Probabilistic Forecasting

Tilmann Gneiting[1] and Matthias Katzfuss[2]

[1]Institut für Angewandte Mathematik, Universität Heidelberg, 69120 Heidelberg, Germany; email: t.gneiting@uni-heidelberg.de

[2]Department of Statistics, Texas A&M University, College Station, TX 77843; email: katzfuss@gmail.com

Annu. Rev. Stat. Appl. 2014. 1:125–51

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-062713-085831

Keywords

calibration, consistent scoring function, ensemble forecast, proper scoring rule, distributional regression

Abstract

A probabilistic forecast takes the form of a predictive probability distribution over future quantities or events of interest. Probabilistic forecasting aims to maximize the sharpness of the predictive distributions, subject to calibration, on the basis of the available information set. We formalize and study notions of calibration in a prediction space setting. In practice, probabilistic calibration can be checked by examining probability integral transform (PIT) histograms. Proper scoring rules such as the logarithmic score and the continuous ranked probability score serve to assess calibration and sharpness simultaneously. As a special case, consistent scoring functions provide decision-theoretically coherent tools for evaluating point forecasts. We emphasize methodological links to parametric and nonparametric distributional regression techniques, which attempt to model and to estimate conditional distribution functions; we use the context of statistically postprocessed ensemble forecasts in numerical weather prediction as an example. Throughout, we illustrate concepts and methodologies in data examples.

1. INTRODUCTION

1.1. Probabilistic Forecasting—A New Paradigm

Probabilistic forecast: a forecast in the form of a probability distribution over future quantities or events

A common desire of all humankind is to make predictions for an uncertain future. Clearly then, forecasts should be probabilistic, i.e., they should take the form of probability distributions over future quantities or events. Taking a historical perspective, Stigler (1975) describes the transition from point estimation to distribution estimation in the nineteenth century. Today, we are witnessing a paradigm shift, shown by a transdisciplinary transition from single-valued or point forecasts to distributional or probabilistic forecasts (Gneiting 2008). In a nutshell, probabilistic forecasts serve to quantify the uncertainty in a prediction, and they are an essential ingredient of optimal decision making.

Although probability forecasts for binary events (e.g., an 80% chance of rain today, a 10% chance of a financial meltdown by the end of the year) have been commonly issued for the past several decades (Gigerenzer et al. 2005), attention has been shifting toward probabilistic forecasts for more general types of variables and events. Critical problems of science and society have been driving this development; these problems include weather and climate prediction (Collins & Knight 2007; Gneiting & Raftery 2005; Palmer 2002, 2012), flood risk assessment (Cloke & Pappenberger 2009; Krzysztofowicz 2001), seismic hazard prediction (Jordan et al. 2011; see sidebar, The L'Aquila Earthquake Trial), predictions about the availability of renewable energy resources (Pinson 2013, Zhu & Genton 2012), economic and financial risk management (Groen et al. 2013, Timmermann 2000), election outcome prediction (Montgomery et al. 2012), demographic and epidemiological projection (Alkema et al. 2007, Raftery et al. 2012), health care management (Jones & Spiegelhalter 2012), and predictive and preventative medicine (Hood et al. 2004). The need for advancement in the theory, methodology, and application of probabilistic forecasting is pronounced, and challenges and opportunities for statistical scientists to become involved and contribute abound.

To give a prominent example, the Bank of England's Monetary Policy Committee has been issuing probabilistic forecasts of inflation rates in the form of two-piece normal distributions for nearly two decades (Bank of England 2013, Wallis 2003). **Figure 1** shows the February 2013 projection of the future UK consumer price index. The fan chart displays the predictive distributions in terms of annual percentage change. The central area depicts a pointwise 10% prediction interval, and the progressively lighter-shaded bands extend this interval by 10% each; the entire fan provides a 90% interval forecast. Following the Bank of England's lead, central banks worldwide have embraced the concept of probabilistic forecasting, including the monetary

THE L'AQUILA EARTHQUAKE TRIAL

In October 2012, an Italian court sentenced six leading scientists and a government official to six years of prison each for providing "incomplete, imprecise and contradictory information" (Hall 2011, p. 266) on the probability and risk of a major seismic event prior to the devastating earthquake that hit the city of L'Aquila on April 6, 2009. Condemned by the scientific community worldwide (Nat. Publ. Group 2012), the L'Aquila verdict serves to highlight the challenges in communicating forecast uncertainty in low-probability high-risk environments. Although there is consensus among the seismological community that "probabilistic forecasts are the best means for transmitting scientific information about future earthquake occurrence to decision makers" (Jordan et al. 2011, p. 348), the implied probabilities for major seismic events at lead times of days to weeks rarely exceed a few percent, thereby imposing a major impediment to civil protection efforts (van Stiphout et al. 2010, Jordan 2013).

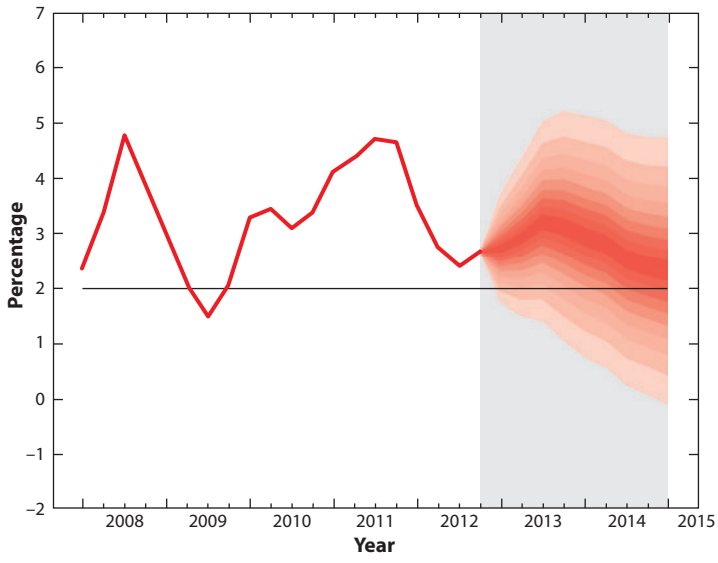

Figure 1

February 2013 Bank of England forecast of inflation in the United Kingdom as a percentage increase in the consumer price index (Bank of England 2013, with permission). The shaded bands in the fan chart show prediction intervals in increments of 10%.

authorities of Australia, Brazil, Canada, Norway, the Philippines, South Africa, Thailand, and Turkey (Hammond 2012). Typically, the forecasts derive from suites of econometric time series models, such as dynamic stochastic general equilibrium approaches.

1.2. Article Overview

Our aim in this review is to give a selective overview of the state of the art in probabilistic forecasting, covering theory, methodology, and a range of applications and focusing on predictions of real-valued quantities, such as inflation rate, temperature, or precipitation accumulation. Throughout, we illustrate concepts and methodologies using a case study on short-term probabilistic forecasts of wind speed at the Stateline wind energy center in the US Pacific Northwest. This case study is described in Section 1.3.

Section 2 reviews theoretical foundations in the setting of a prediction space, i.e., a probability space tailored to the study of distributional forecasts. Probabilistic forecasting has the general goal of maximizing the sharpness of the predictive distributions, subject to calibration. Briefly, calibration concerns the statistical compatibility between the probabilistic forecasts and the realizations. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only.

The issues associated with the generation and evaluation of forecasts are intimately related. This relation motivates our treatment of proper scoring rules, consistent scoring functions, and elicitable functionals in Section 3. A scoring rule assigns a numerical score to a probabilistic forecast based on the predictive distribution and realization. A proper scoring rule is designed such that truth telling (i.e., quoting the true distribution as the forecast distribution) is an optimal strategy in expectation. Similarly, a scoring function assigns a numerical score to a single-valued point forecast. A consistent scoring function is a special case of a proper scoring rule that depends on

Sharpness: the concentration of the predictive distributions in absolute terms; a property exclusive to the forecasts

Calibration: statistical compatibility of probabilistic forecasts and observations; essentially, realizations should be indistinguishable from random draws from predictive distributions

the predictive distribution via a target functional only, such as the mean, the median, or a quantile. We connect this notion with a predictive view of regression, arguing for distributional regression.

Forecasting is rarely a purely statistical exercise. Rather, forecasters must draw on subject matter expertise to issue predictive distributions that condition on judiciously chosen information (Holzmann & Eulert 2013). To illustrate the progress in what is arguably the most advanced application, our review closes with a succinct discussion of the state of the art of probabilistic weather forecasting in Section 4, which showcases the fruitful interplay between analytic-numerical modeling and statistical modeling. A pressing need is to go beyond the univariate, real-valued case, which we review, to the multivariate case, such as in temporal, spatial, and spatiotemporal scenarios and trajectories. Moreover, examinations of the multivariate case should include the development of and quest for forecast evaluation and visualization techniques (Spiegelhalter et al. 2011).

1.3. Example: Probabilistic Forecasts at the Stateline Wind Energy Center

Wind power is the fastest-growing energy source today. Owing to its global proliferation, methods for short-term forecasts of wind resources at wind energy sites are in vigorous development (Jeon & Taylor 2012; Pinson 2012a,b, 2013; Traiteur et al. 2012; Zhu & Genton 2012). Throughout this review, we illustrate methods and concepts using probabilistic forecasts of wind speed at the Stateline wind energy center in the US Pacific Northwest (Gneiting et al. 2006). Specifically, we consider two-hour-ahead forecasts of the hourly average wind speed in May to November 2003 at Vansycle, Oregon, in the immediate vicinity of the Stateline wind energy center. The information set consists of current and past observations of wind speed and wind direction at Vansycle and at two off-site locations, Kennewick and Goodnoe Hills in southern Washington.

