distribution

$$F_t = \frac{1}{2} (\mathcal{N}(\mu_t, 1) + \mathcal{N}(\mu_t + \tau_t, 1)),$$

Table 2: The three major modes of calibration are logically independent of each other and may occur in any combination. For instance, the unfocused forecaster in Example 3 is probabilistically calibrated (P), but neither exceedance calibrated ( $\bar{E}$ ) nor marginally calibrated ( $\bar{P}$ ).

Properties	Example
PEM	Example 1 (ideal forecaster)
PE $\bar{M}$	$G_t = F_t = \mathcal{N}(t, 1)$
P $\bar{E}M$	Example 2 (climatological forecaster)
P $\bar{E}\bar{M}$	Example 3 (unfocused forecaster)
$\bar{P}EM$	Example 4 (mean-biased forecaster)
$\bar{P}E\bar{M}$	Example 5 (sign-biased forecaster)
$\bar{P}EM$	Example 6 (mixed forecaster)
$\bar{P}\bar{E}\bar{M}$	$G_t = \mathcal{N}(0, 1), F_t = \mathcal{N}(1, 1)$

where  $\tau_t$  is either 1 or  $-1$ , with equal probabilities, and independently of  $\mu_t$ . This forecaster is probabilistically calibrated, but neither exceedance calibrated nor marginally calibrated. To prove the claim for probabilistic calibration, put  $\Phi_{\pm}(x) = \frac{1}{2}(\Phi(x) + \Phi(x \mp 1))$  and note that

$$\frac{1}{T} \sum_{t=1}^T G_t \circ F_t^{-1}(p) \longrightarrow \frac{1}{2} \left[ \Phi \circ \Phi_+^{-1}(p) + \Phi \circ \Phi_-^{-1}(p) \right] = p,$$

where the equality follows upon putting  $p = \Phi_+(x)$ , substituting and simplifying. Exceedance calibration does not hold, because

$$\frac{1}{T} \sum_{t=1}^T G_t^{-1} \circ F_t(x) \longrightarrow \frac{1}{2} \left[ \Phi^{-1} \circ \Phi_+(x) + \Phi^{-1} \circ \Phi_-(x) \right] \neq x$$

in general. The marginal calibration condition is violated, because nature's limit distribution,  $\bar{G} = \mathcal{N}(0, 2)$ , does not equal  $\bar{F} = \frac{1}{2}\mathcal{N}(0, 2) + \frac{1}{4}\mathcal{N}(-1, 2) + \frac{1}{4}\mathcal{N}(1, 2)$ .

**Example 4 (mean-biased forecaster)** This forecaster issues the probabilistic forecast  $F_t = \mathcal{N}(\mu_t + \tau_t, 1)$ , where, again,  $\tau_t$  is either 1 or  $-1$ , with equal probabilities, and independently of  $\mu_t$ . The mean-biased forecaster is exceedance calibrated but neither probabilistically calibrated nor marginally calibrated.

**Example 5 (sign-biased forecaster)** The predictive distribution of the sign-biased forecaster is  $F_t = \mathcal{N}(-\mu_t, 1)$ . This forecaster is exceedance calibrated and marginally calibrated, but not probabilistically calibrated.

**Example 6 (mixed forecaster)** The mixed forecaster randomizes between the climatological and the sign-biased forecast, with equal probabilities and independently of  $\mu_t$ . This forecaster is marginally calibrated, but neither probabilistically calibrated nor exceedance calibrated.

The examples in this section show that probabilistic calibration, exceedance calibration and marginal calibration are logically independent of each other and may occur in any combination. Table 2 summarizes the respective results.

### 2.3 Hamill's forecaster

We add a discussion of Hamill's forecaster. As previously, nature picks  $G_t = \mathcal{N}(\mu_t, 1)$ , where  $\mu_t$  is standard normal. Hamill's forecaster is a master forecaster who assigns the prediction task with equal probability to any of three student forecasters, each of whom is biased, as described in Table 1. For Hamill's forecaster,

$$\frac{1}{T} \sum_{t=1}^T G_t \circ F_t^{-1}(p) \longrightarrow \frac{1}{3} \left[ \Phi \left( \Phi^{-1}(p) - \frac{1}{2} \right) + \Phi \left( \frac{13}{10} \Phi^{-1}(p) \right) + \Phi \left( \Phi^{-1}(p) + \frac{1}{2} \right) \right] = p + \epsilon(p),$$

where  $|\epsilon(p)| \leq 0.0032$  for all  $p$  but  $\epsilon(p) \neq 0$  in general. The probabilistic calibration condition (3) is violated, but only slightly so, resulting in deceptively uniform PIT histograms. As for exceedance calibration, we note that

$$\frac{1}{T} \sum_{t=1}^T G_t^{-1} \circ F_t(p) \longrightarrow \frac{1}{3} \left[ \left( x + \frac{1}{2} \right) + \frac{10}{13} x + \left( x - \frac{1}{2} \right) \right] = \frac{12}{13} x$$

for  $x \in \mathbb{R}$ . Hence, Hamill's forecaster is not exceedance calibrated either, nor marginally calibrated, given that  $\bar{G} = \mathcal{N}(0, 2)$  while  $\bar{F} = \frac{1}{3}(\mathcal{N}(-\frac{1}{2}, 2) + \mathcal{N}(\frac{1}{2}, 2) + \mathcal{N}(0, \frac{269}{100}))$ .

### 2.4 Sharpness principle

In view of our assumption that forecaster's information basis is at most that of nature, the best situation that we can possibly hope for is the equality (1) of  $F_t$  and  $G_t$  that characterizes the ideal forecaster. Operationally, we adopt the paradigm of maximizing the sharpness of the predictive distributions subject to calibration. Our conjectured sharpness principle contends that the two goals — ideal forecasts and the maximization of sharpness subject to calibration — are equivalent. This conjectured equivalence, which we deliberately state loosely, could be explained in two ways. One explanation is that sufficiently stark notions of calibration, such as complete strong calibration across many dynamic subsequences, imply asymptotic equivalence to the ideal forecaster. Strong calibration alone, without the completeness condition, does not seem to impose enough restrictions, but we are unaware of a counterexample and would like to know of one. An alternative and weaker explanation states that any sufficiently calibrated forecaster is at least as spread out as the ideal forecaster.

With respect to this latter explanation, none of probabilistic, exceedance or marginal calibration alone is sufficiently stark. In the examples below it will be convenient to consider a probabilistic calibration condition,

$$\frac{1}{T} \sum_{t=1}^T G_t \circ F_t^{-1}(p) = p \quad \text{for all } p \in (0, 1), \tag{5}$$

for finite sequences  $(F_t)_{1 \leq t \leq T}$  relative to  $(G_t)_{1 \leq t \leq T}$ , and similarly for exceedance calibration and marginal calibration. The examples extend to countable sequences in obvious ways. Now suppose that  $\sigma > 0$ ,  $a > 1$ ,  $0 < \lambda < 1/a$  and  $T = 2$ . Let  $G_1$  and  $G_2$  be continuous and strictly increasing distributions functions with associated densities that are symmetric about zero and have finite variance,  $\text{var}(G_1) = \sigma^2$  and  $\text{var}(G_2) = \lambda\sigma^2$ . If we define

$$F_1(x) = \frac{1}{2} \left( G_1(x) + G_2 \left( \frac{x}{a} \right) \right), \quad F_2(x) = F_1(ax),$$

then

$$\text{var}(F_1) + \text{var}(F_2) = \frac{1}{2} \left(1 + \frac{1}{a^2}\right) (1 + a^2 \lambda^2) \sigma^2 < (1 + \lambda^2) \sigma^2 = \text{var}(G_1) + \text{var}(G_2),$$

even though the finite probabilistic calibration condition (5) holds. A similar example can be given for exceedance calibration. Suppose that  $\sigma > 0$ ,  $0 < a < 1$  and

$$0 < \lambda < a \left( \frac{3+a}{1+3a} \right)^{1/2}.$$

Let  $G_1$  and  $G_2$  be as above and define

$$F_1(x) = G_1\left(\frac{2x}{1+a}\right), \quad F_2(x) = G_2\left(\frac{2ax}{1+a}\right).$$

Then

$$\text{var}(F_1) + \text{var}(F_2) = \frac{1}{4} (1+a)^2 \left(1 + \frac{\lambda^2}{a^2}\right) \sigma^2 < (1 + \lambda^2) \sigma^2 = \text{var}(G_1) + \text{var}(G_2),$$

even though the finite exceedance calibration condition holds. Evidently, a forecaster can be marginally calibrated yet sharper than the ideal forecaster.

For climatological forecasts, however, finite probabilistic calibration and finite marginal calibration are equivalent, and a weak form of the sharpness principle holds.

**Theorem 1** Suppose that  $G_1, \dots, G_T$  and  $F_1 = \dots = F_T = F$  have second moments and satisfy the finite probabilistic calibration condition (5). Then

$$\frac{1}{T} \sum_{t=1}^T \text{var}(F_t) = \text{var}(F) \geq \frac{1}{T} \sum_{t=1}^T \text{var}(G_t)$$

with equality if and only if  $E(G_1) = \dots = E(G_T)$ .

The proof of Theorem 1 is given in the appendix. We are unaware of any other results in this direction; in particular, we do not know whether a non-climatological forecaster can be probabilistically calibrated and marginally calibrated yet sharper than the ideal forecaster.

