J. R. Statist. Soc. B (2007)
69, *Part* 2, *pp.* 243–268

Probabilistic forecasts, calibration and sharpness

Tilmann Gneiting,

University of Washington, Seattle, USA

Fadoua Balabdaoui

Georg-August-Universität Göttingen, Germany

and Adrian E. Raftery

University of Washington, Seattle, USA

[Received May 2005. Final revision October 2006]

Summary. Probabilistic forecasts of continuous variables 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 allows us to distinguish between probabilistic calibration, exceedance calibration and marginal calibration. We propose and study tools for checking calibration and sharpness, among them the probability integral transform 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 centre in the US 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 (Granger, 2006), and the rapidly growing area of financial risk management is dedicated to probabilistic forecasts of portfolio values (Duffie and Pan, 1997). 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

Address for correspondence: Tilmann Gneiting, Department of Statistics, University of Washington, Seattle, WA 98195-4322, USA.
E-mail: tilmann@stat.washington.edu

1369–7412/07/69243

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; Corradi and Swanson, 2006). 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 whereas the observations are real valued.

We employ the following, simple theoretical framework to provide guidance in methodological work. At times or instances $t=1,2,\ldots,$ 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 CDF 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 basis of information is at most that of nature. Hence, if

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

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, regardless 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 PIT is a necessary condition for the forecaster to be ideal, and checks for its uniformity have formed a corner-stone 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 that is described in Table 1. At times or instances $t=1,2,\ldots,$ nature draws a standard normal random number μ_t and picks the data-generating distribution $G_t = \mathcal{N}(\mu_t, 1)$. In the context of weather forecasts, we might think of μ_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 μ_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

Table 1. Scenario for the simulation study†

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

†At times $t = 1, 2, \ldots, 10000$, nature picks a distribution G_t, and the 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 μ and variance σ^2. The sequences $(\mu_t)_{t=1,2,\ldots}$, $(\tau_t)_{t=1,2,\ldots}$ and $(\delta_t, \sigma_t^2)_{t=1,2,\ldots}$ are independent identically distributed and independent of each other.

forecaster observes the current state μ_t but adds a mixture component to the forecast, which can be interpreted as distributional 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. Fig. 1 shows that the PIT histograms for the four forecasters are essentially uniform.

In view of the reliance on the PIT 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 PIT cannot 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 ofthe predic-*

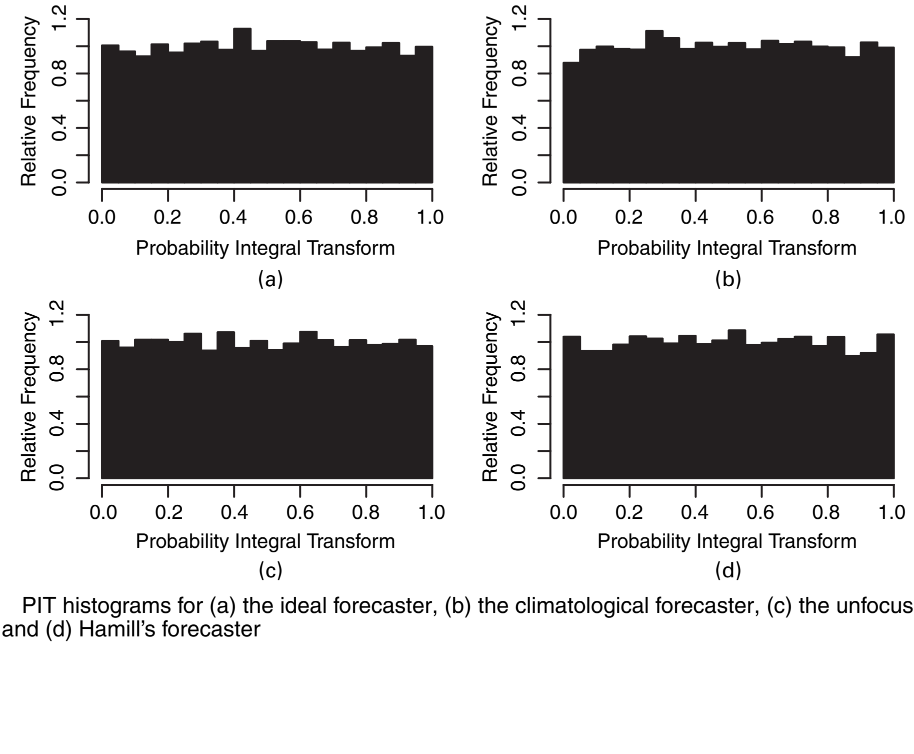


Fig. 1. PIT histograms for (a) the ideal forecaster, (b) the climatological forecaster, (c) the unfocused forecaster and (d) Hamill's forecaster

tive 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 are, 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 counter-examples, 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 us to rank competing forecast procedures. Section 4 turns to a case-study on probabilistic forecasts at the Stateline wind energy centre in the US 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 non-parametric tools for statistical model selection and model 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

Our theoretical framework is as follows. At times or instances $t = 1, 2, \ldots$, nature picks a probability distribution G_t, and the 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 basis of information 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 $\mathbb{R}$. Evidently, G_t is not observed in practice, and any operational evaluation needs to be performed on the basis of the forecasts F_t and the outcomes x_t only.

In comparing forecasters, we take the pragmatic standpoint of a user who is to rank and choose between a number of competitors, as exemplified in the case-study 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 bases of information.

Our approach seems slightly broader than Dawid's (1984) prequential framework, in that we think of $(F_t)_{t=1,2,\ldots}$ as a general countable sequence of forecasts, with the index referring to time, space and/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, $(F_t)_{t=1,2,\ldots}$ and $(G_t)_{t=1,2,\ldots}$ denote sequences of continuous and strictly increasing CDFs, possibly depending on stochastic parameters. We think of $(G_t)_{t=1,2,\ldots}$ as the true data-generating process and of $(F_t)_{t=1,2,\ldots}$ 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 $(F_t)_{t=1,2,\ldots}$ and $(G_t)_{t=1,2,\ldots}$ might depend on stochastic parameters, convergence is understood as almost sure convergence, as $T \to \infty$, and is denoted by an arrow. For now, these notions are of theoretical interest only; in Section 3, 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

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

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

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

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

$$\bar{G}(x) = \lim_{T\to\infty}\left\{\frac{1}{T}\sum_{t=1}^{T} G_t(x)\right\}$$

and

$$\bar{F}(x) = \lim_{T\to\infty}\left\{\frac{1}{T}\sum_{t=1}^{T} F_t(x)\right\}$$

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 researchers have studied calibration in the context of probability forecasts for sequences of binary events (DeGroot and Fienberg, 1982; Dawid, 1982, 1985a, b; 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). This 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, whereas the other loses. No goal is 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 of the existence of certain types of strategy for the forecaster. Shafer and Vovk considered probability forecasts for dichotomous events, rather than distributional forecasts of real-valued quantities, and they did not consider the problem that is tackled here, namely the comparative evaluation of competing forecasters, for which they hint at future work (Shafer and Vovk (2001), page 50).