RST: regime-switching space-time

In this context, Gneiting et al. (2006) introduced the regime-switching space-time (RST) technique, which conditions on the forecast regime and on both on-site and off-site information, to provide probabilistic forecasts in the form of truncated normal predictive distributions. The RST approach distinguishes westerly and easterly forecast regimes, uses regime-dependent parameters, and estimates the predictive model on a rolling 45-day training period. **Figure 2** illustrates the two-hour-ahead RST forecasts of hourly average wind speed at Vansycle for the 7-day period beginning July 5, 2003.

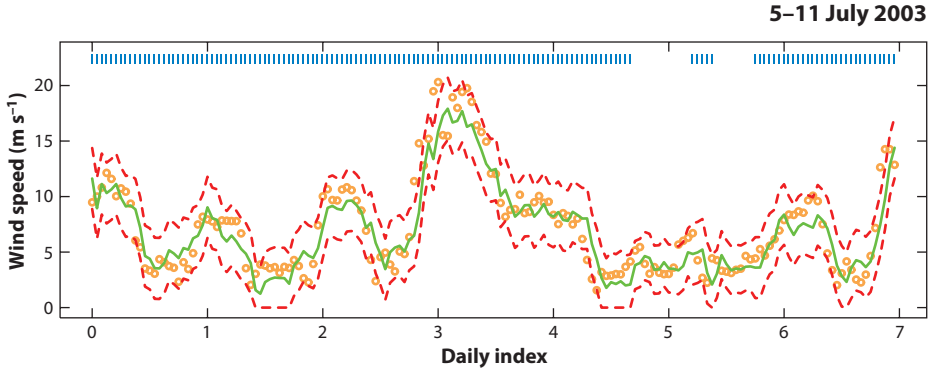

Figure 2

Two-hour-ahead regime-switching space-time forecasts of hourly average wind speed at the Stateline wind energy center for the 7-day period beginning July 5, 2003. The predictive mean is shown in green, the 90% central prediction interval in red, and the realized wind speed in orange. The blue marks at the top indicate forecasts in the prevalent westerly regime.

Throughout this review, we compare the RST forecasts to three reference forecasts. The simplest point forecast is the persistence or no-change (NC) forecast, namely, the most recent observed value. We also consider the Gaussian autoregressive (AR) time series model proposed by Brown et al. (1984), which results in forecasts in the form of a normal distribution. Finally, we consider the simple similarity-based (SB) method described in Section 3.4.

NC: no-change or persistence

AR: autoregressive

SB: similarity-based

2. PREDICTION SPACES, CALIBRATION, AND SHARPNESS

In an important paper, Murphy & Winkler (1987) called for the consideration of the joint distribution of the forecast and the observation. In contrast to their work, which focused on the setting of point forecasts, we follow Gneiting & Ranjan (2013) and introduce the key tool of a prediction space. We review the notions of calibration, dispersion, and sharpness, and we argue that probabilistic forecasting should aim to maximize the sharpness of the predictive distributions, subject to calibration (Gneiting et al. 2007, Murphy & Winkler 1987). Calibration concerns the statistical compatibility between the probabilistic forecasts and the realizations; essentially, the observations should be indistinguishable from random draws from the predictive distributions. Sharpness refers to the concentration of the predictive distributions and thus is a property of the forecasts only. The sharper, the better, provided the predictive distributions are calibrated.

2.1. Prediction Spaces

A prediction space is a probability space tailored to the study of distributional forecasts (Gneiting & Ranjan 2013). Here we focus on the case of a real-valued observation, Y, for which a probabilistic forecast, F, can be identified with the associated cumulative distribution function (CDF) on the real line, $\mathbb{R}$. The prediction space setting considers the joint distribution of the probabilistic forecasts and the observations. In the simplest case, the elements of the sample space can be identified with tuples of the form (F, Y), where the probabilistic forecast F is a CDF-valued random quantity that utilizes a certain information basis or information set $\mathcal{A}$, which comprises the training data, expertise, theories, and assumptions at hand. For readers familiar with measure theory, the information set $\mathcal{A}$ can be viewed as a sigma field; then, F is a CDF-valued random quantity that is measurable with respect to $\mathcal{A}$ (i.e., F is based on only accessible and permitted information). Here and throughout this article we use the symbol $\mathcal{L}$ generically to denote an unconditional or conditional distribution.

CDF: cumulative distribution function

> **Definition 1:** The CDF-valued random quantity F is ideal relative to the information set encoded by $\mathcal{A}$ if $F = \mathcal{L}(Y \mid \mathcal{A})$.

Thus, an ideal probabilistic forecast makes the best possible use of the information at hand. For example, suppose that $Y|\mu \sim \mathcal{N}(\mu, 1)$ and $\mu \sim \mathcal{N}(0, 1)$. Then the probabilistic forecast $F = \mathcal{N}(\mu, 1) = \mathcal{L}(Y|\mu)$ is ideal relative to the information set generated by the random variable μ. The forecast $F = \mathcal{N}(0, 2)$ is ideal relative to the trivial information set. In general, a prediction space specifies the joint distribution of tuples of the form $(F_1, \ldots, F_n, Y)$, where the probabilistic forecasts $F_1, \ldots, F_n$ are CDF-valued random quantities. All subsequent definitions and theoretical results pertain to this setting.

2.2. Calibration and Dispersion

Using the so-called probability integral transform, this section introduces the concepts of calibration and dispersion, which concern the statistical compatibility between probabilistic forecasts and the corresponding realizations.

PIT: probability integral transform

2.2.1. Probability integral transform. If F denotes a fixed, nonrandom predictive CDF for an observation Y, the probability integral transform (PIT) is the random variable $Z_F = F(Y)$. If F is continuous and $Y \sim F$, then Z_F is standard uniform. Under a more general, randomized version of the PIT, the uniformity result holds under arbitrary, but not necessarily continuous nonrandom CDFs (Czado et al. 2009, Rüschendorf 2009). In the prediction space setting, we work with the further extension given in Definition 2, which allows F to be a CDF-valued random quantity (Gneiting & Ranjan 2013).

Definition 2: Let V be a standard uniformly distributed variable that is independent of the CDF-valued random quantity F and the observation Y. For $y \in \mathbb{R}$, let $F(y-) = \lim F_{x \uparrow y}(x)$. Then

$$Z_F = F(Y-) + V\,(F(Y) - F(Y-))$$

is the PIT of the probabilistic forecast F.

In a nutshell, the PIT is the value that the predictive CDF attains at the observation, with suitable adaptions at any points of discontinuity.

Dispersion: the concentration of the predictive distributions relative to the observations; a joint property of forecasts and observations

2.2.2. Notions of calibration and dispersion. As we review the notions of calibration and dispersion defined by Gneiting & Ranjan (2013), we use the terms CDF-valued random quantity and forecast interchangeably.

Definition 3: Let F and G be CDF-valued random quantities with PITs Z_F and Z_G, respectively.

(*a*) The forecast F is marginally calibrated if $\mathbb{E}[F(y)] = \mathbb{P}(Y \leq y)$ for all $y \in \mathbb{R}$.
(*b*) The forecast F is probabilistically calibrated if its PIT Z_F has a standard uniform distribution.
(*c*) The forecast F is overdispersed if $\text{var}(Z_F) < \frac{1}{12}$ and underdispersed if $\text{var}(Z_F) > \frac{1}{12}$.
(*d*) The forecast F is more dispersed than the forecast G if $\text{var}(Z_F) < \text{var}(Z_G)$.

Calibration and dispersion thus concern facets of the joint law of the probabilistic forecast and the observation. If F is probabilistically calibrated, then $\text{var}(Z_F) = \frac{1}{12}$ and F is well dispersed. The following result highlights the role of ideal forecasts (Gneiting & Ranjan 2013).

Theorem 1: A forecast that is ideal relative to some information set is both marginally calibrated and probabilistically calibrated.

Let us revisit the above example where $Y \,|\, \mu \sim \mathcal{N}(\mu, 1)$ and $\mu \sim \mathcal{N}(0, 1)$. The forecasts $F = \mathcal{N}(\mu, 1)$ and $F = \mathcal{N}(0, 2)$ are ideal, so they are both probabilistically and marginally calibrated. The forecast $F = \mathcal{N}(0, \sigma^2)$ is underdispersed if $\sigma^2 < 2$ and overdispersed if $\sigma^2 > 2$. In contrast, some forecasts are calibrated either probabilistically or marginally, but not both (Gneiting et al. 2007, Gneiting & Ranjan 2013).