### 3 Diagnostic tools

We now discuss diagnostic tools for the evaluation of predictive performance. In accordance with Dawid's (1984) prequential principle, the assessment of probabilistic forecasts needs to be based on the predictive distributions and the observations only. Previously, we defined notions of calibration in terms of the asymptotic consistency between the probabilistic forecasts and the data generating distributions, which are unavailable in practice. Hence, we turn to sample versions, by substituting empirical distribution functions based on the outcomes, resulting in methodological tools that stand in their own right as well. In the following, this program is carried out for probabilistic and marginal calibration. Exceedance calibration does not allow for an obvious sample analogue, and it is not clear whether such exists. We discuss graphical displays of sharpness and propose the use of proper scoring rules, that assign numerical measures of predictive performance and find key applications in the ranking of competing forecast procedures.

### 3.1 Assessing probabilistic calibration

The probability integral transform (PIT) is the value that the predictive CDF attains at the observation. Specifically, if  $F_t$  is the predictive distribution and  $x_t$  materializes, the transform is defined as  $p_t = F_t(x_t)$ . The literature usually refers to Rosenblatt (1952), although the probability integral transform can be traced back at least to Pearson (1933). The connection to probabilistic calibration is established by substituting the empirical distribution function  $\mathbf{1}\{x_t \leq x\}$  for the data generating distribution  $G_t(x)$ ,  $x \in \mathbb{R}$  in the probabilistic calibration condition (3), and noting that  $x_t \leq F_t^{-1}(p)$  if and only if  $p_t \leq p$ . The following theorem characterizes the asymptotic uniformity of the empirical sequence of PIT values in terms of probabilistic calibration. We state this result under the assumption of a \*-mixing sequence of observations (Blum *et al.* 1963). The proof is deferred to the appendix.

**Theorem 2** Let  $(F_t)_{t=1,2,\dots}$  and  $(G_t)_{t=1,2,\dots}$  be sequences of continuous, strictly increasing distribution functions. Suppose that  $x_t$  has distribution  $G_t$  and that the  $x_t$  form a \*-mixing sequence of random variables. Then

$$\frac{1}{T} \sum_{t=1}^T \mathbf{1}\{p_t < p\} \longrightarrow p \quad \text{almost surely for all } p \quad (6)$$

if and only if  $(F_t)_{t=1,2,\dots}$  is probabilistically calibrated with respect to  $(G_t)_{t=1,2,\dots}$ .

We emphasize that condition (6) stands in its own right as a criterion for the validity of probabilistic forecasts, independently of our theoretical framework. Indeed, following the lead of Dawid (1984) and Diebold *et al.* (1998), checks for the uniformity of the PIT values have formed a cornerstone of forecast evaluation.

Uniformity is usually assessed in an exploratory sense, and one way of doing this is by plotting the empirical CDF of the PIT values and comparing to the identity function. This approach is adequate for small sample sizes and notable departures from uniformity, and its proponents include Staël von Holstein (1970, p. 142), Seillier-Moiseiwitsch (1993), Hoeting (1994, p. 33), Frühwirth-Schnatter (1996), Raftery, Madigan and Hoeting (1997), Clements and Smith (2000), Moyeed and Papritz (2002), Wallis (2003) and Boero and Marrocu (2004). Histograms of the PIT values accentuate departures from uniformity when the sample size is large and the deviations from uniformity are small. This alternative type of display has been used by Diebold *et al.* (1998), Weigend and Shi (2000), Bauwens *et al.* (2004) and Gneiting *et al.* (2005), among others, and 10 or 20 histogram bins generally seem adequate. Figure 1 employs 20 bins and shows the PIT histograms for the various forecasters in our initial simulation study. The histograms are essentially uniform. Table 3 shows the empirical coverage of the associated central 50% and 90% prediction intervals. This information is redundant, since the empirical coverage can be read off the PIT histogram, as the area under the 10 and 18 central bins, respectively.

Probabilistic weather forecasts are typically based on ensemble prediction systems, which generate a set of perturbations of the best estimate of the current state of the atmosphere, run each of them forward in time using a numerical weather prediction model, and use the resulting set of forecasts as a sample from the predictive distribution of future weather quantities (Palmer 2002; Gneiting and Raftery 2005). The principal device for assessing the calibration of ensemble forecasts is the verification rank histogram or Talagrand diagram, proposed independently by Anderson (1996), Hamill and Colucci (1997) and Talagrand *et al.* (1997), and extensively used since. To obtain a verification rank histogram, find the rank of the observation when pooled within the ordered

Table 3: Empirical coverage of central prediction intervals. The nominal coverage is 50% and 90%, respectively.

Interval	50%	90%
Perfect forecaster	51.2%	90.0%
Climatological forecaster	51.3%	90.7%
Unfocused forecaster	50.1%	90.1%
Hamill’s forecaster	50.9%	89.5%

ensemble values and plot the histogram of the ranks. If we identify the predictive distribution with the empirical CDF of the ensemble values, this technique is seen to be equivalent to plotting a PIT histogram. A similar procedure could be drawn on fruitfully to assess samples from posterior predictive distributions obtained by Markov chain Monte Carlo techniques. Shephard (1994, p. 129) gave an instructive example of how this could be done.

Visual inspection of a PIT or rank histogram can provide hints to the reasons for forecast deficiency. Hump shaped histograms indicate overdispersed predictive distributions with prediction intervals that are too wide on average. U-shaped histograms often correspond to predictive distributions that are too narrow. Triangle-shaped histograms are seen when the predictive distributions are biased. Formal tests of uniformity can be employed and have been studied by Anderson (1996), Talagrand *et al.* (1997), Noceti *et al.* (2003), Garratt *et al.* (2003), Wallis (2003) and Candille and Talagrand (2005), among others. However, the use of formal tests is often hindered by complex dependence structures, particularly in cases in which the PIT values are spatially aggregated. Hamill (2001) gave a thoughtful discussion of the associated issues and potential fallacies.

In the time series context, the observations are sequential, and the predictive distributions correspond to sequential  $k$ -step ahead forecasts. The probability integral transforms for ideal  $k$ -step ahead forecasts are at most  $(k - 1)$ -dependent, and this assumption can be checked empirically, by plotting the sample autocorrelation function for the PIT values and their moments (Diebold *et al.* 1998). Smith (1985), Frühwirth-Schnatter (1996) and Berkowitz (2001) proposed an assessment of independence based on the transformed PIT values,  $\Phi^{-1}(p_t)$ , which are Gaussian under the assumption of ideal forecasts. This further transformation has obvious advantages when formal tests of independence are employed, and seems to make little difference otherwise.

### 3.2 Assessing marginal calibration

Marginal calibration concerns the equality of forecast climate and actual climate. To assess marginal calibration, we propose a comparison of the average predictive CDF,

$$\bar{F}_T(x) = \frac{1}{T} \sum_{t=1}^T F_t(x), \quad x \in \mathbb{R}, \quad (7)$$

to the empirical CDF of the observations,

$$\hat{G}_T(x) = \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{x_t \leq x\}, \quad x \in \mathbb{R}. \quad (8)$$

Indeed, if we substitute the indicator function  $\mathbf{1}\{x_t \leq x\}$  for the data generating distribution  $G_t(x)$ ,  $x \in \mathbb{R}$ , in the definition of marginal calibration, we are led to the asymptotic equality of  $\bar{F}_T$  and  $\hat{G}_T$ .

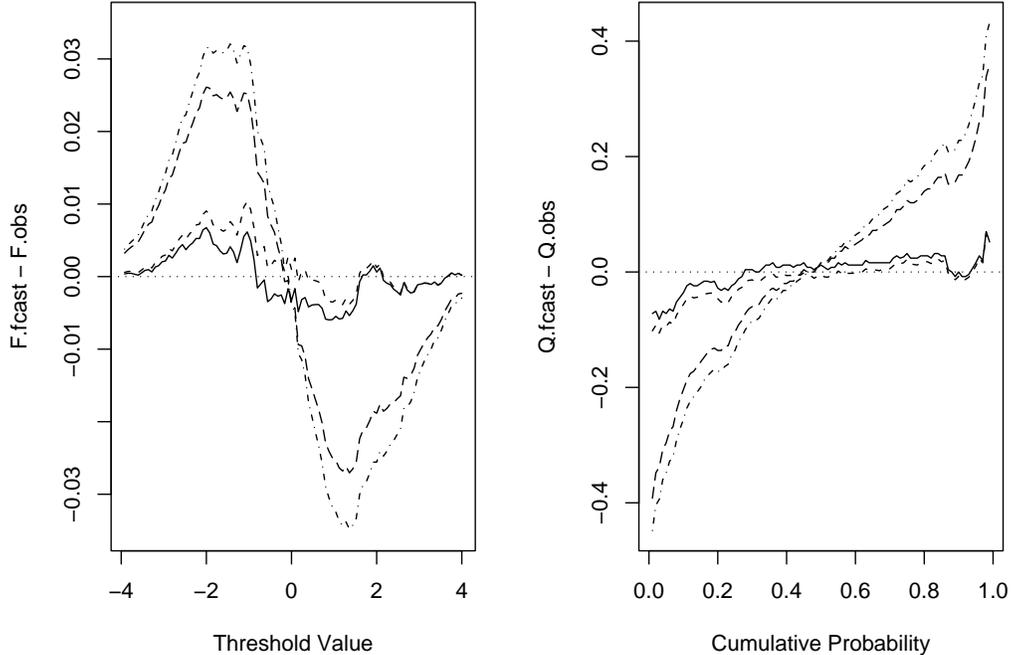


Figure 2: Marginal calibration plot for the ideal forecaster (solid line), climatological forecaster (short dashes), unfocused forecaster (dot-dashed line) and Hamill’s forecaster (long dashes). The presentation is in terms of CDFs (left) and in terms of quantiles (right), respectively.