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, $(\mu_t)_{t=1,2,\ldots}$, $(\sigma_t)_{t=1,2,\ldots}$ and $(\tau_t)_{t=1,2,\ldots}$ denote independent sequences of independent identically distributed random variables. We write $\mathcal{N}(\mu,\sigma^2)$ for the normal distribution with mean μ and variance σ^2, identify distributions and CDFs, and let Φ denote the standard normal CDF. In each example, nature draws a random number $\mu_t \sim \mathcal{N}(0,1)$ that corresponds to the basis of information at time t and picks the data-generating distribution $G_t = \mathcal{N}(\mu_t, 1)$. We recall that convergence is understood as almost sure convergence, as $T \to \infty$, with respect to the law of the aforementioned sequences.

2.2.1. *Example 1 (ideal forecaster)*

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

2.2.2. *Example 2 (climatological forecaster)*

The climatological forecaster issues the distributional forecast $F_t = \mathcal{N}(0,2)$, regardless of t. This forecaster is probabilistically calibrated and marginally calibrated, as can be seen by using arguments based on densities. However,

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

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 marginal 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 expression (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.

2.2.3. *Example 3 (unfocused forecaster)*

The predictive distribution of the unfocused forecaster is the mixture distribution

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

where τ_t is either 1 or -1, with equal probabilities, and independent of μ_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

$$\frac{1}{T}\sum_{t=1}^{T} G_t \circ F_t^{-1}(p) \to \frac{1}{2}\{\Phi \circ \Phi_{+}^{-1}(p) + \Phi \circ \Phi_{-}^{-1}(p)\} = p,$$

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

$$\frac{1}{T}\sum_{t=1}^{T} G_t^{-1} \circ F_t(x) \to \frac{1}{2}\{\Phi^{-1} \circ \Phi_{+}(x) + \Phi^{-1} \circ \Phi_{-}(x)\} \neq x$$

Table 2. The three major modes of calibration are logically independent of each other and may occur in any combination†

Properties	*Example*
PEM	Example 1 (ideal forecaster)
PEM̄	$G_t = F_t = \mathcal{N}(t, 1)$
PĒM	Example 2 (climatological forecaster)
PĒM̄	Example 3 (unfocused forecaster)
P̄EM̄	Example 4 (mean-biased forecaster)
P̄EM	Example 5 (sign-biased forecaster)
P̄ĒM	Example 6 (mixed forecaster)
P̄ĒM̄	$G_t = \mathcal{N}(0, 1)$, $F_t = \mathcal{N}(1, 1)$

†For instance, the unfocused forecaster in example 3 is probabilistically calibrated (P), but neither exceedance calibrated (Ē) nor marginally calibrated (P̄).

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

2.2.4 Example 4 (mean-biased forecaster)

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

2.2.5 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.

2.2.6 Example 6 (mixed forecaster)

The mixed forecaster randomizes between the climatological and the sign-biased forecast, with equal probabilities and independent of μ_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 these 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 μ_t is a standard normal random number. 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,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} G_t \circ F_t^{-1}(p) &\rightarrow \frac{1}{3}\left[\Phi\left\{\Phi^{-1}(p) - \frac{1}{2}\right\} + \Phi\left\{\frac{13}{10}\Phi^{-1}(p)\right\} + \Phi\left\{\Phi^{-1}(p) + \frac{1}{2}\right\}\right] \\ &= p + \varepsilon(p),\end{aligned}$$

where $|\varepsilon(p)| \leqslant 0.0032$ for all p but $\varepsilon(p) \neq 0$ 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

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

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

2.4. Sharpness principle

In view of our assumption that the forecaster's basis of information 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 counter-example 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,

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

for finite sequences $(F_t)_{1\leqslant t\leqslant T}$ relative to $(G_t)_{1\leqslant t\leqslant T}$, 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 CDFs with associated densities that are symmetric about zero and have finite variances, $\operatorname{var}(G_1)=\sigma^2$ and $\operatorname{var}(G_2)=\lambda\sigma^2$. If we define

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

$$F_2(x)=F_1(ax),$$

then

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

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

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

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

Then

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

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

For climatological forecasts, finite probabilistic calibration and finite marginal calibration are equivalent, and a weak type of the sharpness principle holds, in the form of a lower bound on the variance of the predictive distribution.

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

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

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

The proof of theorem 1 is given in Appendix A. 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 hypothetical in practice. Hence, we turn to sample versions, by substituting empirical distribution functions based on the outcomes. The resulting methodological tools stand in their own right; however, our theoretical framework lends support and reassurance. In what follows, this programme is carried out for probabilistic calibration and marginal calibration. Exceedance calibration does not allow for an obvious sample analogue, and it is not clear whether such an analogue 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 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 PIT can be traced back at least to Pearson (1933). The connection to probabilistic calibration is established by substituting the empirical distribution function $\mathbf{1}(x_t \leqslant x)$ for the data-generating distribution $G_t(x)$, $x \in \mathbb{R}$, in the probabilistic calibration condition (3), and noting that $x_t \leqslant F_t^{-1}(p)$ if and only if $p_t \leqslant p$. 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 Appendix A.