2.2.3. Diagnostic checks and tests for probabilistic calibration. Because probabilistic calibration is considered critical in forecasting, checks for the uniformity of the PIT have been routine in density forecast evaluation (Dawid 1984, Diebold et al. 1998, Gneiting et al. 2007). In typical practice, one observes a sample from the joint distribution of the probabilistic forecast

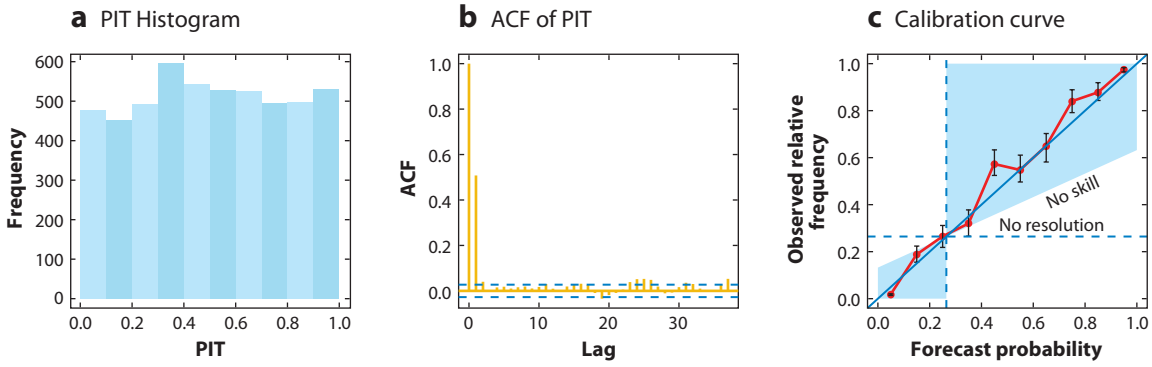

Figure 3

(*a*) Probability integral transform (PIT) histogram, (*b*) sample autocorrelation function (ACF) for the PIT values, and (*c*) calibration curve for exceedance of 10 m/s with bootstrap confidence intervals for two-hour-ahead regime-switching space-time forecasts of hourly average wind speed at the Stateline wind energy center (see Section 1.3). The plot in panel *c* was created using the R package VERIFICATION (Natl. Cent. Atmos. Res. Res. Appl. Prog. 2010) and subsequently modified slightly.

and the observation and then assesses uniformity graphically. This assessment is commonly done by examining histograms of the PIT values. For a probabilistically calibrated forecast, the PIT histogram is statistically uniform. U-shaped histograms indicate underdispersed predictive distributions, whereas hump or inverse-U-shaped histograms correspond to overdispersed predictive distributions (Diebold et al. 1998, Gneiting et al. 2007, Hamill 2001). For example, **Figure 3*a*** shows the PIT histogram for the RST forecasts of hourly average wind speed at the Stateline wind energy center described in Section 1.3. The histogram appears uniform, well in line with the empirical coverage of the central 50% and 90% prediction intervals; these coverage values are 51.2% and 88.4%, respectively. The PIT histogram for the AR forecasts (not shown) is slightly hump-shaped. The use of diagnostic tools for the evaluation of marginal calibration has been illustrated by Gneiting et al. (2007) in this context.

Formal tests of the hypothesis that a given forecasting method is probabilistically calibrated can also be employed, provided that these tests account for typically complex dependence structures, particularly in the case of time series forecasts (Corradi & Swanson 2006, Knüppel 2011). In time series settings, one typically considers sequential k-step-ahead forecasts. The PITs for ideal k-step-ahead forecasts are at most $(k-1)$-dependent, and in addition to tests for the uniformity of the PIT values, the assumption of independence can be checked by examining the sample autocorrelation function (ACF) of the PIT values (Diebold et al. 1998). In the Stateline case study, e.g., $k = 2$. Thus, the PIT values for ideal forecasts are at most 1-dependent, and the sample ACF of the RST forecasts appears to be compatible with this assumption (**Figure 3*b***).

In the case of a binary outcome, we can identify a CDF-valued random quantity with the probability forecast p for a success. Then p is conditionally calibrated if, conditional on p, the binary event materializes with probability p. In this setting, probabilistic and conditional calibration are equivalent (Gneiting & Ranjan 2013) and can be examined by means of a reliability diagram or calibration curve. A reliability diagram plots conditional event frequencies against binned forecast probabilities; deviations from the diagonal indicate violations of the conditional calibration criterion (Dawid 1986, Murphy & Winkler 1992, Ranjan & Gneiting 2010). For example, **Figure 3*c*** shows a calibration curve for the induced RST probability forecasts for whether wind speeds exceed 10 m/s in the Stateline case study, along with bootstrap pointwise confidence intervals.

2.3. Sharpness

Sharpness refers to the concentration of the predictive distributions, and thus it is exclusive to the forecasts. In contrast, notions of dispersion also consider the observations. In the case of density forecasts for a real-valued variable, sharpness can be assessed in terms of the associated prediction intervals. The mean widths of these intervals should be as short as possible, subject to the empirical coverage being at the nominal level. In the Stateline case study described in Section 1.3, the mean widths of the central 50% prediction intervals are 2.26 and 2.74 for RST and AR, respectively; for the 90% intervals, these widths are 5.44 and 6.55.

2.4. Combining Predictive Distributions

In many situations, probabilistic forecasts from distinct experts, organizations, or statistical models are available and may need to be aggregated into a single combined predictive distribution. In generating this distribution, one typically specifies an aggregation method, i.e., a family of combination formulas of the form

$$G_\theta : \mathcal{F} \times \cdots \times \mathcal{F} \to \mathcal{F}, \quad (F_1, \ldots, F_n) \mapsto G_\theta(F_1, \ldots, F_n),$$

where $\mathcal{F}$ is a suitable class of (nonrandom) CDFs, and the parameter θ is estimated from training data (Gneiting & Ranjan 2013).

To date, individual combination formulas have been studied in terms of certain theoretical characteristics, such as the strong setwise function property and the external Bayes property (Genest & Zidek 1986). Gneiting & Ranjan (2013) focused on calibration and dispersion and considered families of combination formulas in the prediction space setting. They derived the following useful result, which is stated in terms of PITs and concerns the ubiquitous linear pool (Geweke & Amisano 2011, Krüger 2013, Stone 1961).

> **Theorem 2:** Consider the linearly combined forecast $G = \sum_{i=1}^{n} w_i F_i$ with distinct components $F_1, \ldots, F_n$ and strictly positive weights $w_1, \ldots, w_n$, where $n \geq 2$. Then
>
> $$\mathrm{var}(Z_G) < \min_{i=1,\ldots,n} \mathrm{var}(Z_{F_i});$$
>
> i.e., the linearly combined forecast G is more dispersed than the least dispersed of the component distributions $F_1, \ldots, F_n$.

This result explains the success of linear aggregation in an overwhelming range of applications for which the component distributions tend to be underdispersed. However, it also shows that linear pools fail to be flexibly dispersive in a certain well-defined sense (Gneiting & Ranjan 2013), suggesting that more general, nonlinear aggregation methods, such as spread-adjusted and beta-transformed linear pools, may improve predictive performance (Gneiting & Ranjan 2013, Ranjan & Gneiting 2010).

3. PROPER SCORING RULES, CONSISTENT SCORING FUNCTIONS, AND A PREDICTIVE VIEW OF REGRESSION

Proper scoring rules provide summary measures of the predictive performance that allow for the joint assessment of calibration and sharpness. Generally, we take scores to be negatively oriented penalties that forecasters wish to minimize. We consider a generic convex class $\mathcal{F}$ of probability distributions on $\mathbb{R}$, which we identify with their respective CDFs.

3.1. Proper Scoring Rules

A scoring rule assigns a numerical score $S(F, y)$ to each pair (F, y), where $F \in \mathcal{F}$ is a probabilistic forecast and $y \in \mathbb{R}$ is the realized value.

3.1.1. Propriety. Proper scoring rules encourage forecasters to provide honest and careful quotes (Gneiting & Raftery 2007). To give a formal definition of proper scoring rules, we write

$$S(F, G) = \mathbb{E}_G[S(F, Y)]$$

for the expected score under G when the probabilistic forecast is F, for $F, G \in \mathcal{F}$, assuming tacitly that the expectation is well defined. The extended real line is denoted by $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$.

Definition 4: The scoring rule $S : \mathcal{F} \times \mathbb{R} \to \bar{\mathbb{R}}$ is proper relative to the class $\mathcal{F}$ if

$$S(G, G) \leq S(F, G) \qquad 1.$$

for all $F, G \in \mathcal{F}$. It is strictly proper if Equation 1 holds with equality only if $F = G$.

Thus, a proper scoring rule is designed such that quoting the true distribution as the forecast distribution is an optimal strategy in expectation. This property is critically important, as the use of improper scoring rules can lead to grossly misguided inferences about predictive performance (Gneiting 2011a, Gneiting & Raftery 2007, Hilden & Gerds 2013).

Given a proper scoring rule, we refer to the expected score function $e(F) = S(F, F)$ as the associated entropy and to the function $d(F, G) = S(F, G) - S(G, G) \geq 0$ as the corresponding divergence. Under slight regularity conditions, which we omit for brevity, the result given in Theorem 3 (Forbes 2012, Gneiting & Raftery 2007, Hendrickson & Buehler 1971) characterizes proper scoring rules using the tools and language of convex analysis (Rockafellar 1970).

Theorem 3: The scoring rule S is proper relative to the class $\mathcal{F}$ if and only if the expected score function $e(F) = S(F, F)$ is concave and $S(F, \cdot)$ is a supergradient of e at the point F, for all $F \in \mathcal{F}$.

For example, let $\mathcal{F}$ be the class of the probability measures with a square-integrable Lebesgue density, f. Consider the quadratic score (QS)

QS: quadratic score

$$QS(f, y) = -2f(y) + \int_{\mathbb{R}} f^2(z)\, dz.$$

Then $e(f)$ is concave with supergradient $QS(f, \cdot)$; thus, the QS is proper. An interesting observation here is that although the linear score, $S(f, y) = -f(y)$, has the same expected score function $e(f)$, the linear score is not a supergradient and is thus improper (Ovcharov 2013).