Theorem 3 provides a rigorous version of this correspondence. Under mild regularity conditions, marginal calibration is a necessary and sufficient condition for the asymptotic equality of  $\hat{G}_T$  and  $\bar{F}_T$ . The proof of this result is deferred to the appendix.

**Theorem 3** Let  $(F_t)_{t=1,2,\dots}$  and  $(G_t)_{t=1,2,\dots}$  be sequences of continuous, strictly increasing distribution functions. Suppose that each  $x_t$  has distribution  $G_t$  and that the  $x_t$  form a  $*$ -mixing sequence of random variables. Suppose furthermore that  $\bar{F}(x) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T F_t(x)$  exists for all  $x \in \mathbb{R}$  and that the limit function is strictly increasing on  $\mathbb{R}$ . Then

$$\hat{G}_T(x) = \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{x_t \leq x\} \longrightarrow \bar{F}(x) \quad \text{almost surely for all } x \in \mathbb{R} \quad (9)$$

if and only if  $(F_t)_{t=1,2,\dots}$  is marginally calibrated with respect to  $(G_t)_{t=1,2,\dots}$ .

We note that (9) stands in its own right as a criterion for the validity of probabilistic forecasts. Still, Theorems 2 and 3 provide reassurance, in that (6) and (9) will be satisfied almost surely if the forecaster issues the same sequence of distributions that nature uses to generate the outcomes, assuming mixing conditions. These results are also of interest because they characterize situations under which (6) and (9) lead us to accept as valid forecasts that might in fact be far from ideal.

The most obvious graphical device in the assessment of marginal calibration is a plot of  $\hat{G}_T(x)$  and  $\bar{F}_T(x)$  versus  $x$ . However, it is often more instructive to plot the difference of the two CDFs, as on the left-hand side of Figure 2, which shows the difference

$$\bar{F}_T(x) - \hat{G}_T(x), \quad x \in \mathbb{R}, \quad (10)$$

Table 4: Average width of central prediction intervals. The nominal coverage is 50% and 90%, respectively.

Interval	50%	90%
Ideal forecaster	1.35	3.29
Climatological forecaster	1.91	4.65
Unfocused forecaster	1.52	3.68
Hamill's forecaster	1.49	3.62

for the various forecasters in our initial simulation study. We call this type of display a marginal calibration plot. Under the hypothesis of marginal calibration, we expect minor fluctuations about zero only, and this is indeed the case for the ideal forecaster and the climatological forecaster. The unfocused forecaster and Hamill's forecaster lack marginal calibration, resulting in major excursions from zero. The same information can be visualized in terms of quantiles, as on the right-hand side of Figure 2, which shows the difference

$$Q(\bar{F}_T, q) - Q(\hat{G}_T, q), \quad q \in (0, 1), \quad (11)$$

of the quantile functions for  $\bar{F}_T$  and  $\hat{G}_T$ , respectively. Under the hypothesis of marginal calibration, we again expect minor fluctuations about zero only, and this is the case for the ideal forecaster and the climatological forecaster. The unfocused forecaster and Hamill's forecaster show quantile difference functions that increase from negative to positive values, indicating forecast climates that are too spread out.

### 3.3 Assessing sharpness

Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only. The more concentrated the predictive distributions, the sharper the forecasts, and the sharper the better, subject to calibration. To assess sharpness, we use numerical and graphical summaries of the width of prediction intervals. For instance, Table 4 shows the average width of the central 50% and 90% prediction intervals for the forecasters in our simulation study. The ideal forecaster is the sharpest, followed by Hamill's, the unfocused and the climatological forecaster. In our simplistic simulation study, the width of the prediction intervals is fixed, except for Hamill's forecaster, and the tabulation is perfectly adequate. In real-world applications, conditional heteroscedasticity often leads to considerable variability in the width of the prediction intervals. The average width then is insufficient to characterize sharpness, and we follow Bremnes (2004) in proposing boxplots as a more instructive graphical device. We refer to this type of display as a sharpness diagram, and an example thereof is shown in Figure 9 below.

### 3.4 Proper scoring rules

Scoring rules assign numerical scores to probabilistic forecasts and form attractive summary measures of predictive performance, in that they address calibration and sharpness simultaneously. We write  $s(F, x)$  for the score assigned when the forecaster issues the predictive distribution  $F$  and  $x$  materializes, and we take scores to be penalties that the forecaster wishes to minimize. A scoring

Table 5: Average logarithmic score (LogS) and continuous ranked probability score (CRPS).

	LogS	CRPS
Ideal forecaster	1.41	0.56
Climatological forecaster	1.75	0.78
Unfocused forecaster	1.53	0.63
Hamill's forecaster	1.52	0.61

rule is proper if the expected value of the penalty  $s(F, x)$  for an observation  $x$  drawn from  $G$  is minimized if  $F = G$ . It is strictly proper if the minimum is unique. Winkler (1977) gave an interesting discussion of the ways in which proper scoring rules encourage honest and sharp forecasts.

The logarithmic score is the negative of the logarithm of the predictive density evaluated at the observation (Good 1952; Bernardo 1979). This scoring rule is proper and has many desirable properties (Roulston and Smith 2002), but lacks robustness (Selten 1998; Gneiting and Raftery 2004). The continuous ranked probability score is defined directly in terms of the predictive cumulative distribution function,  $F$ , namely as

$$\text{crps}(F, x) = \int_{-\infty}^{\infty} (F(y) - \mathbf{1}(y \geq x))^2 dy, \quad (12)$$

and provides a more robust alternative. Gneiting and Raftery (2004) gave an alternative representation and showed that

$$\text{crps}(F, x) = \mathbb{E}_F |X - x| - \frac{1}{2} \mathbb{E}_F |X - X'|, \quad (13)$$

where  $X$  and  $X'$  are independent copies of a random variable with cumulative distribution function  $F$  and finite first moment. The representation (13) shows that the continuous ranked probability score generalizes the absolute error, to which it reduces if  $F$  is a point forecast. It is reported in the same unit as the observations. The continuous ranked probability score is proper, and we rank competing forecast procedures based on its average,

$$\text{CRPS} = \frac{1}{T} \sum_{t=1}^T \text{crps}(F_t, x_t) = \int_{-\infty}^{\infty} \text{BS}(y) dy, \quad (14)$$

where  $\text{BS}(y) = \frac{1}{T} \sum_{t=1}^T (F_t(y) - \mathbf{1}\{x_t \leq y\})^2$  denotes the Brier score (Brier 1950) for probability forecasts of the binary event at the threshold value  $y \in \mathbb{R}$ . Like all proper scoring rules for binary probability forecasts, the Brier score allows for the distinction of a calibration component and a refinement component (Murphy 1972; DeGroot and Fienberg 1983; Dawid 1986). Candille and Talagrand (2005) discuss calibration-sharpness decompositions of the continuous ranked probability score.

Table 5 shows the logarithmic score and the continuous ranked probability score for the various forecasters in our initial simulation study, averaged over the 10000 replicates of the prediction experiment. As expected, both scoring rules rank the ideal forecaster highest, followed by Hamill's, the unfocused and the climatological forecaster. Figure 3 plots the Brier score for the associated binary forecasts in dependence on the threshold value, illustrating the integral representation on the right-hand side of (14). This type of display was proposed by Gerds (2002, Section 2.3) and Schumacher *et al.* (2003) who called the graphs prediction error curves.

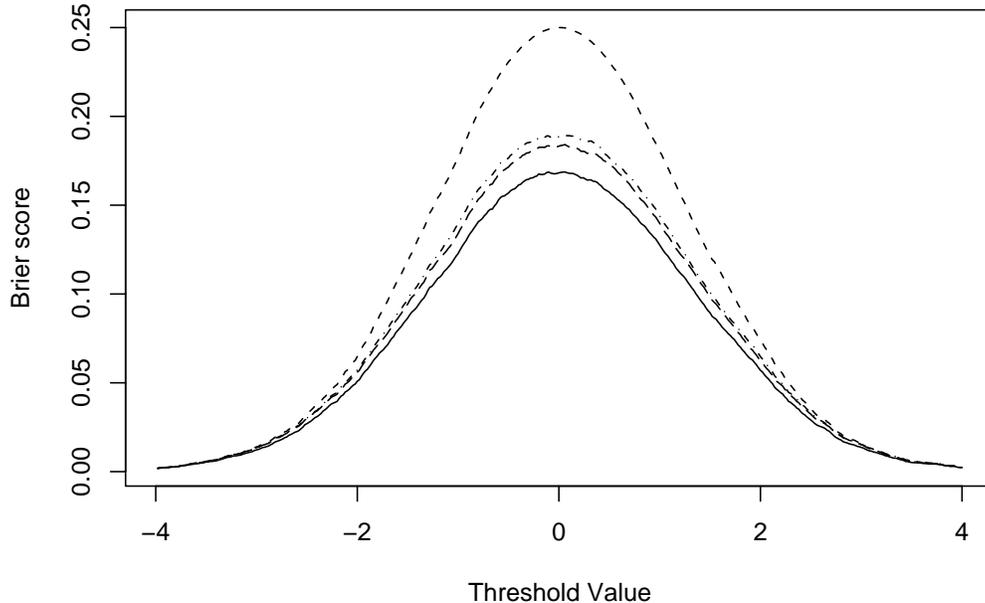


Figure 3: Brier score plot for the ideal forecaster (solid line), climatological forecaster (short dashes), unfocused forecaster (dot-dashed line), and Hamill’s forecaster (long dashes). The graphs show the Brier score as a function of the threshold value. The area under each forecaster’s curve equals the CRPS value (14).