Theorem 2. Let $(F_t)_{t=1,2,\ldots}$ and $(G_t)_{t=1,2,\ldots}$ 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

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}(p_t \leqslant p) \rightarrow p \qquad \text{almost surely for all } p \tag{6}$$

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

We emphasize that condition (6) stands in its own right as a criterion for the validity of probabilistic forecasts, independently of our theoretical framework, in which it is interpreted as a sample version of condition (3). Indeed, following the lead of Dawid (1984) and Diebold *et al.* (1998), checks for the uniformity of the PIT values have formed a corner-stone 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 it with the CDF of the uniform distribution. This approach is adequate for small sample sizes and notable departures from uniformity, and its proponents include Staël von Holstein (1970), page 142, Seillier-Moiseiwitsch (1993), Hoeting (1994), page 33, Brocklehurst and Littlewood (1995), Frühwirth-Schnatter (1996), Raftery *et al.* (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. Fig. 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,

Table 3. Empirical coverage of central prediction intervals in the simulation study†

Forecaster	*Coverage (%) for the following intervals:*	
	50%	*90%*
Ideal	51.2	90.0
Climatological	51.3	90.7
Unfocused	50.1	90.1
Hamill's	50.9	89.5

†The nominal coverage is 50% and 90%.

run each of them forwards in time by 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, which was proposed independently by Anderson (1996), Hamill and Colucci (1997) and Talagrand *et al.* (1997) and has been extensively used since. To obtain a verification rank histogram, find the rank of the observation when pooled within the ordered 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 analogous 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), page 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), Candille and Talagrand (2005) and Corradi and Swanson (2006), 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 case-study in Section 4 provides an example in which $k=2$. The PITs for ideal k-step-ahead forecasts are at most $k-1$ dependent, and this assumption can be checked empirically, by plotting the sample autocorrelation functions 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,

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

with the empirical CDF of the observations,

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

Indeed, if we substitute the indicator function $\mathbf{1}(x_t\leqslant x)$ 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 $\hat{G}_T$. 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 Appendix A.

Theorem 3. Let $(F_t)_{t=1,2,\ldots}$ and $(G_t)_{t=1,2,\ldots}$ 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

$$\bar{F}(x) = \lim_{T\to\infty} \left\{ \frac{1}{T} \sum_{t=1}^{T} F_t(x) \right\}$$

exists for all $x \in \mathbb{R}$ and that the limit function is strictly increasing on $\mathbb{R}$. Then

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

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

We note that condition (9) stands in its own right as a criterion for the validity of probabilistic forecasts, independently of our theoretical framework. Still, the theoretical frame and theorems 2 and 3 provide reassurance, in that conditions (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 conditions (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 in Fig. 2(a), which shows the difference

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

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

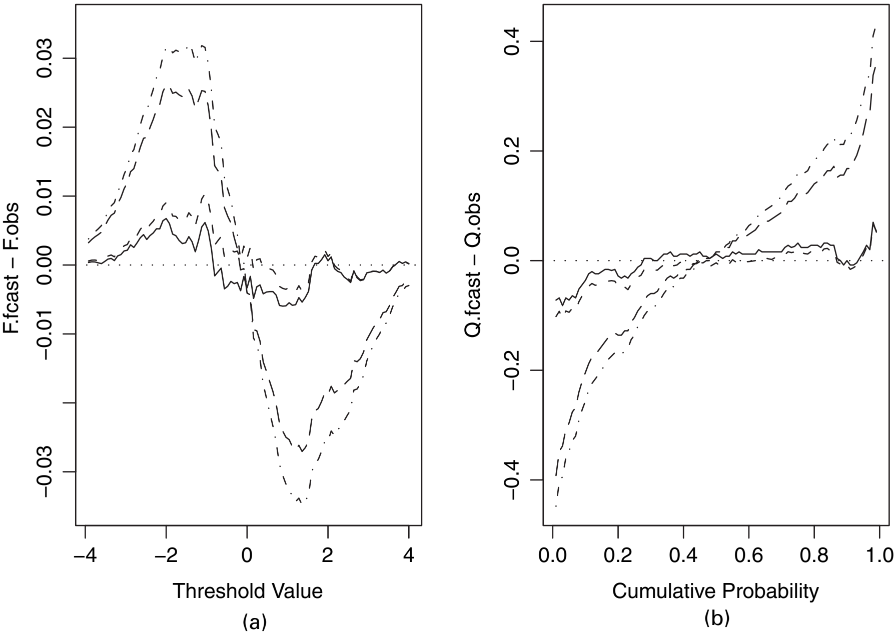


Fig. 2. Marginal calibration plot for the ideal forecaster (———), climatological forecaster (-------), unfocused forecaster (·-·-·-) and Hamill's forecaster (– – –): (a) CDFs; (b) quantiles

about 0 only, and this is indeed so for the ideal forecaster and the climatological forecaster. The unfocused forecaster and Hamill's forecaster lack marginal calibration, resulting in major excursions from 0. The same information can be visualized in terms of quantiles, as in Fig. 2(b), which shows the difference

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

of the quantile functions for $\bar{F}_T$ and $\hat{G}_T$. Under the hypothesis of marginal calibration, we again expect minor fluctuations about 0 only, and this is so 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 initial 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 box plots as a more instructive graphical device. We refer to this type of display as a sharpness diagram, and an example thereof is shown in Fig. 9 in Section 4.3.

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 that is 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 rule is proper if the expected value of the penalty $s(F,x)$ for an observation

Table 4. Average width of central prediction intervals in the simulation study†

Forecaster	*Average width for the following intervals:*	
	50%	*90%*
Ideal	1.35	3.29
Climatological	1.91	4.65
Unfocused	1.52	3.68
Hamill's	1.49	3.62

†The nominal coverage is 50% and 90%.

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 it lacks robustness (Selten, 1998; Gneiting and Raftery, 2006). The continuous ranked probability score is defined directly in terms of the predictive CDF F as

$$\mathrm{crps}(F, x) = \int_{-\infty}^{\infty} \{F(y) - \mathbf{1}(y \geqslant x)\}^2 \,\mathrm{d}y \tag{12}$$

and provides a more robust alternative. Gneiting and Raftery (2006) gave an alternative representation and showed that

$$\mathrm{crps}(F, x) = E_F|X - x| - \tfrac{1}{2}\, E_F|X - X'|, \tag{13}$$

where X and X' are independent copies of a random variable with CDF F and finite first moment. The representation (13) is particularly convenient when F is represented by a sample, possibly based on Markov chain Monte Carlo output or forecast ensembles (Gschlößl and Czado, 2005). Furthermore, the representation 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 on the basis of its average,

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

where

$$\mathrm{BS}(y) = \frac{1}{T} \sum_{t=1}^{T} \{F_t(y) - \mathbf{1}(x_t \leqslant y)\}^2$$

denotes the Brier (1950) score 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) discussed 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. Fig. 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 equation (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.