3.1.2. Local proper scoring rules. Much recent attention has focused on notions of locality. Let k be a nonnegative integer, and let S be a scoring rule for a convex class, $\mathcal{F}$, of probability measures on $\mathbb{R}$ that admit a Lebesgue density, f, with continuous derivatives up to order k. Then S is local of order k if there exists a function $s : \mathbb{R}^{2+k} \to \bar{\mathbb{R}}$ such that $S(f, y) = s(y, f(y), f'(y), \ldots, f^{(k)}(y))$ for all density forecasts $f \in \mathcal{F}$ and $y \in \mathbb{R}$. For example, the logarithmic score (LS),

LS: logarithmic score

$$LS(f, y) = -\log f(y),$$

is a local proper scoring rule of order $k = 0$. Although the LS is the unique rule of this type up to equivalence (Bernardo 1979, Good 1952), nontrivial local proper scoring rules of order $k \geq 2$ exist. The Hyvärinen score (HS),

HS: Hyvärinen score

$$\mathrm{HS}(f, y) = 2\frac{f''(y)}{f(y)} - \left(\frac{f'(y)}{f(y)}\right)^2, \qquad 2.$$

is the most prominent example. In a far-reaching and elegant recent paper, Parry et al. (2012) proved the existence of local proper scoring rules of any even order $k \geq 0$. In related work, Ehm & Gneiting (2012) characterized the local proper scoring rules of order $k = 2$ relative to a broad class $\mathcal{F}$. Interestingly, the HS in Equation 2 can be computed without knowing the normalizing constant for the density forecast f. This property is characteristic of local proper scoring rules other than the logarithmic score (Parry et al. 2012) and allows for the use of M-estimators in situations in which normalizing constants are unavailable (Hyvärinen 2005).

CRPS: continuous ranked probability score

3.1.3. Continuous ranked probability score. The restriction to density forecasts can be impractical, and we now discuss proper scoring rules that are specified directly in terms of predictive CDFs. The continuous ranked probability score (CRPS) is defined as

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbb{1}\{y \leq x\})^2 \, \mathrm{d}x \qquad 3.$$

$$= \mathbb{E}_F|Y - y| - \frac{1}{2}\mathbb{E}_F|Y - Y'|, \qquad 4.$$

where Y and Y' are independent random variables with CDF F and finite first moment (Gneiting & Raftery 2007, Matheson & Winkler 1976). The representation in Equation 4 gives the CRPS in the same unit as the observations, and it generalizes the absolute error. (The CRPS reduces to the absolute error if F is a point forecast, i.e., a point measure.) Thus, the CRPS provides a direct way of comparing point forecasts and probabilistic forecasts. Weighted versions are also available (Gneiting & Ranjan 2011, Matheson & Winkler 1976).

DSS: Dawid–Sebastiani score

3.1.4. Dawid–Sebastiani score. The CRPS has many attractive properties, but it can be hard to compute for complex forecast distributions. A viable alternative that depends on the probabilistic forecast, F, through only its first two central moments, μ_F and σ_F^2, is given by the proper Dawid–Sebastiani score (DSS) (Dawid & Sebastiani 1999),

$$\mathrm{DSS}(F, y) = \frac{(y - \mu_F)^2}{\sigma_F^2} + 2\log\sigma_F.$$

Table 1 shows closed-form expressions for the proper scoring rules described here under Gaussian predictive distributions.

3.1.5. Stateline case study. In practice, forecasting procedures are ranked by averaging scores over a test set. **Table 2** considers the wind forecasting example and competing methods presented in Section 1.3. In terms of all scores, the RST method performs best, whereas the NC point forecast performs worst.

3.2. Consistent Scoring Functions and Elicitable Functionals

As we have argued, forecasts ought to be probabilistic, taking the form of probability distributions over future quantities or events. However, practical situations may require single-valued point

Table 1 Explicit forms of some scoring rules under Gaussian predictive distributions

Scoring rule	$S(\mathcal{N}(\mu, \sigma^2), y)$
Quadratic score	$-\frac{2}{\sigma}\varphi\left(\frac{y-\mu}{\sigma}\right) + \frac{1}{2\sqrt{\pi}\sigma}$
Logarithmic score[a]	$\frac{(y-\mu)^2}{2\sigma^2} + \log\sigma + \frac{1}{2}\log 2\pi$
Hyvärinen score	$\frac{1}{\sigma^2}\left(\frac{(y-\mu)^2}{\sigma^2} - 2\right)$
Continuous ranked probability score	$\sigma\left(\frac{y-\mu}{\sigma}\left(2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1\right) + 2\varphi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right)$

[a]For Gaussian predictive distributions, the Dawid–Sebastiani score is the same as the logarithmic score up to an affine transformation.

forecasts, for reasons of decision making, reporting requirements, or communications, among others. Following the recent work of Gneiting (2011a), we now address this type of situation.

In common practice, competing point forecasts are compared using a nonnegative loss or scoring function $s(x, y)$, which represents the penalty when the point forecast x is issued and the observation y is realized. Given a predictive probability distribution $F \in \mathcal{F}$ for the future observation, the Bayes rule or optimal point forecast is any $\hat{x} \in \mathbb{R}$ such that

$$\hat{x} = \arg\min_x \mathbb{E}_F[s(x, Y)], \qquad 5.$$

where Y is a random variable with distribution F. In most cases of practical interest, Bayes rules exist, and these rules are frequently unique (Ferguson 1967).

3.2.1. Consistent scoring functions. For decision-theoretically coherent point forecasting, forecasts need a directive in the form of a statistical functional (Gneiting 2011a). Formally, a statistical functional (or simply a functional) is a potentially set-valued mapping $T(F)$ from a class of probability distributions, $\mathcal{F}$, to the real line, $\mathbb{R}$, for which the mean or expectation functional, quantiles, and expectiles (Newey & Powell 1987) are key examples.

Table 2 Comparison of the predictive performance of the forecasting methods in the Stateline case study introduced in Section 1.3[a]

	QS[b]	LS[b]	HS[b]	DSS[c]	CRPS	AE	SE	$IS_{0.90}$
NC					1.61	1.61	4.88	32.17
AR	−0.15	2.14	−0.24	2.44	1.12	1.53	4.19	9.11
SB				5.21	1.06	1.46	3.82	8.55
RST	−0.18	1.96	−0.35	2.08	0.96	1.34	3.19	7.55

[a]The scoring rules used in this table are defined and discussed throughout Section 3.
[b]These scores require a predictive density and hence are not defined for the discrete forecasts NC and SB.
[c]These scores require a positive variance and hence are not defined for the point forecast NC. The high value for SB is due to a number of cases in which the predictive variance was close to zero.
Abbreviations: AE, absolute error; AR, autoregressive; CRPS, continuous ranked probability score; DSS, Dawid–Sebastiani score; HS, Hyvärinen score; IS, interval score; LS, logarithmic score; NC, no-change forecast; QS, quadratic score; RST, regime-switching space-time; SB, similarity-based; SE, squared error.

Definition 5: The scoring function s is consistent for the functional T relative to the class $\mathcal{F}$ if

$$\mathbb{E}_F[s(t, Y)] \leq \mathbb{E}_F[s(x, Y)] \quad 6.$$

for all probability distributions $F \in \mathcal{F}$, all $t \in T(F)$, and all $x \in \mathbb{R}$. It is strictly consistent if (*a*) it is consistent and (*b*) equality in Equation 6 implies that $x \in T(F)$.

Consistent scoring function: a special case of a proper scoring rule in the context of point forecasts

Any consistent scoring function induces a proper scoring rule in a straightforward and natural construction (Gneiting 2011a).

Theorem 4: Suppose that the scoring function s is consistent for the functional T relative to the convex class $\mathcal{F}$. For each $F \in \mathcal{F}$, let $t_F \in T(F)$. Then $S(F, y) = s(t_F, y)$ is a proper scoring rule relative to the class $\mathcal{F}$.

3.2.2. Elicitable functionals. We turn to the notion of elicitability, which is a critically important concept in the evaluation of point forecasts. Although the idea dates back to the pioneering work of Osband (1985), the term elicitable was coined only recently by Lambert et al. (2008).

Definition 6: The functional T is elicitable relative to the class $\mathcal{F}$, if there exists a scoring function s that is strictly consistent for T relative to $\mathcal{F}$.

Elicitability: a critical property of a statistical functional that allows for decision-theoretically principled forecast evaluation

SE: squared error

IS_α: interval score for central prediction intervals with nominal coverage $\alpha \in (0, 1)$

Many commonly used functionals, such as means, quantiles, and expectiles (Newey & Powell 1987), are elicitable. For example, the squared error (SE) scoring function, $s(x, y) = (x - y)^2$, is strictly consistent for the mean functional relative to the class of probability distributions on $\mathbb{R}$ with finite second moments. Thus, means or expectations are elicitable. According to Savage's (1971) classic result, subject to weak regularity conditions, the scoring function s is consistent for the mean functional if and only if it is of Bregman form (Banerjee et al. 2005), i.e.,

$$s(x, y) = \phi(y) - \phi(x) - \phi'(x)(y - x), \quad 7.$$

where ϕ is convex with subgradient ϕ'. Generally, if ϕ is strictly convex, the scoring function is strictly consistent, and the SE scoring function arises when $\phi(t) = t^2$.

The asymmetric piecewise linear scoring function, $s(x, y) = (\mathbb{1}\{y < x\} - \alpha)(x - y)$, is consistent for the α-quantile functional ($0 < \alpha < 1$). Scores of this type can be combined to yield a proper scoring rule for interval forecasts,

$$IS_\alpha(l, r; y) = (r - l) + \frac{2}{\alpha}(l - y)\mathbb{1}\{y < l\} + \frac{2}{\alpha}(y - r)\mathbb{1}\{y > r\}, \quad 8.$$

where l and r denote the $\frac{\alpha}{2}$ and $(1 - \frac{\alpha}{2})$ quantiles, respectively, that bound the central $(1 - \alpha)$ prediction interval (Gneiting & Raftery 2007). The interval score (IS) equals a weighted sum (with weights depending on α) of the length of the prediction interval, $r - l$, and the distance between the realization, y, and the interval. Forecasts are rewarded for narrow prediction intervals that capture the realization.