#### 4 Case study: Probabilistic forecasts at the Stateline wind energy center

Wind power is the fastest-growing energy source today. Estimates are that within the next 15 years wind energy will fill about 6% of the electricity supply in the United States. In Denmark, wind energy already meets 20% of the country’s total energy needs. However, arguments against the proliferation of wind energy have been put forth, often focusing on the perceived inability to forecast wind resources with any degree of accuracy. The development of advanced probabilistic forecast methodologies helps address these concerns.

The prevalent approach to short-range forecasts of wind speed and wind power at prediction horizons up to a few hours is based on on-site observations and autoregressive time series models (Brown, Katz and Murphy 1984). Gneiting *et al.* (2004) proposed a novel spatio-temporal approach, the regime-switching space-time (RST) method, that merges meteorological and statistical expertise to obtain fully probabilistic forecasts of wind resources. Henceforth, we illustrate our diagnostic approach to the evaluation of predictive performance by a comparison and ranking of three competing methodologies for two-hour ahead forecasts of hourly average wind speed at the Stateline wind energy center. The evaluation period is May through November 2003, resulting in a total of 5136 probabilistic forecasts.

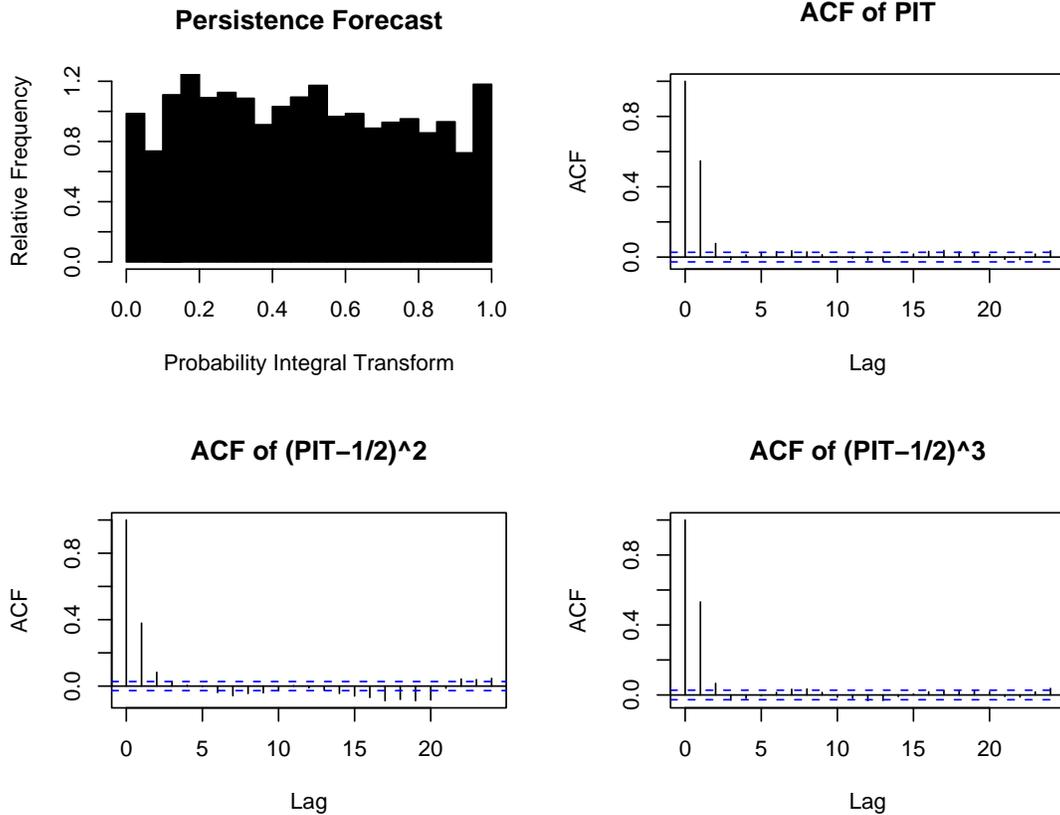


Figure 4: Probability integral transform (PIT) histogram and sample autocorrelation functions for the first three centered moments of the PIT values, for persistence forecasts of hourly average wind speed at the Stateline wind energy center.

#### 4.1 Predictive distributions for hourly average wind speed

We consider three competing statistical prediction algorithms for two-hour ahead probabilistic forecasts of hourly average wind speed,  $w_t$ , at the Stateline wind energy center. Stateline is located on the Vansycle ridge at the border between the states of Oregon and Washington in the U.S. Pacific Northwest. The data source is described in Gneiting *et al.* (2004).

The first method is the persistence forecast, a naive yet surprisingly skillful, nonparametric reference forecast. The persistence point forecast is simply the most recent observed value of hourly average wind speed at Stateline. To obtain a predictive distribution, we dress the point forecast with the 19 most recent observed values of the persistence error, similarly to the approach proposed by Roulston and Smith (2003). Specifically, the predictive CDF for  $w_{t+2}$  is the empirical distribution function of the set

$$\{\max(w_t - w_{t-h} + w_{t-h-2}, 0) : h = 0, \dots, 18\}.$$

The second technique is the autoregressive time series approach, which was proposed by Brown *et al.* (1984) and has found widespread use since. To apply this technique, we fit and extract a diurnal trend component based on a sliding 40-day training period, fit a stationary autoregression

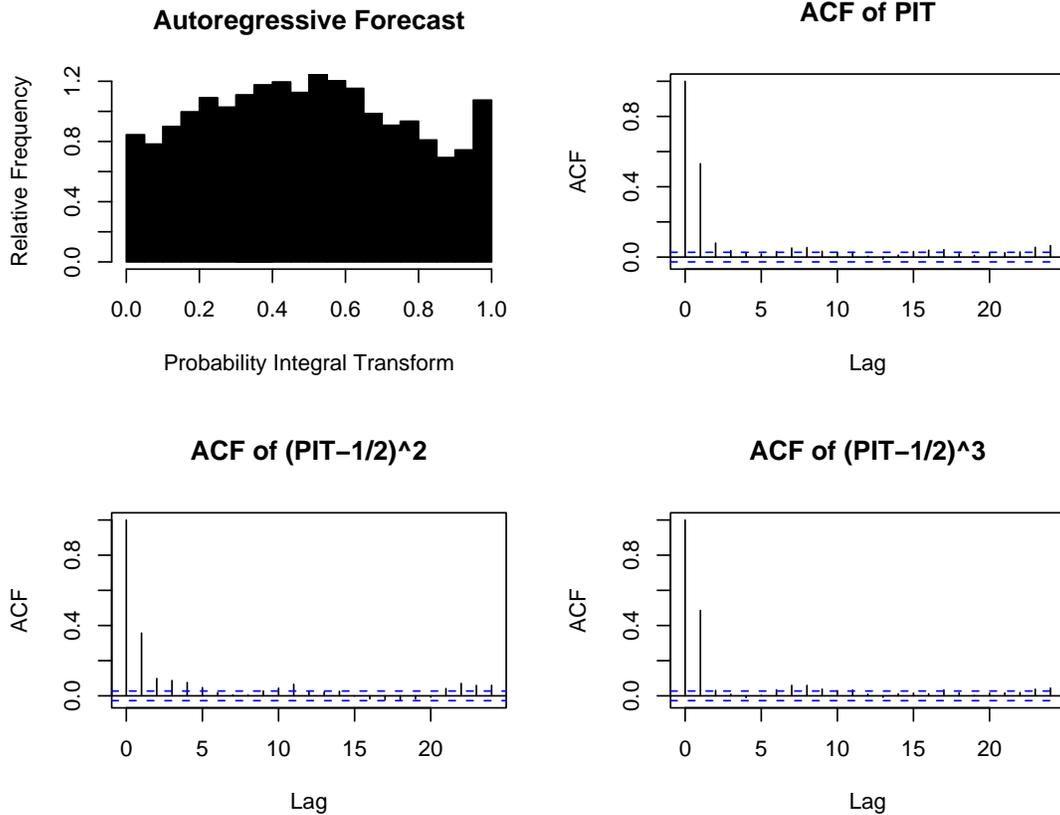


Figure 5: Same as Figure 4, but for autoregressive forecasts.

to the residual component and find a Gaussian predictive distribution in the customary way. The predictive distribution assigns a typically small positive mass to the negative half-axis, and in view of the nonnegativity of the predictand, we redistribute this mass to wind speed zero. The details are described in Gneiting *et al.* (2004), where the method is referred to as the AR-D technique.