4. Case-study: probabilistic forecasts at the Stateline wind energy centre

Wind power is the fastest growing source of energy today. Estimates are that within the next 15 years wind energy will fill about 6% of the electricity supply in the USA. 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

Table 5. Average logarithmic score LogS and continuous ranked probability score CRPS in the simulation study

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

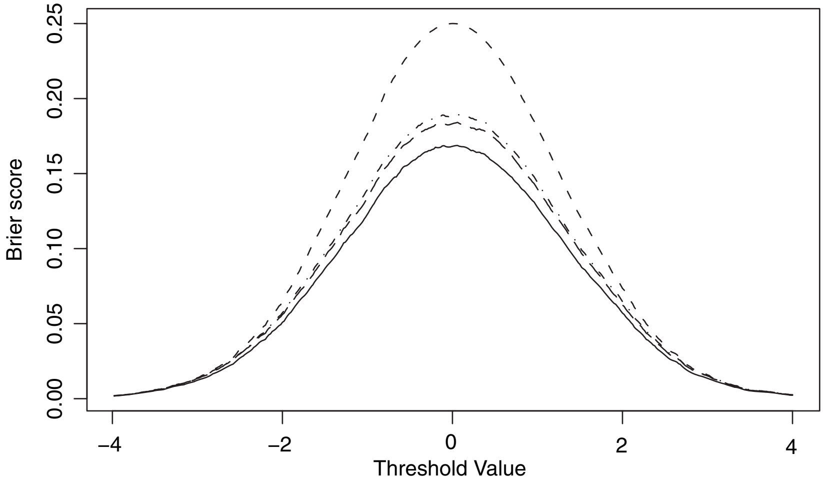


Fig. 3. Brier score plot for the ideal forecaster (———), climatological forecaster (-------), unfocused forecaster (·-·-·-·-) and Hamill's forecaster (– – –): the curves show the Brier score as a function of the threshold value; the area under each forecaster's curve equals the CRPS value (14)

forecast wind resources with any degree of accuracy. The development of advanced probabilistic forecast methodologies helps to 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 *et al.*, 1984). Gneiting *et al.* (2004) proposed a novel spatiotemporal 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 2-hour-ahead forecasts of hourly average wind speed at the Stateline wind energy centre. The evaluation period is May–November 2003, resulting in a total of 5136 probabilistic forecasts.

4.1. Predictive distributions for hourly average wind speed

We consider three competing statistical prediction algorithms for 2-hour-ahead probabilistic forecasts of hourly average wind speed w_t at the Stateline wind energy centre. Stateline is located on the Vansycle ridge at the border between the states of Oregon and Washington in the US Pacific Northwest. The data source is described in Gneiting *et al.* (2004).

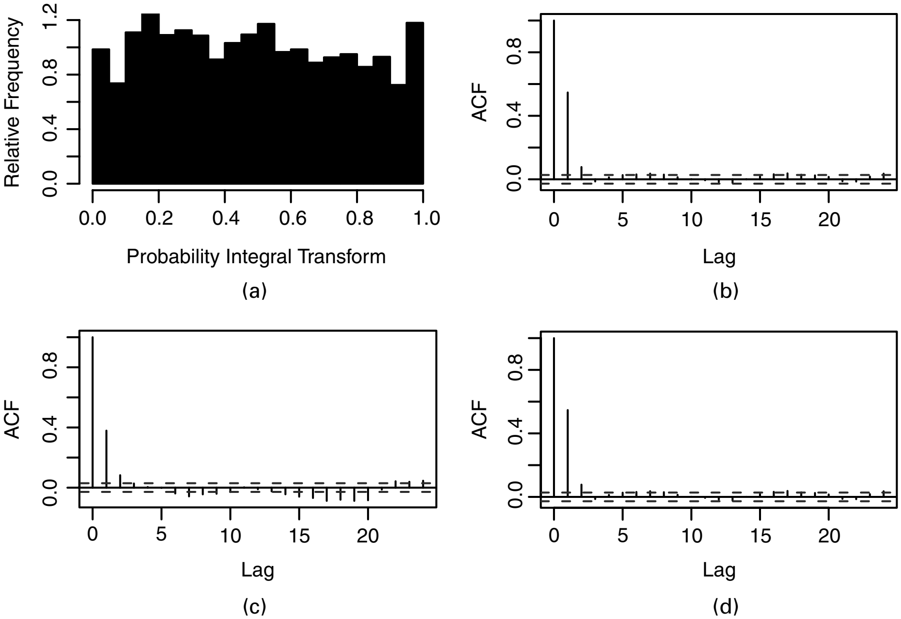


Fig. 4. PIT histogram and sample autocorrelation functions for the first three centred moments of the PIT values, for persistence forecasts of hourly average wind speed at the Stateline wind energy centre: (a) PIT histogram; (b) autocorrelation function of the PIT; (c) autocorrelation function of (PIT $-\frac{1}{2})^2$; (d) autocorrelation function of (PIT $-\frac{1}{2})^3$