Subject to slight regularity conditions, a scoring function is consistent for the α-quantile functional if and only if that function is a generalized piecewise linear (GPL) function of order α, i.e.,

$$s(x, y) = (\mathbb{1}\{y < x\} - \alpha)(g(x) - g(y)), \quad 9.$$

where g is nondecreasing (Gneiting 2011a,b; Thomson 1979). Furthermore, if g is strictly increasing, s is strictly consistent. GPL loss functions arise in a wide range of practically relevant settings, including commodity and energy markets (Basu & Markov 2004, Gneiting 2011b,

Granger & Newbold 1986). The absolute error (AE) scoring function arises when $g(t) = t$ and $\alpha = 1/2$ in Equation 9.

Newey & Powell (1987) introduced the τ-expectile functional ($0 < \tau < 1$) of a probability measure with a finite first moment. Like the α-quantile, which is given by the Bayes rule under the asymmetric piecewise linear function, the τ-expectile is given by the Bayes rule or optimal point forecast from Equation 5 under the asymmetric piecewise quadratic scoring function, $s(x, y) = |\mathbb{1}\{y < x\} - \tau|\,(x-y)^2$. Unsurprisingly, expectiles have properties that resemble those of quantiles (Newey & Powell 1987). Under standard regularity conditions, any scoring function s that is consistent for the τ-expectile takes the form

$$s(x, y) = |\mathbb{1}\{y < x\} - \tau|\,(\phi(y) - \phi(x) - \phi'(x)(y - x)), \qquad 10.$$

where ϕ is convex with subgradient ϕ' (Gneiting 2011a), thereby combining the key characteristics of the Bregman and GPL families in Equations 7 and 9, respectively.

3.2.3. Nonelicitable functionals. The α-conditional value-at-risk or expected shortfall (CVaR_α, $0 < \alpha < 1$) equals the expectation of a random variable with distribution F, conditional on the random variable taking values in its upper $(1-\alpha)$–tail (Rockafellar & Uryasev 2002). The CVaR_α is a popular risk measure in quantitative finance, and one of its elegant and appealing properties is coherency in the sense of Artzner et al. (1999). Unfortunately, this functional is not elicitable (Gneiting 2011a), which challenges its use as a predictive measure of risk. Because consistent scoring functions are not available, how one might assess and compare methodologies for CVaR_α forecasts remains unclear. However, a transition to probabilistic forecasts may be a potential remedy (Ziegel 2013).

3.3. Testing for Equal Predictive Performance

Suppose that we have two competing forecasting methods, say F and G, and we wish to test the hypothesis that these two methods have equal predictive performance in the sense that the expectation of the score differential vanishes. In practice, forecasting procedures are ranked by their average score over a test set. Focusing for now on proper scoring rules, the corresponding average scores are

$$\bar{S}_n^F = \frac{1}{n}\sum_{i=1}^{n} S(F_i, y_i) \quad \text{and} \quad \bar{S}_n^G = \frac{1}{n}\sum_{i=1}^{n} S(G_i, y_i),$$

respectively. The same considerations apply to the special case of consistent scoring functions (Theorem 4).

3.3.1. Diebold–Mariano test. If the forecast cases are independent, a test of equal forecast performance can be based on the statistic

$$t_n = \sqrt{n}\,\frac{\bar{S}_n^F - \bar{S}_n^G}{\hat{\sigma}_n}, \qquad 11.$$

where

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} (S(F_i, y_i) - S(G_i, y_i))^2 \qquad 12.$$

is an estimate of the variance of the score differential. Subject to traditional regularity conditions, the statistic t_n is asymptotically standard normal under the null hypothesis of vanishing expected

AE: absolute error

score differentials, and one- or two-sided asymptotic tail probabilities are readily calculated. If the null hypothesis is rejected, F is preferred if t_n is negative, whereas G is preferred if t_n is positive.

In the case of sequential k-step-ahead time series forecasts, the assumption of independence between the score differentials might be violated. In this setting, Diebold & Mariano (1995) generalize the variance estimate given in Equation 12 to account for autocorrelation. Subject to regularity conditions, the corresponding statistic t_n in Equation 11 remains asymptotically standard normal under the null hypothesis of vanishing expected score differentials. This type of procedure is often referred to as a Diebold–Mariano test (Diebold 2012, Diebold & Mariano 1995). If the sample size is small, permutation tests offer alternatives (D'Agostino et al. 2012, Diebold & Mariano 1995).

3.3.2. Stateline case study. In **Table 2**, we compare the forecasting methods of Section 1.3 in terms of the mean SE for the respective predictive mean, the mean AE for the corresponding predictive median, and the average IS from Equation 8 for the central 90% prediction interval. We also carry out the Diebold–Mariano test to compare the SB (F) and RST (G) methods using the CRPS. Using $k = 2$, $n = 5{,}136$, $\bar{S}_n^F - \bar{S}_n^G = 0.099$, and $\hat{\sigma}_n^2 = 0.62$ results in a test statistic of $t_n = 8.98$. The corresponding p-value for a test of the null hypothesis of vanishing score differentials is essentially zero.

3.4. A Predictive View of Regression

This section provides a succinct, subjective view of regression from a predictive perspective. From a predictive standpoint, regression should aim to model the conditional distribution of a response variable in terms of a collection of explanatory variables that represent the information at hand. Hothorn et al. (2013) comment in this context that

> [t]he ultimate goal of regression analysis is to obtain information about the conditional distribution of a response given a set of explanatory variables. This goal is, however, seldom achieved because most established regression models only estimate the conditional mean as a function of the explanatory variables.

For ease of exposition, we assume a real-valued response variable. We distinguish mean, quantile, and expectile regression, which are based on the use of consistent scoring functions, from distributional regression approaches, which specify full conditional distributions and can be linked to the use of proper scoring rules.

3.4.1. Mean, quantile, and expectile regression. Frequently, one pursues a goal more modest than distributional regression, in that only a specific functional of the conditional predictive distribution is modeled.

In ordinary least squares regression, one seeks the conditional expectation of the response variable given the explanatory variables. Because the SE scoring function is consistent for the mean or expectation functional, it is natural to estimate parameters with least squares techniques. Similarly, generalized linear models (McCullagh & Nelder 1989) relate the mean of a certain specified distribution to a linear function in the explanatory variables via a suitably chosen link function.

Moving beyond mean regression, common approaches focus on conditional quantiles or conditional expectiles (Kneib 2013). Quantile regression models a quantile of the response variable conditional on the explanatory variables (Koenker 2005, Koenker & Bassett 1978). To estimate

the regression coefficients from training data, one uses the asymmetric piecewise linear scoring function, which is consistent for the α-quantile. Expectile regression works analogously but is based on the asymmetric piecewise quadratic scoring function (Efron 1991, Newey & Powell 1987, Sobotka & Kneib 2012). Although in principle one can build full conditional distributions from conditional quantiles or expectiles, any such approach suffers from the problem of quantile or expectile crossing, necessitating major additional effort (Dette & Volgushev 2008, Kneib 2013).

3.4.2. Distributional regression: parametric approaches. As noted, the ultimate goal of regression analysis is to model the conditional distribution of the response variable given a set of explanatory variables. Both parametric and nonparametric approaches are feasible and commonly used.

In the parametric setting, Poisson regression serves as an incidental example of distributional regression because the Poisson distribution is fully specified by a single parameter. Such an approach has limited flexibility, however, and frequently one must account for overdispersion (Lawless 1987). In practice, forecasters should seek specific solutions that are tailored to the problem at hand and that yield parametric models for full predictive distributions. For example, in the Stateline case study discussed in Section 1.3, the response variable is the hourly average wind speed at Vansycle with a prediction horizon of two hours ahead. The explanatory variables are the current and past observations of wind speed and wind direction at Goodnoe Hills, Kennewick, and Vansycle, and the predictive distributions for the RST method are truncated normal. This is an instance of the approach described by Cannon (2012, p. 126) with reference to Cawley et al. (2007), in that

> a parametric probability distribution is specified for the predictand of interest and then some form of regression model, either linear or nonlinear, is used to estimate parameters . . . conditioned upon values of a separate set of predictors.

Another example of such an approach is given in Section 4.2, where we discuss statistical postprocessing techniques for ensemble weather forecasts. Pers et al. (2009) view machine learning techniques from the same perspective. In addition, in time series and/or spatial settings (Cressie & Wikle 2011, Granger & Newbold 1986), the practice of providing a full predictive distribution for the quantity of interest, rather than just a point forecast, is becoming more common.

The parameters of a distributional regression model can be estimated by optimizing a proper score averaged over a training set (Dawid 2007, Gneiting & Raftery 2007, Gneiting et al. 2005, Hothorn et al. 2013), a technique traditionally referred to as minimum contrast estimation (Birgé & Massart 1993, Pfanzagl 1969). In the special case of the logarithmic score, this amounts to maximum likelihood estimation. Wald's (1949) classic proof of the consistency of maximum likelihood estimates depends only on propriety and thus applies to general optimum score estimates.

3.4.3. Distributional regression: nonparametric approaches. We now discuss semiparametric and nonparametric approaches to the modeling of conditional distributions. The simplistic n-nearest-neighbor method uses a suitable notion of distance to compute similarity values between the given values of the explanatory variables and the values of these variables in the training cases. The conditional predictive distribution, then, is the empirical distribution of the response variable among the n nearest neighbors in the training set, for which proximity is defined in terms of the similarity measure.

The success of nearest-neighbor techniques in particular and of similarity-based methods in general (Gentner & Holyoak 1997) can be explained in part by the Cover–Hart inequality (Cover & Hart 1967). According to this inequality, the expected SE of a 1-nearest-neighbor point forecast

has an upper bound of twice the Bayes risk. Perhaps surprisingly, this favorable upper estimate continues to hold under a wide range of scoring functions and proper scoring rules, including the CRPS (Gneiting 2012).