The third method is the regime-switching space-time (RST) approach of Gneiting *et al.* (2004). The RST model is parsimonious, yet takes account of all the salient features of wind speed: alternating atmospheric regimes, temporal and spatial autocorrelation, diurnal and seasonal non-stationarity, conditional heteroscedasticity and non-Gaussianity. The method uses offsite information from the nearby meteorological towers at Goodnoe Hills and Kennewick, identifies atmospheric regimes and fits conditional predictive models for each regime, based on a sliding 45-day training period. Details can be found in Gneiting *et al.* (2004), where the method is referred to as the RST-D-CH technique. Any minor discrepancies in the results reported below and in Gneiting *et al.* (2004) stem from the use of R versus SPLUS and differences in optimization routines.

## 4.2 Assessing calibration

Figures 4, 5 and 6 show the probability integral transform (PIT) histograms for the three forecast techniques, along with the sample autocorrelation functions for the first three centered moments of the PIT values and the respective Bartlett confidence intervals. The PIT histograms for the

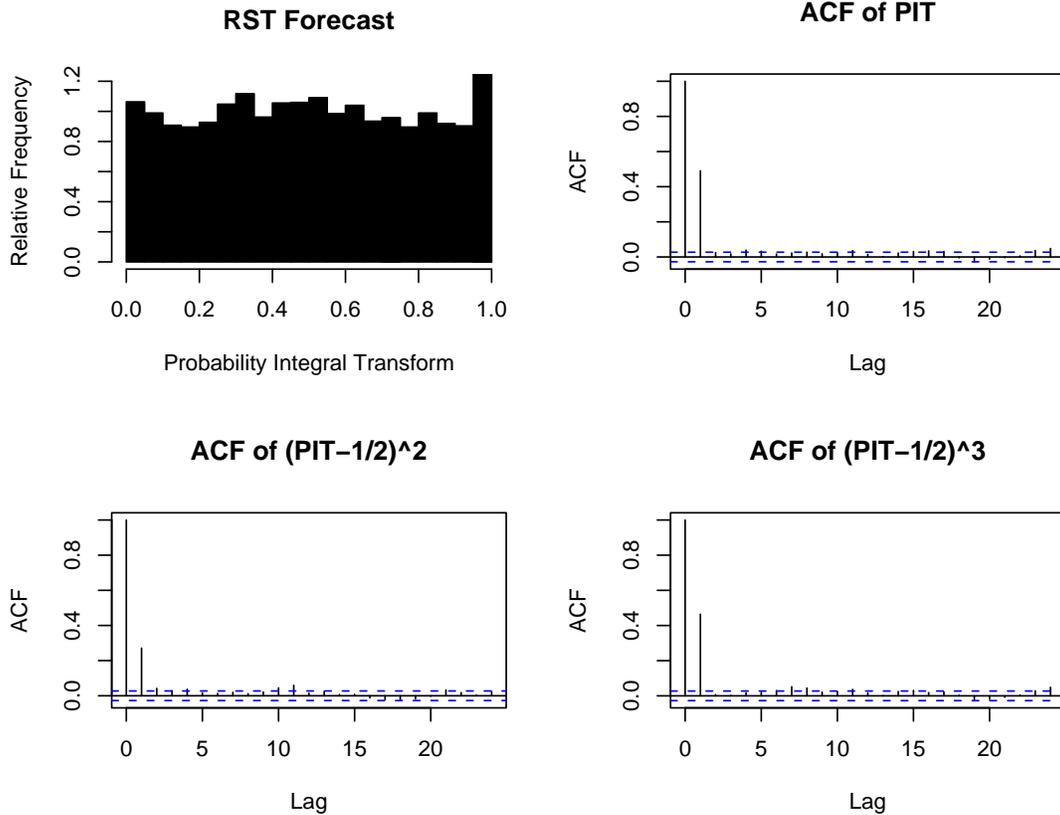


Figure 6: Same as Figure 4, but for RST forecasts.

persistence and RST forecasts appear uniform. The histogram for the autoregressive forecasts is hump shaped, thereby suggesting departures from probabilistic calibration. Table 6 shows the empirical coverage of central prediction intervals.

The PIT values for ideal 2-step ahead forecasts are at most 1-dependent, and the sample autocorrelation functions for the RST forecasts seem compatible with this assumption. The sample autocorrelations for the persistence forecasts were nonnegligible at lag 2, and the centered second moment showed notable negative correlations at lags between 15 and 20 hours. These features indicate a lack of fit of the predictive model, even though they seem hard to interpret diagnostically. The respective sample autocorrelations for the autoregressive forecast were positive and nonnegligible at lags up to five hours, suggesting conditional heteroscedasticity in the wind speed series. Indeed, Gneiting *et al.* (2004) showed that the autoregressive forecasts improve when a conditionally heteroscedastic model is employed. In the current classical autoregressive formulation the predictive variance varies as a result of the sliding training period, but high-frequency changes in predictability are not taken into account.

Figure 7 shows marginal calibration plots for the three forecasts, both in terms of CDFs and in terms of quantiles. The graphs show the differentials (10) and (11) and point at nonnegligible excursions from zero, particularly at small wind speeds and for the autoregressive forecast. The lack of predictive model finds an explanation in Figure 8, which shows the empirical cumulative distribution function,  $\bar{F}_T$ , of hourly average wind speed. Hourly average wind speeds less than 1

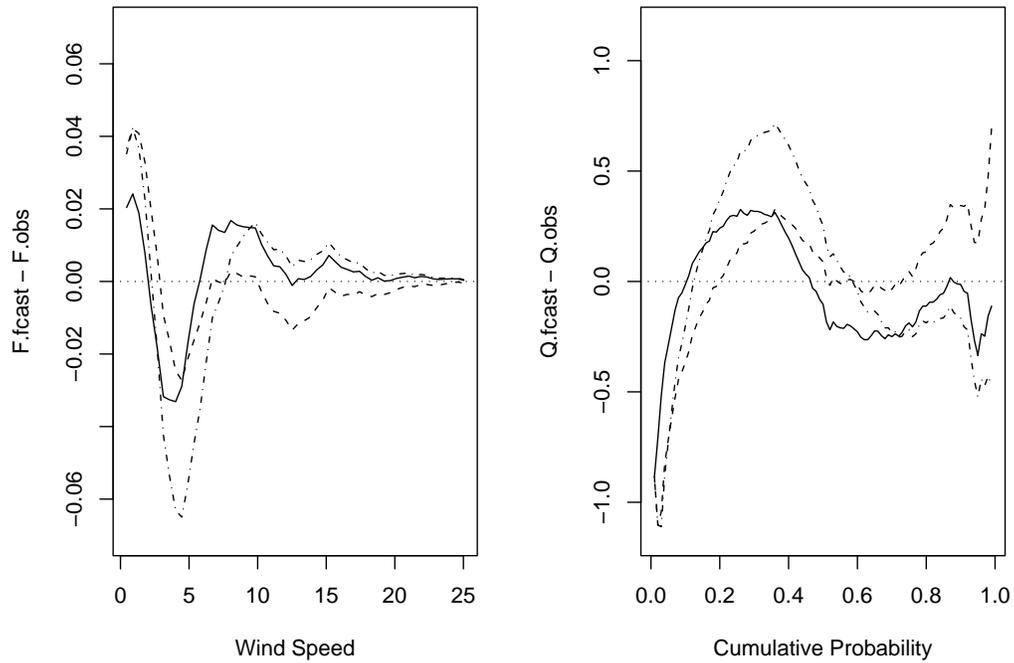


Figure 7: Marginal calibration plot for persistence forecasts (dashed line), autoregressive forecasts (dot-dashed line) and RST forecasts (solid line) of hourly average wind speed at the Stateline wind energy center in terms of CDFs (left) and in terms of quantiles (right), in  $\text{m} \cdot \text{s}^{-1}$ .

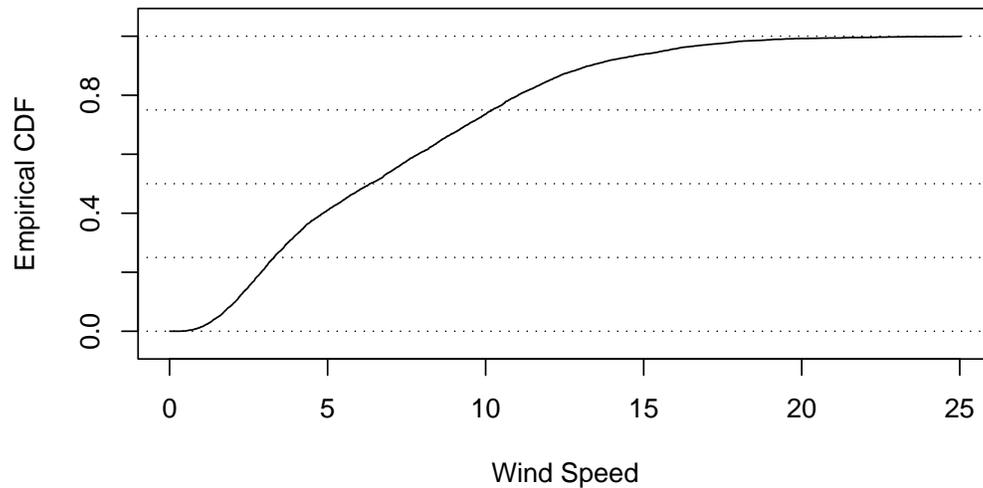


Figure 8: Empirical CDF of hourly average wind speed at the Stateline wind energy center in May through November 2003, in  $\text{m} \cdot \text{s}^{-1}$ .