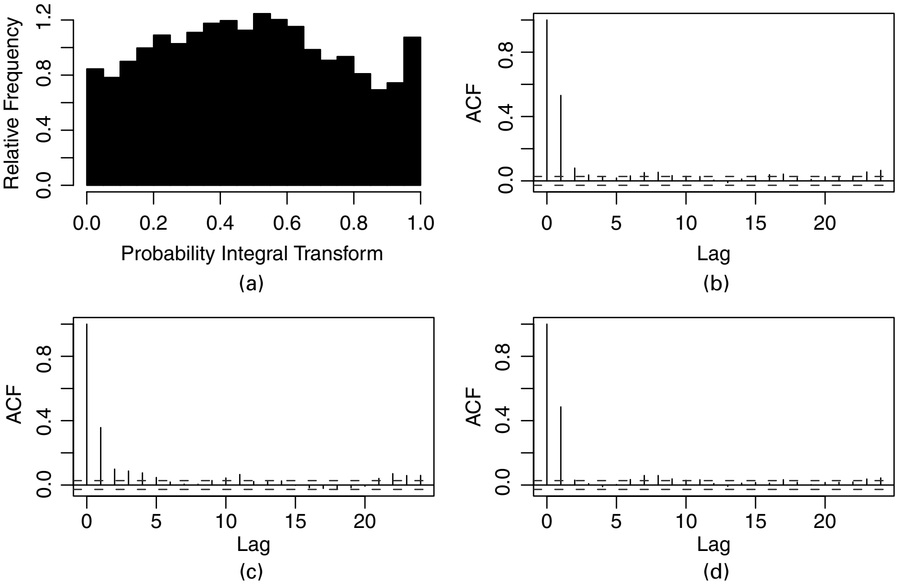
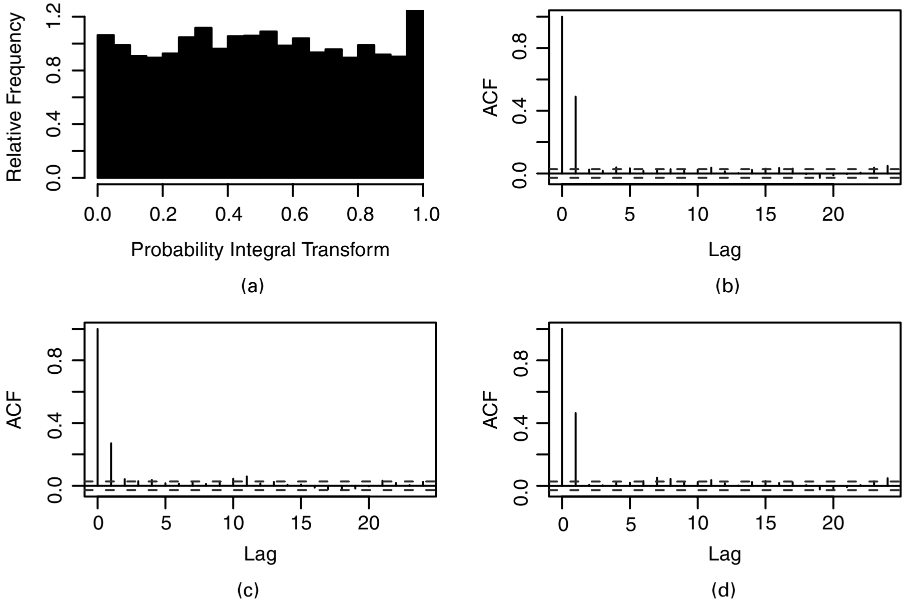


Fig. 5. PIT histogram and sample autocorrelation functions for the first three centred moments of the PIT values, for autoregressive forecasts of hourly average wind speed: (a) PIT histogram; (b) autocorrelation function of the PIT; (c) autocorrelation function of (PIT $-\frac{1}{2})^2$; (d) autocorrelation function of (PIT $-\frac{1}{2})^3$



(a) (b) (c) (d)

Fig. 6. PIT histogram and sample autocorrelation functions for the first three centred moments of the PIT values, for RST forecasts of hourly average wind speed: (a) PIT histogram; (b) autocorrelation function of the PIT; (c) autocorrelation function of $(\mathrm{PIT}-\frac{1}{2})^2$; (d) autocorrelation function of $(\mathrm{PIT}-\frac{1}{2})^3$

Table 6. Empirical coverage of central prediction intervals†

Forecast	*Empirical coverage for the following intervals:*	
	50%	*90%*
Persistence	50.9	89.2
Autoregressive	55.6	90.4
RST	51.2	88.4

†The nominal coverage is 50% and 90%.

The first method is the persistence forecast, a naïve yet surprisingly skilful, non-parametric 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 that was proposed by Roulston and Smith (2003). Specifically, the predictive CDF for w_{t+2} is the empirical distribution function of the set

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

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 to the residual component and find a Gaussian predictive distribution in the customary

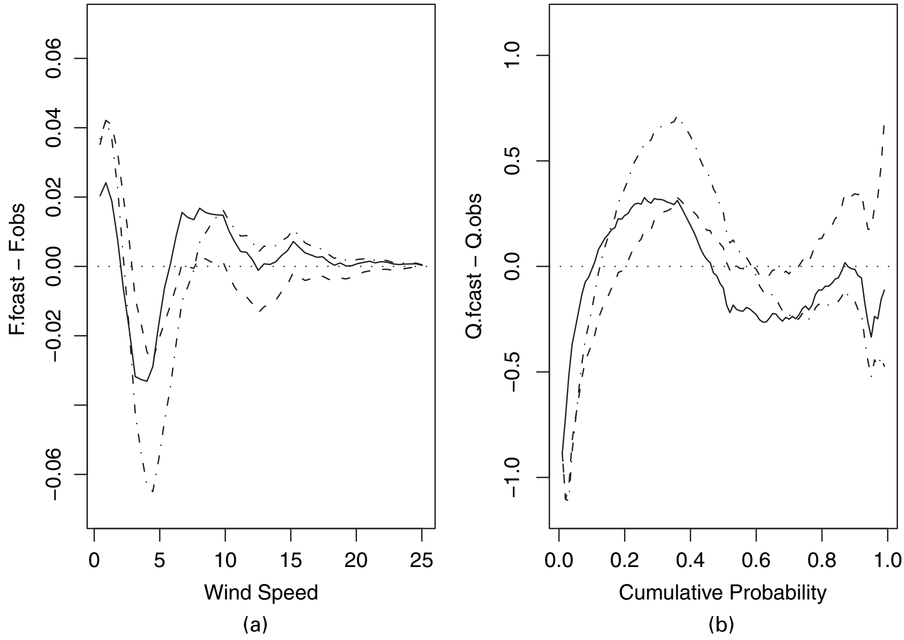


Fig. 7. Marginal calibration plot for persistence forecasts (-------), autoregressive forecasts (·-·-·-) and RST forecasts (———) of hourly average wind speed at the Stateline wind energy centre in terms of (a) CDFs and (b) quantiles, in metres per second

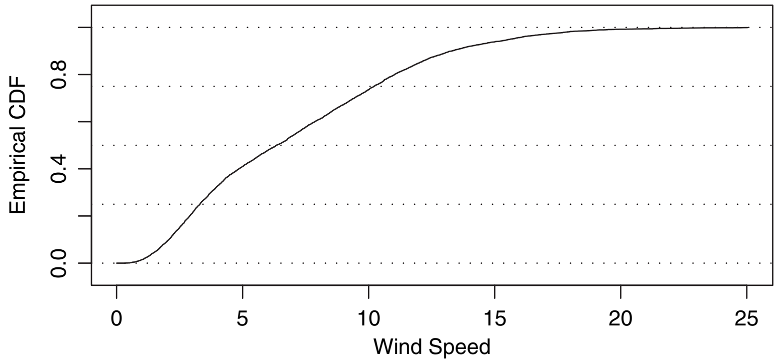


Fig. 8. Empirical CDF of hourly average wind speed at the Stateline wind energy centre in May–November 2003, in metres per second

way. The predictive distribution assigns a typically small positive mass to the negative half-axis, and, in view of the non-negativity of the predictand, we redistribute this mass to wind speed 0. 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 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 off-site 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.