More sophisticated nonparametric methods weight the training cases depending on the similarity values (Stone 1977). If a predictive density is desired, kernel smoothing can be used to convert the discrete distribution into a smooth Lebesgue density. Implementation choices for similarity methods of this kind are intricate and include the choice of the distance measure, the conversion to weights, the smoothing kernel, and the bandwidth (Hall et al. 1999, 2004; Hyndman et al. 1996). In a related Bayesian development, Dunson et al. (2007) discuss density regression.

For the wind forecasting example in Section 1.3 (the Stateline case study), we used a simple similarity-based (SB) method with the current wind speed at Vansycle and the coordinates of the current wind vector at the three meteorological stations as explanatory variables. These variables were standardized to have a variance of 1. The distance measure d was Euclidean distance; the weight was taken as $w_i = f(d_i)/\sum f(d_j)$, where $f(d) = \exp(-d^2)$; and we did not use a smoothing kernel, i.e., the predictive distribution was discrete. Our training set was determined by a 1,000-hour rolling window. Jeon & Taylor (2012) recently proposed a related, more sophisticated approach for probabilistic wind energy forecasting.

4. PROBABILISTIC WEATHER FORECASTING

Arguably, the most mature and successful implementation of probabilistic forecasting methods is in weather prediction. Here we consider medium-range forecasts, for which the lead times are on the order of days.

4.1. Ensemble Forecasts

Ensemble forecast: a collection of point forecasts for a specific quantity or event

For prediction horizons of up to a few hours, statistical methods are the preferred techniques in weather forecasting, as shown by our Stateline case study example in Section 1.3. For lead times of more than a few hours, however, weather centers draw on highly sophisticated numerical models of the atmosphere that are run in real time to produce point forecasts of future atmospheric states. In a strong move toward distributional forecasts, these efforts have been transformed via the operational implementation of ensemble forecasts over the past two decades (Gneiting & Raftery 2005, Palmer 2002). Such efforts have had major economic and societal benefits. An ensemble prediction system consists of multiple runs—typically between 10 and 50—of numerical weather prediction models, which differ either in the initial conditions used or in the model's parameterized analytic representation of the atmosphere. For example, **Figure 4** shows the predictions from four randomly selected members of the 50-member European Center for Medium-Range Weather Forecasts (ECMWF) ensemble (Buizza et al. 1999, 2005; Molteni et al. 1996; Leutbecher & Palmer 2008; Richardson 2000) that were valid April 4, 2011, at 2:00 AM local time for temperature over Germany, promising an agreeable spring night.

4.2. Statistical Postprocessing of Ensemble Forecasts

Ideally, we would like to consider an ensemble forecast as a random sample from the predictive distribution of future states of the atmosphere. However, doing so is rarely feasible in practice because ensemble forecasts are subject to biases and dispersion errors and thereby call for some form of statistical postprocessing (Bröcker & Smith 2008, Gneiting & Raftery 2005, Wilks & Hamill 2007).

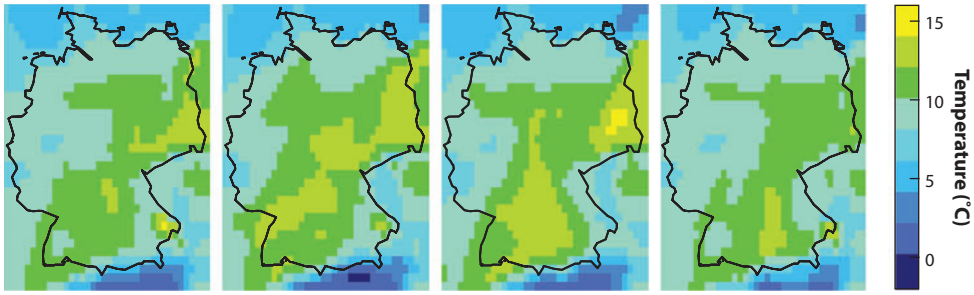

Figure 4

Twenty-four-hour-ahead temperature forecasts valid April 4, 2011, at 2:00 AM local time over Germany for four randomly selected members of the 50-member European Center for Medium-Range Weather Forecasts ensemble.

State-of-the-art techniques for statistical postprocessing include the nonhomogeneous regression (NR) or ensemble model output statistics (EMOS) technique proposed by Gneiting et al. (2005) and the ensemble Bayesian model averaging (BMA) approach developed by Raftery et al. (2005).

To fix the idea of NR, let y denote a real-valued weather variable of interest and write $x_1, \ldots, x_m$ for the corresponding ensemble member forecasts. The NR predictive distribution is a single parametric distribution of the general form

$$y|x_1, \ldots, x_m \sim f(y|x_1, \ldots, x_m).$$

The left side of the equation refers to the conditional distribution of y given the ensemble member forecasts $x_1, \ldots, x_m$, which serve as explanatory variables. The right side shows an instance of a parametric distributional regression approach, where f is a parametric density function for which the parameters depend on the ensemble values in suitable ways.

The BMA method employs a mixture distribution of the general form

BMA: Bayesian model averaging

$$y|x_1, \ldots, x_m \sim \sum_{i=1}^{m} w_i g(y|x_i),$$

where $g(y|x_i)$ denotes a parametric density or kernel that depends on the ensemble member forecast x_i in suitable ways. The mixture weights $w_1, \ldots, w_m$ reflect the skill of their corresponding members over the training period, in keeping with a nonparametric distributional regression technique. In the case of exchangeable members, the constraint $w_1 = \cdots = w_m = 1/m$ applies (Fraley et al. 2010).

In either approach, the parameters of the predictive model are estimated on a rolling training period, which typically consists of the most recent 20 to 40 days, using optimum score techniques. The choice of the NR predictive density f and the BMA component density g depends critically on the weather variable of interest. **Tables 3** and **4** sketch NR and BMA implementations, respectively, for the most important weather variables (Ben Bouallègue 2013; Fraley et al. 2010; Gneiting & Raftery 2005; Kleiber et al. 2011a,b; Raftery et al. 2005; Scheuerer 2013; Sloughter et al. 2007, 2010; Thorarinsdottir & Gneiting 2010; Wilks 2009) and are similar to summaries available elsewhere (Möller et al. 2013, Schefzik et al. 2013). The simplest case arises in the NR model for temperature or pressure (Gneiting et al. 2005). This model employs a Gaussian predictive density, in that

$$y|x_1, \ldots, x_m \sim \mathcal{N}(a + b_1 x_1 + \cdots + b_m x_m, c + ds^2)$$

Table 3 Nonhomogeneous regression implementations

Weather quantity	Range	Functional form
Temperature	$y \in \mathbb{R}$	Normal
Precipitation amount[a]	$y^{1/2} \in \mathbb{R}^+$	Truncated logistic
	$y \in \mathbb{R}^+$	Generalized extreme value
Wind speed	$y \in \mathbb{R}^+$	Truncated normal

[a]We consider the approaches of Wilks (2009) and Scheuerer (2013). In the former case, we refer to $y^{1/2} \in \mathbb{R}^+$ because the truncated logistic distribution applies to root-transformed precipitation accumulations.

with bias parameters a and $b_1, \ldots, b_m$ and spread parameters c and d, where s^2 is the variance of the ensemble member values. If the ensemble members are exchangeable, as for the ECMWF ensemble, one requires that $b_1 = \cdots = b_m$. For an example, see **Figure 5**.

4.3. Copula Approaches to Probabilistic Forecasts of Multivariate Quantities

Statistical postprocessing techniques such as NR and BMA typically apply to a single weather variable at a single location and a single look-ahead time. However, in many applications, dependencies in combined events must be properly accounted for. For example, planning for winter road maintenance requires joint probabilistic forecasts of temperature and precipitation (Berrocal et al. 2010), renewable energy forecasting depends on spatiotemporal weather scenarios (Pinson et al. 2009, Pinson 2013), and using ensemble forecasts to drive hydrologic models relies on physically realistic precipitation patterns (Cloke & Pappenberger 2009). If statistical postprocessing proceeds independently for each weather variable, location, and look-ahead time, dependencies are ignored and must be restored.

Copula: a multivariate cumulative distribution function with standard uniform margins

Standard approaches to statistical postprocessing yield a postprocessed predictive CDF, F_l, for each univariate weather quantity, Y_l, where, say, $l = 1, \ldots, L$. We seek a physically realistic and consistent multivariate or joint predictive CDF, F, with margins $F_1, \ldots, F_L$. The celebrated theorem of Sklar (1959) shows that every multivariate CDF F with margins $F_1, \ldots, F_L$ can be represented in the form

$$F(y_1, \ldots, y_L) = C(F_1(y_1), \ldots, F_L(y_L))$$

for $y_1, \ldots, y_L \in \mathbb{R}$, where $C : [0, 1]^L \to [0, 1]$ is a copula, i.e., a multivariate CDF with standard uniform margins. If each F_l is continuous, the copula is unique. In the case of empirical measures, it suffices to consider empirical copulas, and general considerations remain unchanged (Schefzik et al. 2013).

Table 4 Bayesian model averaging implementations based on ensemble values x_i, where $i = 1, \ldots, m$

Weather quantity	Range	Kernel	Mean	Variance
Temperature	$y \in \mathbb{R}$	Normal	$a_{0i} + a_{1i} x_i$	σ_i^2
Precipitation amount[a]	$y^{1/3} \in \mathbb{R}^+$	Gamma	$a_{0i} + a_{1i}\, x_i^{1/3}$	$b_{0i} + b_{1i}\, x_i$
Wind speed	$y \in \mathbb{R}^+$	Gamma	$a_{0i} + a_{1i}\, x_i$	$b_{0i} + b_{1i}\, x_i$

[a]In the case of precipitation amount, we refer to $y^{1/3} \in \mathbb{R}^+$ because the gamma kernels apply to cube-root-transformed precipitation accumulations (Sloughter et al. 2007).