Table 6: Empirical coverage of central prediction intervals. The nominal coverage is 50% and 90%, respectively.

Interval	50%	90%
Persistence forecast	50.9%	89.2%
Autoregressive forecast	55.6%	90.4%
RST forecast	51.2%	88.4%

Table 7: Average width of central prediction intervals, in  $\text{m} \cdot \text{s}^{-1}$ . The nominal coverage is 50% and 90%, respectively.

Interval	50%	90%
Persistence forecast	2.63	7.51
Autoregressive forecast	2.74	6.55
RST forecast	2.20	5.31

$\text{m} \cdot \text{s}^{-1}$  were almost never observed, even though the predictive distributions assign positive point mass to wind speed zero.

### 4.3 Assessing sharpness

The sharpness diagram in Figure 9 shows boxplots that illustrate the width of central prediction intervals for the 5136 predictive distributions in the evaluation period. The prediction intervals for the persistence forecast varied the most in width, followed by the RST and autoregressive forecasts. Table 7 shows the respective average widths. The RST method was by far the sharpest, with prediction intervals that were about 20% shorter on average than those for the autoregressive technique.

### 4.4 Continuous ranked probability score

Table 8 shows the CRPS value (14) for the various techniques. We report the scores month by month, which allows for an assessment of seasonal effects and straightforward tests of the null hypothesis of no difference in predictive performance. For instance, the RST method showed lower CRPS than the autoregressive technique in each month during the evaluation period, May through November 2003. Under the null hypothesis of equal predictive performance this happens with probability  $(\frac{1}{2})^7 = \frac{1}{128}$  only. Similarly, the autoregressive technique outperformed the persistence method in May through October, but not in November. Various other tests can be employed, but care needs to be taken to avoid dependencies in the forecast differentials. Here, the results for distinct months can be considered independent for all practical purposes. Diebold and Mariano (1995) gave a thoughtful discussion of these issues, and we refer to their work for a comprehensive account of tests of predictive performance. Figure 10 illustrates the Brier score decomposition (14) of the CRPS value for the entire evaluation period. The RST method outperformed the other techniques at all thresholds.

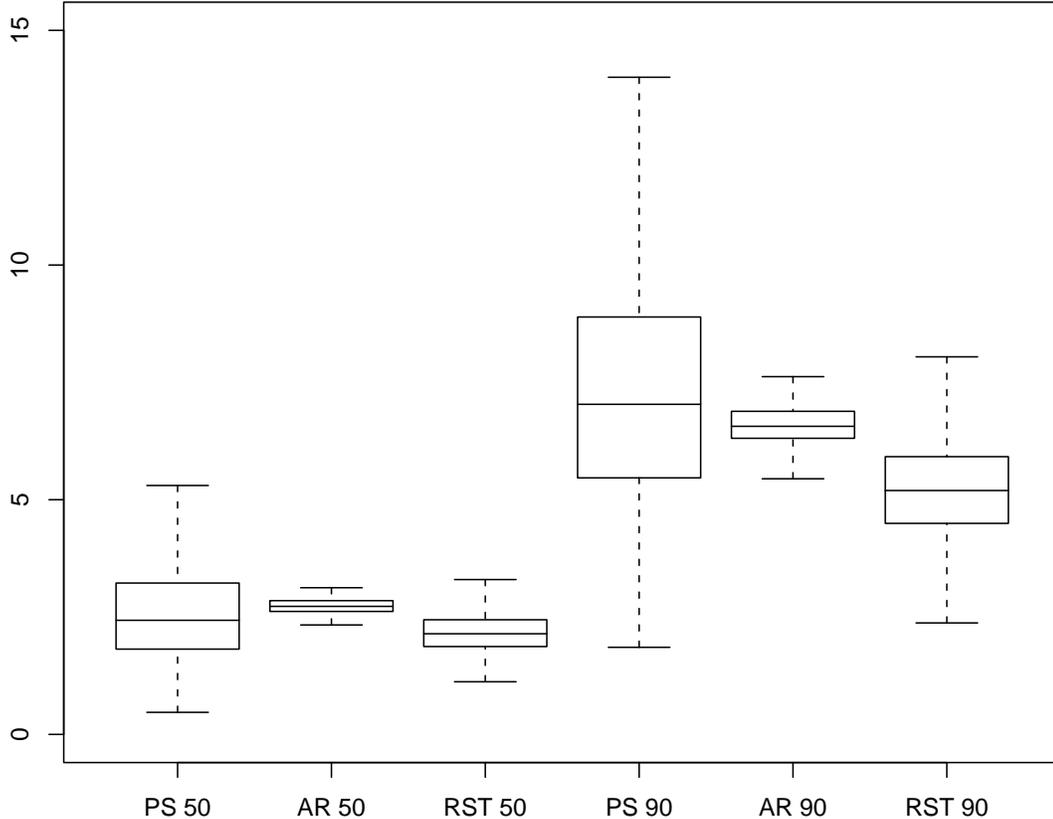


Figure 9: Sharpness diagram for persistence (PS), autoregressive (AR) and RST forecasts of hourly average wind speed at the Stateline wind energy center. The boxplots show the 5th, 25th, 50th, 75th and 95th percentile of the width of the central prediction interval, in  $\text{m} \cdot \text{s}^{-1}$ . The smaller the width, the sharper. The nominal coverage is 50% (left) and 90% (right), respectively.

We noted in Section 3.4 that the continuous ranked probability score generalizes the absolute error, and reduces to the latter for point forecasts. Table 9 shows the respective mean absolute error (MAE) for the persistence, autoregressive and RST point forecasts. The persistence point forecast is the most recent observed value of hourly average wind speed at Vansycle. The autoregressive point forecast is the mean of the respective predictive distribution, and similarly for the RST forecast. The RST method had the lowest MAE, followed by the autoregressive and persistence techniques. The MAE and CRPS values are reported in the same unit as the wind speed observations, that is, in  $\text{m} \cdot \text{s}^{-1}$ , and can be directly compared. The insights that the monthly scores provide are indicative of the potential benefits of thoughtful stratification.

The CRPS and MAE values establish a clear-cut ranking of the forecast methodologies that places the RST method first, followed by the autoregressive and persistence techniques. The RST method also performed best in terms of probabilistic and marginal calibration, and the RST forecasts were by far the sharpest. The diagnostic approach points at forecast deficiencies and suggests potential improvements to the predictive models. In particular, the marginal calibration plots in Figure 7 suggest a modified version of the RST technique that uses truncated normal rather than

Table 8: CRPS value (14) for probabilistic forecasts of hourly average wind speed at the Stateline wind energy center in March through November 2003, month by month and for the entire evaluation period, in  $\text{m} \cdot \text{s}^{-1}$ .

CRPS	May	Jun	Jul	Aug	Sep	Oct	Nov	Mar–Nov
Persistence forecast	1.16	1.08	1.29	1.21	1.20	1.29	1.16	1.20
Autoregressive forecast	1.12	1.02	1.10	1.11	1.11	1.22	1.13	1.12
RST forecast	0.96	0.85	0.95	0.95	0.97	1.08	1.00	0.97

Table 9: Mean absolute error (MAE) for point forecasts of hourly average wind speed at the Stateline wind energy center in March through November 2003, month by month and for the entire evaluation period, in  $\text{m} \cdot \text{s}^{-1}$ .

MAE	May	Jun	Jul	Aug	Sep	Oct	Nov	Mar–Nov
Persistence forecast	1.60	1.45	1.74	1.68	1.59	1.68	1.51	1.61
Autoregressive forecast	1.53	1.38	1.50	1.54	1.53	1.68	1.54	1.53
RST forecast	1.32	1.18	1.33	1.31	1.36	1.48	1.37	1.34

cut-off normal predictive distributions. This modification yields small but consistent improvements in predictive performance (Gneiting *et al.* 2006).

## 5 Discussion

Our paper addressed the important issue of evaluating predictive performance for probabilistic forecasts of continuous variables. Following the lead of Dawid (1984) and Diebold *et al.* (1998), predictive distributions have traditionally been evaluated by checking the uniformity of the probability integral transform (PIT). We introduced the pragmatic and flexible paradigm of *maximizing the sharpness of the predictive distributions subject to calibration*. Calibration refers to the statistical consistency between the predictive distributions and the associated observations and is a joint property of the predictions and the values that materialize. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only.

We interpreted probabilistic forecasting within a simple theoretical framework that allowed us to distinguish probabilistic, exceedance and marginal calibration, and that lends support to the use of diagnostic tools in evaluating and comparing probabilistic forecasters. Probabilistic calibration corresponds to the uniformity of the PIT values, and the PIT histogram remains a key tool in the diagnostic approach to forecast evaluation. In addition, we proposed the use of marginal calibration plots, sharpness diagrams and proper scoring rules, which form powerful tools for learning about forecast deficiencies and ranking competing forecast methodologies. Our own applied work on probabilistic forecasting has benefitted immensely from these tools, as documented in Section 4 and in the partial applications in Gneiting *et al.* (2004), Raftery *et al.* (2005) and Gneiting *et al.* (2005). Predictive distributions can be reduced to point forecasts, or to probability forecasts of binary events, and the associated forecasts can be assessed using the diagnostic devices described by Murphy, Brown and Chen (1989) and Murphy and Winkler (1992), among others.