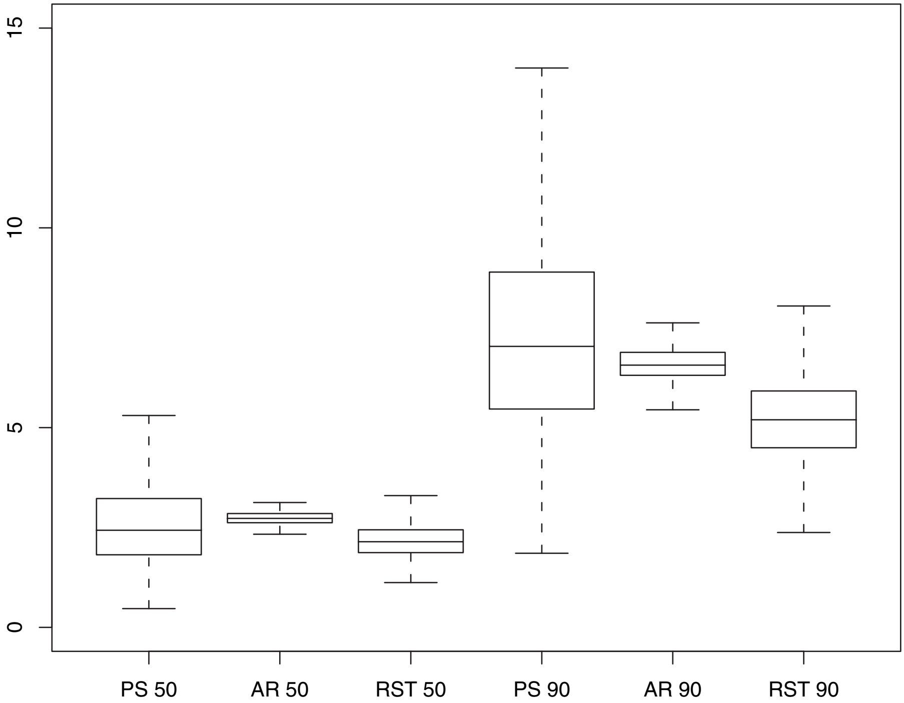


Fig. 9. Sharpness diagram for persistence (PS), autoregressive (AR) and RST forecasts of hourly average wind speed at the Stateline wind energy centre: the box plots show the fifth, 25th, 50th, 75th and 95th percentiles of the width of the central prediction interval, in metres per second; the smaller the width, the sharper (the nominal coverage is 50% (left) and 90% (right))

Table 7. Average width of central prediction intervals†

Forecast	*Average widths ($m\ s^{-1}$) for the following intervals:*	
	50%	*90%*
Persistence	2.63	7.51
Autoregressive	2.74	6.55
RST	2.20	5.31

†The nominal coverage is 50% and 90%.

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 that are reported below and in Gneiting *et al.* (2004) stem from the use of R rather than S-PLUS and differences in optimization routines.

4.2. Assessing calibration

Figs 4–6 show the PIT histograms for the three forecast techniques, along with the sample autocorrelation functions for the first three centred moments of the PIT values and the respective Bartlett confidence intervals. The PIT histograms for the persistence and RST forecasts appear

Table 8. CRPS value (14) for probabilistic forecasts of hourly average wind speed at the Stateline wind energy centre in March–November 2003, month by month and for the entire evaluation period

Forecast	*CRPS* $(m\ s^{-1})$ *for the following months:*							
	May	*June*	*July*	*August*	*September*	*October*	*November*	*March–November*
Persistence	1.16	1.08	1.29	1.21	1.20	1.29	1.16	1.20
Autoregressive	1.12	1.02	1.10	1.11	1.11	1.22	1.13	1.12
RST	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 centre in March–November 2003, month by month and for the entire evaluation period

Forecast	*MAE* $(m\ s^{-1})$ *for the following months:*							
	May	*June*	*July*	*August*	*September*	*October*	*November*	*March–November*
Persistence	1.60	1.45	1.74	1.68	1.59	1.68	1.51	1.61
Autoregressive	1.53	1.38	1.50	1.54	1.53	1.68	1.54	1.53
RST	1.32	1.18	1.33	1.31	1.36	1.48	1.37	1.34

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 for the aforementioned evaluation period.

The PIT values for ideal two-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 non-negligible at lag 2, and the centred second moment showed notable negative correlations at lags between 15 and 20 h. These features indicate a lack of fit of the predictive model, even though they seem difficult to interpret diagnostically. The respective sample autocorrelations for the autoregressive forecast were positive and non-negligible at lags up to 5 h, 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.

Fig. 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 to non-negligible excursions from 0, particularly at small wind speeds and for the autoregressive forecast. The lack of predictive model fit finds an explanation in Fig. 8, which shows the empirical CDF $\bar{F}_T$ of hourly average wind speed. Hourly average wind speeds less than 1 m s^{-1} were almost never observed, even though the predictive distributions assign positive point mass to wind speed 0.

4.3. Assessing sharpness

The sharpness diagram in Fig. 9 shows box plots that illustrate the width of central prediction intervals for the 5136 predictive distributions in the evaluation period, May–November 2003.

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. Under the null hypothesis of equal predictive performance this happens with probability $\left(\frac{1}{2}\right)^7 = 1/128$ only. Similarly, the autoregressive technique outperformed the persistence method in May–October, but not in November. Various other tests can be employed, but care needs to be taken to avoid dependences 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. Fig. 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.

We noted in Section 3.4 that the continuous ranked probability score generalizes the absolute error and reduces to it 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 units as the wind speed observations, i.e. in metres per second, and can be directly compared. The insights that the monthly scores provide are indicative of the potential benefits of thoughtful stratification.