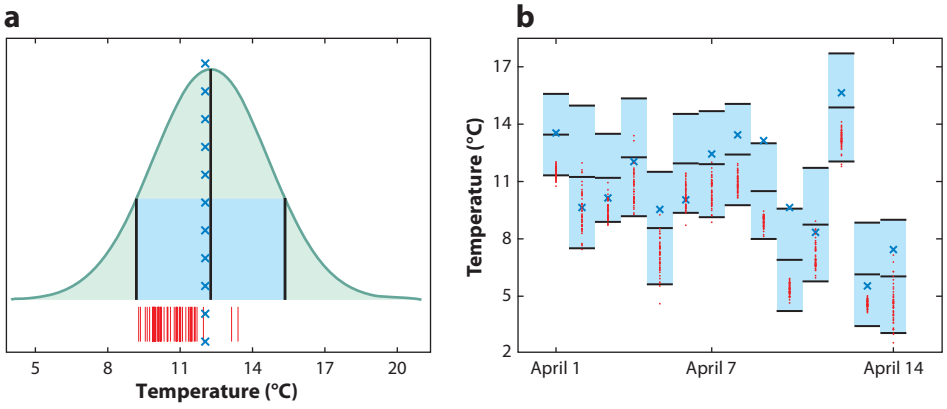

Figure 5

Twenty-four-hour-ahead nonhomogeneous regression (NR) postprocessed predictive distributions for temperature valid (*a*) April 4, 2011 and (*b*) April 1–14, 2011 at 2:00 AM local time in Frankfurt, Germany. Predictive distributions are based on the 50-member European Center for Medium-Range Weather Forecasts ensemble. The ensemble member forecasts are shown in red, the NR medians and 90% central prediction intervals are shown in black, and the realized temperatures are shown in blue.

Sklar's theorem demonstrates that standard approaches to statistical postprocessing can accommodate any type of joint dependence structure, provided that a suitable copula function is specified. If the dimension L is small, or if a specific type of structure, such as temporal or spatial structure, can be exploited, parametric families of copulas can be fitted. Gaussian copulas are a popular choice (Berrocal et al. 2008, Gel et al. 2004, Möller et al. 2013, Pinson et al. 2009, Schuhen et al. 2012); this approach is similar to the Gaussian copula regression approach described by Masarotto & Varin (2012).

In contrast, if L is large and no specific structure can be exploited, one needs to resort to nonparametric approaches, adopting dependence structures either from records of historical weather observations or from the ensemble forecast at hand, as embodied in empirical copulas. The Schaake shuffle borrows the rank order structure from suitably chosen past weather records (Clark et al. 2004), whereas the ensemble copula coupling (ECC) approach draws on rank dependence information supplied by the available ensemble forecast (Schefzik et al. 2013). The ECC technique generates a postprocessed ensemble forecast with the same number of members and the same rank-order structure as the original ensemble, which **Figure 6** and various recent case studies illustrate (Flowerdew 2012, Pinson 2012b, Roulin & Vannitsem 2012, Schefzik et al. 2013). Schefzik et al. (2013) show that, essentially, ECC applies the empirical copula of the original ensemble to samples from the postprocessed predictive distributions.

Schaake shuffle: a method that adopts the empirical copula from a record of observations

Ensemble copula coupling (ECC): a method that adopts the empirical copula from the ensemble forecast at hand

4.4. Discussion

Although we have restricted this section to discussion of ensembles of numerical weather prediction models, ensemble forecasts have been gaining importance across scientific disciplines (Araújo & New 2006, Cloke & Pappenberger 2009, Lozano et al. 2011). Indeed, the need for fruitful interplay between analytic-numerical and statistical modeling is becoming ubiquitous, similar to the recent surge of interest in the emerging field of uncertainty quantification at the cutting edge of the interface among statistics, applied mathematics, and application domains. Both ensemble forecasting and uncertainty quantification seek to provide physically realistic, calibrated, and sharp

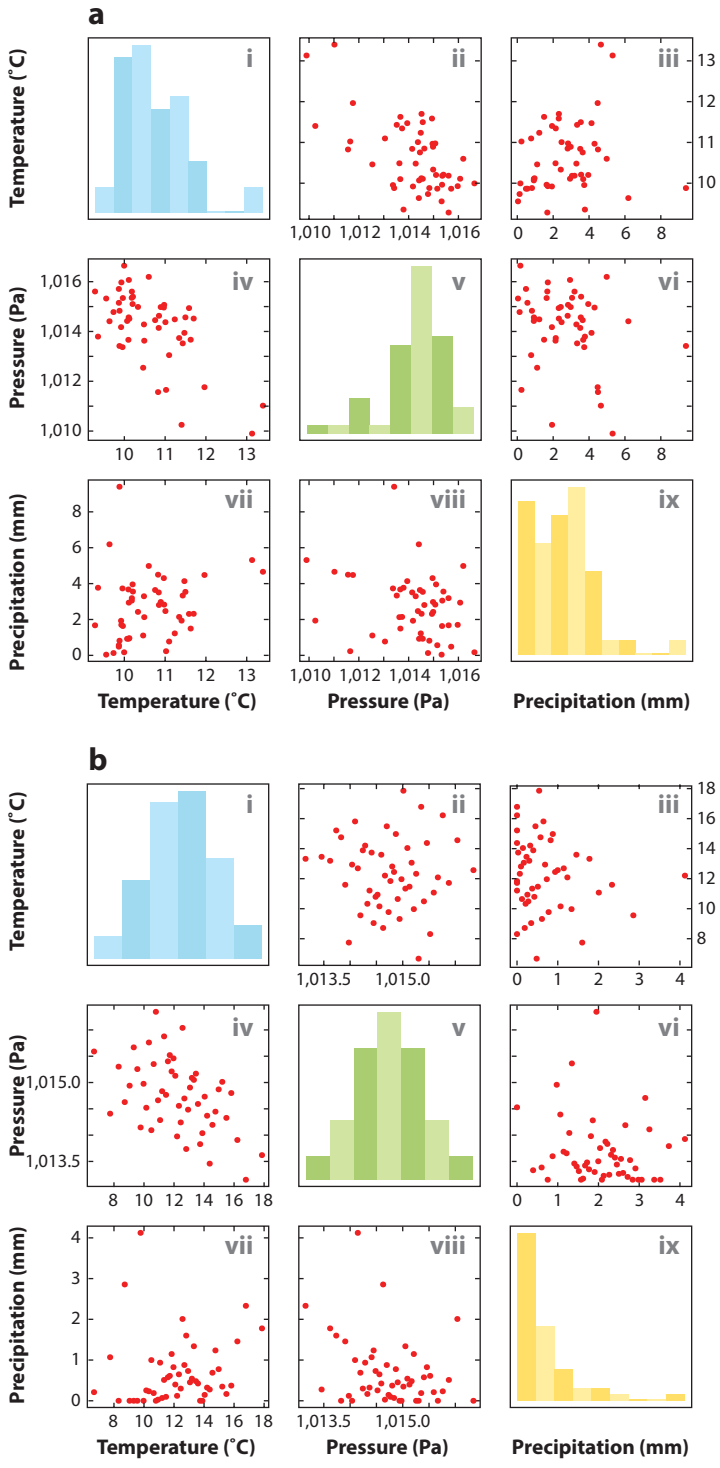
a
i
ii
iii
iv
v
vi
vii
viii
ix
Temperature (°C)
Pressure (Pa)
Precipitation (mm)
b
i
ii
iii
iv
v
vi
vii
viii
ix
Temperature (°C)
Pressure (Pa)
Precipitation (mm)

Figure 6

Twenty-four-hour-ahead ensemble forecasts valid April 4, 2011, at 2:00 AM local time in Frankfurt, Germany, for temperature, pressure, and six-hour accumulated precipitation. (*a*,*b*) Subpanels i, v, and ix show marginal histograms (*blue*, temperature; *green*, pressure; *yellow*, precipitation); panels ii, iii, iv, vi, vii, and viii show bivariate scatterplots. (*a*) Unprocessed forecasts from the 50-member European Center for Medium-Range Weather Forecasts. (*b*) Panels ii, iii, and vi show a statistical sample of size 50 from independently Bayesian model averaged postprocessed marginal predictive cumulative distribution functions. The forecast in panels iv, vii, and viii has been subjected to ensemble copula coupling, which retains the margins while restoring the rank dependence structure from the unprocessed ensemble (Schefzik et al. 2013).

probabilistic forecasts of multivariate quantities, including forecasts for temporal, spatial, and/or spatiotemporal scenarios and trajectories. Theoretically principled and practically useful tools for evaluating probabilistic forecasts in such settings are in great demand (Gneiting et al. 2008, Pinson 2013, Pinson & Girard 2012).

SUMMARY POINTS

1. Across scientific disciplines, we are witnessing a transition from point forecasts to probabilistic forecasts, which take the form of probability distributions over future quantities or events.
2. Probabilistic forecasts aim to maximize their sharpness, subject to calibration. Calibration concerns the statistical compatibility between the probabilistic forecasts and the realized observations; sharpness refers to the concentration of the predictive distributions and thus is a property exclusive to the forecasts.
3. In practice, calibration can be examined via PIT histograms.
4. Scoring rules assess calibration and sharpness simultaneously. These rules must be proper to encourage honest and careful forecasting. An especially attractive example is the CRPS.
5. The scoring function used to evaluate a point forecast must be consistent for the task at hand; e.g., the AE is consistent for the median, and the SE is consistent for the mean.
6. Regression can be viewed from a predictive perspective. Again, distributional predictions are gaining importance. The parameters of the predictive distributions can be estimated from training data using proper scoring rules; maximum likelihood estimation is a special case thereof.
7. Ensemble forecasts have been gaining importance in a wealth of applications.
8. Weather prediction arguably provides the most advanced practical example of real-time probabilistic forecasting. Statistical postprocessing techniques such as NR and BMA supplement and improve numerical, atmospheric physics–based ensemble forecasts.