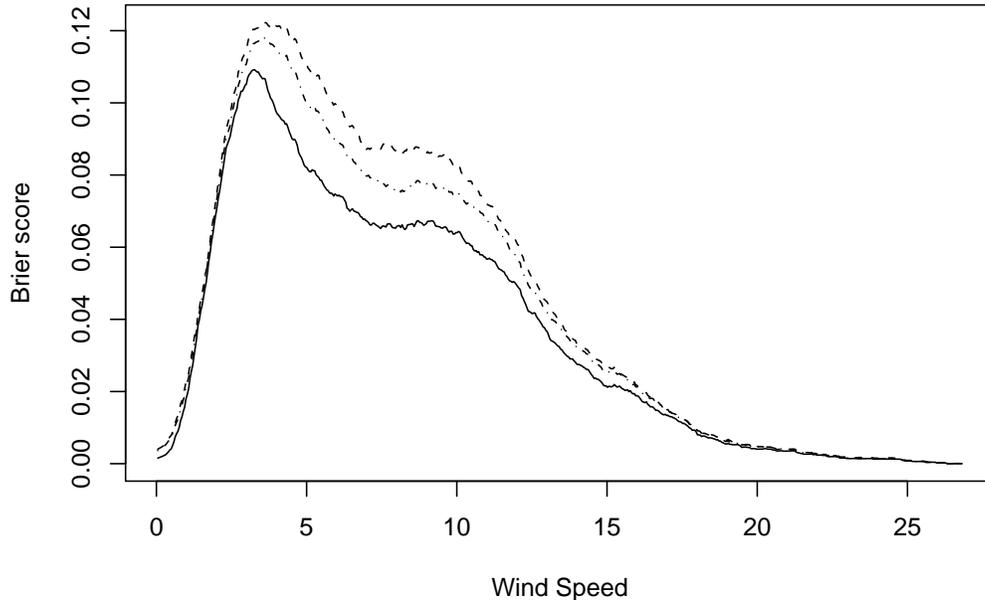


Figure 10: Brier score plot for persistence forecasts (dashed line), autoregressive forecasts (dot-dashed line) and RST forecasts (solid line) of hourly average wind speed at the Stateline wind energy center, in  $\text{m}\cdot\text{s}^{-1}$ . The graphs show the Brier score as a function of the threshold value. The area under each forecast’s curve equals the CRPS value (14).

If we were to reduce our conclusions to a single recommendation, we would close with a call for the assessment of sharpness, particularly when the goal is that of ranking. Previous comparative studies of the predictive performance of probabilistic forecasts have focused on calibration. For instance, Moyeed and Papritz (2002) compared spatial prediction techniques, Clements and Smith (2000) and Boero and Marrocu (2004) evaluated linear and non-linear time series models, Garrat *et al.* (2003) assessed macroeconomic forecast models, and Bauwens *et al.* (2004) studied the predictive performance of financial duration models. In each of these works, the assessment was best on the predictive performance of the associated point forecasts, and on the uniformity of the PIT values. We contend that comparative studies of these types call for routine assessments of sharpness, in the form of sharpness diagrams and through the use of proper scoring rules.

Despite the frequentist flavor of our diagnostic approach, calibration and sharpness are properties that are relevant to Bayesian forecasters as well. Rubin (1984, pp. 1161 and 1160) argued that “the probabilities attached to Bayesian statements do have frequency interpretations that tie the statements to verifiable real world events.” Consequently, a “Bayesian is calibrated if his probability statements have their asserted coverage in repeated experience.” Gelman *et al.* (1996) developed Rubin’s posterior predictive approach, proposed posterior predictive checks as Bayesian counterparts to the classical tests for goodness of fit, and advocated their use in judging the fit of Bayesian models. This relates to our diagnostic approach, which emphasizes the need for understanding the ways in which predictive distributions fail or succeed. Indeed, the diagnostic devices posited herein form powerful tools for Bayesian as well as frequentist model diagnostics and model choice. Tools such as the PIT histogram, marginal calibration plots, sharpness diagrams and proper scoring rules are widely applicable, since they are nonparametric, do not depend on nested models,

allow for structural change, and apply to predictive distributions that are represented by samples, as they arise in a rapidly growing number of Markov chain Monte Carlo methodologies and ensemble prediction systems. In the time series context, the predictive framework is natural and model fit can be assessed through the performance of the time-forward predictive distributions (Smith 1985; Shephard 1994; Frühwirth-Schnatter 1996). In other types of situations, cross-validatory approaches often apply fruitfully (Dawid 1984, p. 288; Gneiting and Raftery 2004).

## Appendix

### Proof of Theorem 1

Consider the random variable  $U = F(x_1)^{z_1} F(x_2)^{z_2} \dots F(x_T)^{z_T}$  where  $x_1 \sim G_1, \dots, x_T \sim G_T$  and  $(z_1, \dots, z_T)'$  is multinomial with equal probabilities. The finite probabilistic calibration condition implies that  $U$  is uniformly distributed. By the variance decomposition formula,

$$\text{var}(F) = \text{var}(F^{-1}(U)) = \text{E} \left[ \text{var} \left( F^{-1}(U) \mid z_1, \dots, z_T \right) \right] + \text{var} \left[ \text{E} \left( F^{-1}(U) \mid z_1, \dots, z_T \right) \right].$$

The first term in the decomposition equals

$$\frac{1}{T} \sum_{t=1}^T \text{var}(x_t) = \frac{1}{T} \sum_{t=1}^T \text{var}(G_t)$$

and the second term is nonnegative and vanishes if and only if  $\text{E}(G_1) = \dots = \text{E}(G_T)$ . ■

### Proof of Theorem 2

For  $p \in (0, 1)$  and  $t = 1, 2, \dots$ , put  $Y_t = \mathbf{1}\{p_t < p\} - G_t \circ F_t^{-1}(p)$  and note that  $E(Y_t) = 0$ . By Theorem 2 of Blum *et al.* (1963),

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T Y_t = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \left( \mathbf{1}\{p_t < p\} - G_t \circ F_t^{-1}(p) \right) = 0$$

almost surely. The uniqueness of the limit implies that (6) is equivalent to the probabilistic calibration condition (3). ■

### Proof of Theorem 3

For  $x \in \mathbb{R}$  let  $q = \bar{F}(x)$ , and for  $t = 1, 2, \dots$  put  $q_t = \bar{F}(x_t)$ . Then

$$\hat{G}_T(x) = \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{x_t \leq x\} = \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{q_t \leq q\}.$$

By Theorem 2 with  $F_t = \bar{F}$  for  $t = 1, 2, \dots$ , we have that  $\frac{1}{T} \sum_{t=1}^T \mathbf{1}\{q_t \leq q\} \rightarrow q$  almost surely if and only if  $\frac{1}{T} \sum_{t=1}^T G_t \circ \bar{F}^{-1}(q) \rightarrow q$  almost surely. Hence, marginal calibration is equivalent to (9). ■

## Acknowledgements

We thank John B. Bremnes, Barbara G. Brown, Richard E. Chandler, Theo Eicher, Hans R. Künsch, Kristin Larson, Jon A. Wellner, Anton H. Westveld and Kenneth Westrick for discussions, comments and references, and we are grateful to an anonymous reviewer for a wealth of helpful suggestions. Our work has been supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745. Tilmann Gneiting furthermore acknowledges support by the National Science Foundation under Award no. 0134264 and by the Washington Technology Center, and thanks the Soil Physics group at Universität Bayreuth, Germany, where part of this work was performed.

## References

- Anderson, J. L. (1996), A method for producing and evaluating probabilistic forecasts from ensemble model integrations, *Journal of Climate*, **9**, 1518–1530.
- Bauwens, L., Giot, P., Grammig, J. and Veredas, D. (2004), A comparison of financial duration models via density forecasts, *International Journal of Forecasting*, **20**, 589–609.
- Berkowitz, J. (2001), Testing density forecasts, with applications to risk management, *Journal of Business and Economic Statistics*, **19**, 465–474.
- Bernardo, J. M. (1979), Expected information as expected utility, *Annals of Statistics*, **7**, 686–690.
- Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995), Bayesian computing and stochastic systems (with discussion and rejoinder), *Statistical Science*, **10**, 3–66.
- Blum, J. R., Hanson, D. L. and Koopmans, L. H. (1963), On the strong law of large numbers for a class of stochastic processes, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **2**, 1–11.
- Boero, G. and Marrocu, E. (2004), The performance of SETAR models: A regime conditional evaluation of point, interval and density forecasts, *International Journal of Forecasting*, **20**, 305–320.
- Bremnes, J. B. (2004), Probabilistic forecasts of precipitation in terms of quantiles using NWP model output, *Monthly Weather Review*, **132**, 338–347.
- Brier, G. W. (1950), Verification of forecasts expressed in terms of probability, *Monthly Weather Review*, **78**, 1–3.
- Brown, B. G., Katz, R. W. and Murphy, A. H. (1984), Time series models to simulate and forecast wind speed and wind power, *Journal of Climate and Applied Meteorology*, **23**, 1184–1195.
- Campbell, S. D. and Diebold, F. X. (2005), Weather forecasting for weather derivatives, *Journal of the American Statistical Association*, **100**, 6–16.
- Candille, G. and Talagrand, O. (2005), Evaluation of probabilistic prediction systems for a scalar variable, *Quarterly Journal of the Royal Meteorological Society*, **131**, 2131–2150.
- Christoffersen, P. F. (1998), Evaluating interval forecasts, *International Economic Review*, **39**, 841–862.
- Clements, M. P. and Smith, J. (2000), Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment, *Journal of Forecasting*, **19**, 255–276.