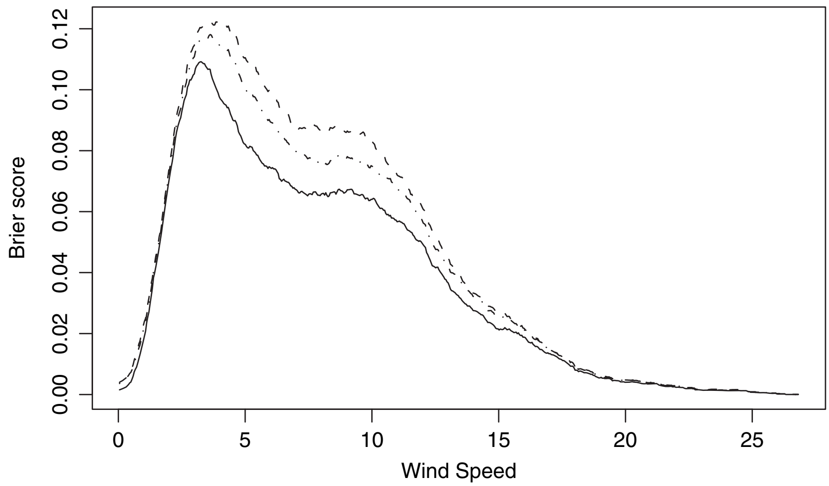


Fig. 10. Brier score plot for persistence forecasts (-------), autoregressive forecast (·-·-·-) and RST forecasts (———) of hourly average wind speed at the Stateline wind energy centre, in metres per second: the graphs show the Brier score as a function of the threshold value; the area under each forecast's curve equals the CRPS value (14)

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 Fig. 7 suggest a modified version of the RST technique that uses truncated normal rather than 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 PIT. Here we have 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 benefited 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 by using the diagnostic devices that were described by Murphy *et al.* (1989) and Murphy and Winkler (1992), among others.

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, Garratt *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 based 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 flavour of our diagnostic approach, calibration and sharpness are properties that are relevant to Bayesian forecasters as well. Rubin (1984), pages 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 that are 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 non-parametric, 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 can often be applied fruitfully (Dawid (1984), page 288, and Gneiting and Raftery (2006)).

Acknowledgements

We thank John B. Bremnes, Barbara G. Brown, Richard E. Chandler, Claudia Czado, 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 a reviewer for a wealth of helpful suggestions. The data that were used in Section 4 were obtained from Oregon State University Energy Resources Research Laboratory under the direction of Stel N. Walker. Our work has been supported by the Department of Defense 'Multidisciplinary university research initiative' programme administered by the Office of Naval Research under grant N00014-01-10745. Tilmann Gneiting furthermore acknowledges support by the National Science Foundation under award 0134264 and by the Washington Technology Center, and thanks the Soil Physics group at Universität Bayreuth, Germany, where part of this work was performed.

Appendix A

A.1. Proof of theorem 1

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

$$\begin{aligned}\mathrm{var}(F) &= \mathrm{var}\{F^{-1}(U)\}\\ &= E[\mathrm{var}\{F^{-1}(U)|z_1, \ldots, z_T\}] + \mathrm{var}[E\{F^{-1}(U)|z_1, \ldots, z_T\}].\end{aligned}$$

The first term in the decomposition equals

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

and the second term is non-negative and vanishes if and only if $E(G_1) = \ldots = E(G_T)$.

A.2. Proof of theorem 2

For $p \in (0,1)$ and $t = 1, 2, \ldots$, 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),

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

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

A.3. Proof of theorem 3

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

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

By theorem 2 with $F_t = \bar{F}$ for $t = 1, 2, \ldots$, we have that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}(q_t \leqslant q) \to q \qquad \text{almost surely}$$

if and only if

$$\frac{1}{T}\sum_{t=1}^{T} G_t \circ \bar{F}^{-1}(q) \to q \qquad \text{almost surely.}$$

Hence, marginal calibration is equivalent to condition (9).

References

Anderson, J. L. (1996) A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Clim.*, **9**, 1518–1530.

Bauwens, L., Giot, P., Grammig, J. and Veredas, D. (2004) A comparison of financial duration models via density forecasts. *Int. J. Forecast.*, **20**, 589–609.

Berkowitz, J. (2001) Testing density forecasts, with applications to risk management. *J. Bus. Econ. Statist.*, **19**, 465–474.

Bernardo, J. M. (1979) Expected information as expected utility. *Ann. Statist.*, **7**, 686–690.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995) Bayesian computing and stochastic systems (with discussion). *Statist. Sci.*, **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. *Z. Wahrsch. Ver. Geb.*, **2**, 1–11.

Boero, G. and Marrocu, E. (2004) The performance of SETAR models: a regime conditional evaluation of point, interval and density forecasts. *Int. J. Forecast.*, **20**, 305–320.

Bremnes, J. B. (2004) Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mnthly Weath Rev.*, **132**, 338–347.

Brier, G. W. (1950) Verification of forecasts expressed in terms of probability. *Mnthly Weath. Rev.*, **78**, 1–3.

Brocklehurst, S. and Littlewood, B. (1995) Techniques for prediction analysis and recalibration. In *Handbook of Software Reliability Engineering* (ed. M. R. Lyu). New York: McGraw-Hill.

Brown, B. G., Katz, R. W. and Murphy, A. H. (1984) Time series models to simulate and forecast wind speed and wind power. *J. Clim. Appl. Meteorol.*, **23**, 1184–1195.

Candille, G. and Talagrand, O. (2005) Evaluation of probabilistic prediction systems for a scalar variable. *Q. J. R. Meteorol. Soc.*, **131**, 2131–2150.

Christoffersen, P. F. (1998) Evaluating interval forecasts. *Int. Econ. Rev.*, **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. *J. Forecast.*, **19**, 255–276.

Corradi, V. and Swanson, N. R. (2006) Predictive density evaluation. In *Handbook of Economic Forecasting*, vol. 1 (eds G. Elliott, C. W. J. Granger and A. Timmermann), pp. 197–284. Amsterdam: Elsevier.

Dawid, A. P. (1982) The well-calibrated Bayesian. *J. Am. Statist. Ass.*, **77**, 605–610.

Dawid, A. P. (1984) Statistical theory: the prequential approach (with discussion). *J. R. Statist. Soc.* A, **147**, 278–292.

Dawid, A. P. (1985a) The impossibility of inductive inference. *J. Am. Statist. Ass.*, **80**, 340–341.

Dawid, A. P. (1985b) Calibration-based empirical probability (with discussion). *Ann. Statist.*, **13**, 1251–1285.

Dawid, A. P. (1986) Probability forecasting. In *Encyclopedia of Statistical Sciences*, vol. 7 (eds S. Kotz, N. L. Johnson and C. B. Read), 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*, vol. 1 (eds S. S. Gupta and J. O. Berger), pp. 291–314. New York: Academic Press.

DeGroot, M. H. and Fienberg, S. E. (1983) The comparison and evaluation of forecasters. *Statistician*, **12**, 12–22.