- Dawid, A. P. (1982), The well-calibrated Bayesian, *Journal of the American Statistical Association*, **77**, 605–610.
- Dawid, A. P. (1984), Statistical theory: The prequential approach (with discussion), *Journal of the Royal Statistical Society Ser. A*, **147**, 278–292.
- Dawid, A. P. (1985a), The impossibility of inductive inference, *Journal of the American Statistical Association*, **80**, 340–341.
- Dawid, A. P. (1985b), Calibration-based empirical probability (with discussion and rejoinder), *Annals of Statistics*, **13**, 1251–1285.
- Dawid, A. P. (1986), Probability forecasting, in *Encyclopedia of Statistical Sciences* (eds. S. Kotz, N. L. Johnson and C. B. Read), vol. 7, pp. 210–218, New York: Wiley.
- Dawid, A. P. and Vovk, V. G. (1999), Prequential probability: Principles and properties, *Bernoulli*, **5**, 125–162.
- DeGroot, M. H. and Fienberg, S. E. (1982), Assessing probability assessors: Calibration and refinement, in *Statistical Decision Theory and Related Topics III*, (eds. S. S. Gupta and J. O. Berger), vol. 1, pp. 291–314, New York: Academic Press.
- DeGroot, M. H. and Fienberg, S. E. (1983), The comparison and evaluation of probability forecasters, *Statistician*, **12**, 12–22.
- Diebold, F. X. and Mariano, R. S. (1995), Comparing predictive accuracy, *Journal of Business and Economic Statistics*, **13**, 253–263.
- Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998), Evaluating density forecasts with applications to financial risk management, *International Economic Review*, **39**, 863–883.
- Duffie, D. and Pan, J. (1997), An overview of value at risk, *Journal of Derivatives*, **4**, 7–49.
- Foster, D. P. and Vohra, R. V. (1998), Asymptotic calibration, *Biometrika*, **85**, 379–390.
- Frühwirth-Schnatter, S. (1996), Recursive residuals and model diagnostics for normal and non-normal state space models, *Environmental and Ecological Statistics*, **3**, 291–309.
- Garratt, A., Lee, K., Pesaran, M. H. and Shin, Y. (2003), Forecast uncertainties in macroeconomic modelling: An application to the UK economy, *Journal of the American Statistical Association*, **98**, 829–838.
- Gelman, A., Meng, X.-L. and Stern, H. (1996), Posterior predictive assessment of model fitness via realized discrepancies, *Statistica Sinica*, **6**, 733–807.
- Gerds, T. (2002), Nonparametric efficient estimation of prediction error for incomplete data models, Ph. D. Thesis, Mathematische Fakultät, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany.
- Gneiting, T. and Raftery, A. E. (2004), Strictly proper scoring rules, prediction, and estimation, Technical Report no. 463, Department of Statistics, University of Washington.
- Gneiting, T. and Raftery, A. E. (2005), Weather forecasting with ensemble methods, *Science*, **310**, 248–249.
- Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005), Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, *Monthly Weather Review*, **133**, 1098–1198.
- Gneiting, T., Larson, K., Westrick, K., Genton, M. G. and Aldrich, E. (2004), Calibrated probabilistic forecasting at the Stateline wind energy center: The regime-switching space-time (RST)

- method, Technical Report no. 464, Department of Statistics, University of Washington.
- Gneiting, T., Larson, K., Westrick, K., Genton, M. G. and Aldrich, E. (2006), Calibrated probabilistic forecasting at the Stateline wind energy center: The regime-switching space-time (RST) method, *Journal of the American Statistical Association*, in press.
- Good, I. J. (1952), Rational decisions, *Journal of the Royal Statistical Society Ser. B*, **14**, 107–114.
- Granger, C. W. J. (2006), Preface: Some thoughts on the future of forecasting, *Oxford Bulletin of Economics and Statistics*, **67S**, 707–711.
- Hamill, T. M. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Monthly Weather Review*, **129**, 550–560.
- Hamill, T. M. and Colucci, S. J. (1997), Verification of Eta-RSM short-range ensemble forecasts, *Monthly Weather Review*, **125**, 1312–1327.
- Hoeting, J. (1994), Accounting for model uncertainty in linear regression, Ph. D. Thesis, Department of Statistics, University of Washington, Seattle, Washington.
- Jolliffe, I. T. and Stephenson, D. B., eds. (2003), *Forecast Verification. A Practitioner's Guide in Atmospheric Science*, Chichester, United Kingdom: Wiley.
- Krzysztofowicz, R. (1999), Bayesian theory of probabilistic forecasting via deterministic hydrologic model, *Water Resources Research*, **35**, 2739–2750.
- Krzysztofowicz, R. and Sigrest, A. A. (1999), Calibration of probabilistic quantitative precipitation forecasts, *Weather and Forecasting*, **14**, 427–442.
- Moyeed, R. A. and Papritz, A. (2002), An empirical comparison of kriging methods for nonlinear spatial point prediction, *Mathematical Geology*, **34**, 365–386.
- Murphy, A. H. (1972), Scalar and vector partitions of the probability score. Part I: Two-state situation, *Journal of Applied Meteorology*, **11**, 273–278.
- Murphy, A. H. and Winkler, R. L. (1987), A general framework for forecast verification, *Monthly Weather Review*, **115**, 1330–1338.
- Murphy, A. H. and Winkler, R. L. (1992), Diagnostic verification of probability forecasts, *International Journal of Forecasting*, **7**, 435–455.
- Murphy, A. H., Brown, B. G. and Chen, Y.-S. (1989), Diagnostic verification of temperature forecasts, *Weather and Forecasting*, **4**, 485–501.
- Noceti, P., Smith, J. and Hodges, S. (2003), An evaluation of tests of distributional forecasts, *Journal of Forecasting*, **22**, 447–455.
- Oakes, D. (1985), Self-calibrating priors do not exist, *Journal of the American Statistical Association*, **80**, 339.
- Palmer, T. N. (2002), The economic value of ensemble forecasts as a tool for risk assessment: From days to decades, *Quarterly Journal of the Royal Meteorological Society*, **128**, 747–774.
- Pearson, K. (1933), On a method of determining whether a sample of size  $n$  supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random, *Biometrika*, **25**, 379–410.
- Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997), Bayesian model averaging for linear regression models, *Journal of the American Statistical Association*, **92**, 179–191.
- Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Monthly Weather Review*, **133**, 1155–1174.

- Rosenblatt, M. (1952), Remarks on a multivariate transformation, *Annals of Mathematical Statistics*, **23**, 470–472.
- Roulston, M. S. and Smith, L. A. (2002), Evaluating probabilistic forecasts using information theory, *Monthly Weather Review*, **130**, 1653–1660.
- Roulston, M. S. and Smith, L. A. (2003), Combining dynamical and statistical ensembles, *Tellus*, **A55**, 16–30.
- Rubin, D. B. (1984), Bayesianly justifiable and relevant frequency calculations for the applied statistician, *Annals of Statistics*, **12**, 1151–1172.
- Sandroni, A., Smorodinsky, R. and Vohra, R. V. (2003), Calibration with many checking rules, *Mathematics of Operations Research*, **28**, 141–153.
- Schervish, M. J. (1985), Comment, *Journal of the American Statistical Association*, **80**, 341–342.
- Schervish, M. J. (1989), A general method for comparing probability assessors, *Annals of Statistics*, **17**, 1856–1879.
- Schumacher, M., Graf, E. and Gerds, T. (2003), How to assess prognostic models for survival data: A case study in oncology, *Methods of Information in Medicine*, **42**, 564–571.
- Seillier-Moisewitsch, F. (1993), Sequential probability forecasts and the probability integral transform, *International Statistical Review*, **61**, 395–408.
- Selten, R. (1998), Axiomatic characterization of the quadratic scoring rule, *Experimental Economics*, **1**, 43–62.
- Shafer, G. and Vovk, V. (2001), *Probability and Finance: It's Only a Game!*, New York: Wiley.
- Shephard, N. (1994), Partial non-Gaussian state space, *Biometrika*, **81**, 115–131.
- Smith, J. Q. (1985), Diagnostic checks of non-standard time series models, *Journal of Forecasting*, **4**, 283–291.
- Staël von Holstein, C.-A. S. (1970), *Assessment and Evaluation of Subjective Probability Distributions*, Stockholm, Sweden: Economics Research Institute, Stockholm School of Economics.
- Talagrand, O., Vautard, R. and Strauss, B. (1997), Evaluation of probabilistic prediction systems, in *Proceedings of a Workshop held at ECMWF on Predictability*, 20–22 October 1997, Reading, United Kingdom: European Centre for Medium-Range Weather Forecasts, pp. 1–25.
- Vovk, V. and Shafer, G. (2005), Good randomized sequential probability forecasting is always possible, *Journal of the Royal Statistical Society Ser. B*, **67**, 747–763.
- 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.
- Weigend, A. S. and Shi, S. (2000), Predicting daily probability distributions of S&P500 returns, *Journal of Forecasting*, **19**, 375–392.
- Winkler, R. L. (1977), Rewarding expertise in probability assessment, in *Decision Making and Change in Human Affairs*, Jungermann, H. and de Zeeuw, G., eds., Dordrecht, Holland: D. Reidel, pp. 127–140.