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998) Evaluating density forecasts with applications to financial risk management. *Int. Econ. Rev.*, **39**, 863–883.

Diebold, F. X. and Mariano, R. S. (1995) Comparing predictive accuracy. *J. Bus. Econ. Statist.*, **13**, 253–263.

Duffie, D. and Pan, J. (1997) An overview of value at risk. *J. Deriv.*, **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. *Environ. Ecol. Statist.*, **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. *J. Am. Statist. Ass.*, **98**, 829–838.

Gelman, A., Meng, X.-L. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sin.*, **6**, 733–807.

Gerds, T. (2002) Nonparametric efficient estimation of prediction error for incomplete data models. *PhD Thesis*. Mathematische Fakultät, Albert-Ludwigs-Universität Freiburg, Freiburg.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G. and Aldrich, E. (2004) Calibrated probabilistic forecasting at the Stateline wind energy centre: the regime-switching space-time (RST) method. *Technical Report 464*. Department of Statistics, University of Washington, Seattle.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G. and Aldrich, E. (2006) Calibrated probabilistic forecasting at the Stateline wind energy centre: the regime-switching space-time (RST) method. *J. Am. Statist. Ass.*, **101**, 968–979.

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

Gneiting, T. and Raftery, A. E. (2006) Strictly proper scoring rules, prediction and estimation. *J. Am. Statist. Ass.*, to be published.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mnthly Weath. Rev.*, **133**, 1098–1198.

Good, I. J. (1952) Rational decisions. *J. R. Statist. Soc.* B, **14**, 107–114.

Granger, C. W. J. (2006) Some thoughts on the future of forecasting. *Oxf. Bull. Econ. Statist.*, **67S**, 707–711.

Gschlößl, S. and Czado, C. (2005) Spatial modelling of claim frequency and claim size in insurance. *Discussion Paper 461*. Sonderforschungsbereich 386, Ludwig-Maximilians-Universität München, Munich.

Hamill, T. M. (2001) Interpretation of rank histograms for verifying ensemble forecasts. *Mnthly Weath. Rev.*, **129**, 550–560.

Hamill, T. M. and Colucci, S. J. (1997) Verification of Eta-RSM short-range ensemble forecasts. *Mnthly Weath. Rev.*, **125**, 1312–1327.

Hoeting, J. (1994) Accounting for model uncertainty in linear regression. *PhD Thesis*. Department of Statistics, University of Washington, Seattle.

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

Krzysztofowicz, R. (1999) Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Wat. Resour. Res.*, **35**, 2739–2750.

Krzysztofowicz, R. and Sigrest, A. A. (1999) Calibration of probabilistic quantitative precipitation forecasts. *Weath. Forecast.*, **14**, 427–442.

Moyeed, R. A. and Papritz, A. (2002) An empirical comparison of kriging methods for nonlinear spatial point prediction. *Math. Geol.*, **34**, 365–386.

Murphy, A. H. (1972) Scalar and vector partitions of the probability score: Part I, Two-state situation. *J. Appl. Meteorol.*, **11**, 273–278.

Murphy, A. H., Brown, B. G. and Chen, Y.-S. (1989) Diagnostic verification of temperature forecasts. *Weath. Forecast.*, **4**, 485–501.

Murphy, A. H. and Winkler, R. L. (1987) A general framework for forecast verification. *Mnthly Weath. Rev.*, **115**, 1330–1338.

Murphy, A. H. and Winkler, R. L. (1992) Diagnostic verification of probability forecasts. *Int. J. Forecast.*, **7**, 435–455.

Noceti, P., Smith, J. and Hodges, S. (2003) An evaluation of tests of distributional forecasts. *J. Forecast.*, **22**, 447–455.

Oakes, D. (1985) Self-calibrating priors do not exist. *J. Am. Statist. Ass.*, **80**, 339.

Palmer, T. N. (2002) The economic value of ensemble forecasts as a tool for risk assessment: from days to decades. *Q. J. R. Meteorol. Soc.*, **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., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Mnthly Weath. Rev.*, **133**, 1155–1174.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997) Bayesian model averaging for linear regression models. *J. Am. Statist. Ass.*, **92**, 179–191.

Rosenblatt, M. (1952) Remarks on a multivariate transformation. *Ann. Math. Statist.*, **23**, 470–472.

Roulston, M. S. and Smith, L. A. (2002) Evaluating probabilistic forecasts using information theory. *Mnthly Weath. Rev.*, **130**, 1653–1660.

Roulston, M. S. and Smith, L. A. (2003) Combining dynamical and statistical ensembles. *Tellus* A, **55**, 16–30.

Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151–1172.

Sandroni, A., Smorodinsky, R. and Vohra, R. V. (2003) Calibration with many checking rules. *Math. Oper. Res.*, **28**, 141–153.

Schervish, M. J. (1985) Comment. *J. Am. Statist. Ass.*, **80**, 341–342.

Schervish, M. J. (1989) A general method for comparing probability assessors. *Ann. Statist.*, **17**, 1856–1879.

Schumacher, M., Graf, E. and Gerds, T. (2003) How to assess prognostic models for survival data: a case study in oncology. *Meth. Inform. Med.*, **42**, 564–571.

Seillier-Moiseiwitsch, F. (1993) Sequential probability forecasts and the probability integral transform. *Int. Statist. Rev.*, **61**, 395–408.

Selten, R. (1998) Axiomatic characterization of the quadratic scoring rule. *Exptl Econ.*, **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. *J. Forecast.*, **4**, 283–291.

Staël von Holstein, C.-A. S. (1970) *Assessment and Evaluation of Subjective Probability Distributions*. Stockholm: Economics Research Institute.

Talagrand, O., Vautard, R. and Strauss, B. (1997) Evaluation of probabilistic prediction systems. In *Proc. Wrkshp Predictability*, pp. 1–25. Reading: European Centre for Medium-Range Weather Forecasts.

Vovk, V. and Shafer, G. (2005) Good randomized sequential probability forecasting is always possible. *J. R. Statist. Soc.* B, **67**, 747–763.

Wallis, K. F. (2003) Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts. *Int. J. Forecast.*, **19**, 165–175.

Weigend, A. S. and Shi, S. (2000) Predicting daily probability distributions of S&P500 returns. *J. Forecast.*, **19**, 375–392.

Winkler, R. L. (1977) Rewarding expertise in probability assessment. In *Decision Making and Change in Human Affairs* (eds H. Jungermann and G. de Zeeuw), pp. 127–140. Dordrecht: Reidel.