FUTURE ISSUES

1. Parametric and nonparametric copula techniques for probabilistic forecasts of multivariate quantities and events must be developed further and compared in terms of their predictive performance.

2. There is a pressing need for the development of decision-theoretically principled methods for the evaluation of probabilistic forecasts of multivariate variables.
3. Suitable modes of calibration other than probabilistic and marginal calibration need to be defined and studied.
4. Proper scoring rules and consistent scoring functions form convex cones. This calls for the development of Choquet representations other than the Schervish decomposition (Gneiting & Raftery 2007, Schervish 1989) in the case of probability forecasts of a binary event.
5. Local proper scoring rules allow for optimum score estimation of nonnormalized statistical models, offering thus far underappreciated potential in statistical inference.
6. Popular financial risk measures such as the expected shortfall fail to be elicitable, posing critical challenges to scientists and regulators alike and calling for a transition to distributional forecasts.
7. The need for further development of distributional or density regression techniques in both parametric and nonparametric settings is pronounced.
8. Strong methodological ties between probabilistic forecasting, regression, and the emerging field of uncertainty quantification can be fruitfully explored and utilized.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank the Bank of England for permission to use **Figure 1**, Roman Schefzik for providing **Figures 4–6**, and Evgeni Ovcharov, Johanna Ziegel, and an anonymous reviewer for comments. Tilmann Gneiting acknowledges funding from the European Union Seventh Framework Program under grant agreement no. 290976.

LITERATURE CITED

Alkema L, Raftery AE, Clark SJ. 2007. Probabilistic projections of HIV prevalence using Bayesian melding. *Ann. Appl. Stat.* 1:229–48

Araújo MB, New M. 2006. Ensemble forecasting of species distributions. *Trends Ecol. Evol.* 22:42–47

Artzner P, Delbaen F, Eber J-M, Heath D. 1999. Coherent measures of risk. *Math. Finance* 9:203–28

Banerjee A, Guo X, Wang H. 2005. On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. Inf. Theory* 51:2664–69

Bank of England. 2013. *Inflation Report February 2013*. London: Bank of England. **http://www.bankofengland.co.uk/publications/Documents/inflationreport/2013/ir13feb.pdf**

Basu S, Markov S. 2004. Loss function assumptions in rational expectations tests on financial analysts' earnings forecasts. *J. Account. Econ.* 38:171–203

Ben Bouallègue Z. 2013. Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Weather Forecast.* 28:515–24

Bernardo JM. 1979. Expected information as expected utility. *Ann. Stat.* 7:686–90

Berrocal VJ, Raftery AE, Gneiting T. 2008. Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Ann. Appl. Stat.* 2:1170–93

Berrocal VJ, Raftery AE, Gneiting T, Steed RC. 2010. Probabilistic weather forecasting for winter road maintenance. *J. Am. Stat. Assoc.* 105:522–37

Birgé L, Massart P. 1993. Rates of convergence for minimum contrast estimators. *Probab. Theory Relat. Fields* 97:113–50

Bröcker J, Smith LA. 2008. From ensemble forecasts to predictive distribution functions. *Tellus A* 60:663–78

Brown BG, Katz RW, Murphy AH. 1984. Time series models to simulate and forecast wind speed and wind power. *J. Clim. Appl. Meteorol.* 23:1184–95

Buizza R, Houtekamer PL, Toth Z, Pellerin G, Wei M, Zhu Y. 2005. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Weather Rev.* 133:1076–97

Buizza R, Miller M, Palmer TN. 1999. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* 125:2887–908

Cannon AJ. 2012. Neural networks for probabilistic environmental prediction: conditional density estimation network creation and evaluation (CaDENCE) in R. *Comput. Geosci.* 41:126–35

Cawley GC, Janacek GJ, Haylock MR, Dorling SR. 2007. Predictive uncertainty in environmental modelling. *Neural Netw.* 20:537–49

Clark M, Gangopadhyay S, Hay L, Rajagopalan B, Wilby R. 2004. The Schaake shuffle: a method for reconstructing space-time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* 5:243–62

Cloke HL, Pappenberger F. 2009. Ensemble flood forecasting: a review. *J. Hydrol.* 375:613–26

Collins M, Knight S, eds. 2007. Theme Issue: Ensembles and probabilities: a new era in the prediction of climate change. *Philos. Trans. R. Soc. A* 365(1857)

Corradi V, Swanson NR. 2006. Predictive density evaluation. In *Handbook of Economic Forecasting*, ed. G Elliott, CWJ Granger, A Timmermann, pp. 197–284. Amsterdam: Elsevier

Cover T, Hart P. 1967. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13:21–27

Cressie NAC, Wikle CK. 2011. *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley

Czado C, Gneiting T, Held L. 2009. Predictive model assessment for count data. *Biometrics* 65:1254–61

D'Agostino A, McQuinn K, Whelan K. 2012. Are some forecasters really better than others? *J. Money Credit Bank.* 44:715–32

Dawid AP. 1984. Present position and potential developments: some personal views: statistical theory: the prequential approach. *J. R. Stat. Soc. A* 147:278–90

Dawid AP. 1986. Probability forecasting. In *Encyclopedia of Statistical Sciences*, Vol. 7, ed. S Kotz, NL Johnson, CB Read, pp. 210–18. New York: Wiley

Dawid AP. 2007. The geometry of proper scoring rules. *Ann. Inst. Stat. Math.* 59:77–93

Dawid AP, Sebastiani P. 1999. Coherent dispersion criteria for optimal experimental design. *Ann. Stat.* 27:65–81

Dette H, Volgushev S. 2008. Non-crossing non-parametric estimates of quantile curves. *J. R. Stat. Soc. B* 70:609–27

Diebold FX. 2012. *Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of Diebold–Mariano tests*. Natl. Bur. Econ. Res. Work. Pap. No. 18391, Natl. Bur. Econ. Res., Cambridge, MA. **http://www.nber.org/papers/w18391.pdf**

Diebold FX, Gunther TA, Tay AS. 1998. Evaluating density forecasts with applications to financial risk management. *Int. Econ. Rev.* 39:863–83

Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *J. Bus. Econ. Stat.* 13:253–63

Introduced the now commonly used Diebold–Mariano test for equal predictive performance.

Dunson DB, Pillai N, Park J-H. 2007. Bayesian density regression. *J. R. Stat. Soc. B* 69:163–83

Efron B. 1991. Regression percentiles using asymmetric squared error loss. *Stat. Sin.* 1:93–125

Ehm W, Gneiting T. 2012. Local proper scoring rules of order two. *Ann. Stat.* 40:609–37

Ferguson TS. 1967. *Mathematical Statistics: A Decision-Theoretic Approach*. New York: Academic

Flowerdew J. 2012. *Calibration and combination of medium-range ensemble precipitation forecasts*. Forec. Res. Tech. Rep. 567, Met Office, Exeter, UK. **http://www.metoffice.gov.uk/media/pdf/h/6/FRTR567.pdf**

Forbes PGM. 2012. Compatible weighted proper scoring rules. *Biometrika* 99:989–94
Fraley C, Raftery AE, Gneiting T. 2010. Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Weather Rev.* 138:190–202
Gel Y, Raftery AE, Gneiting T. 2004. Calibrated probabilistic mesoscale weather field forecasting: the geostatistical output perturbation (GOP) method (with discussion). *J. Am. Stat. Assoc.* 99:575–90
Genest C, Zidek JV. 1986. Combining probability distributions: a critique and an annotated bibliography. *Stat. Sci.* 1:114–35
Gentner D, Holyoak KJ. 1997. Reasoning and learning by analogy: introduction. *Am. Psychol.* 52:32–34
Geweke J, Amisano G. 2011. Optimal prediction pools. *J. Econom.* 164:130–41
Gigerenzer G, Hertwig R, Van Den Broeck E, Fasolo B, Katsikoupolos KV. 2005. "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Anal.* 25:623–29
Gneiting T. 2008. Editorial: probabilistic forecasting. *J. R. Stat. Soc. A* 171:319–21
Gneiting T. 2011a. Making and evaluating point forecasts. *J. Am. Stat. Assoc.* 106:746–62
Gneiting T. 2011b. Quantiles as optimal point forecasts. *Int. J. Forecast.* 27:197–207
Gneiting T. 2012. On the Cover–Hart inequality: What's a sample of size one worth? *Stat* 1:12–17
Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. B* 67:243–68
Gneiting T, Larson K, Westrick K, Genton MG, Aldrich E. 2006. Calibrated probabilistic forecasting at the Stateline wind energy center: the regime-switching space-time method. *J. Am. Stat. Assoc.* 101:968–79
Gneiting T, Raftery AE. 2005. Weather forecasting with ensemble methods. *Science* 310:248–49
Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* 102:359–78
Gneiting T, Raftery AE, Westveld AH III, Goldman T. 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* 133:1098–118
Gneiting T, Ranjan R. 2011. Comparing density forecasts using threshold- and quantile-weighted proper scoring rules. *J. Bus. Econ. Stat.* 29:411–22
Gneiting T, Ranjan R. 2013. Combining predictive distributions. *Electron. J. Stat.* 7:1747–82
Gneiting T, Stanberry LI, Grimit EP, Held L, Johnson NA. 2008. Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds. *Test* 17:211–64
Good IJ. 1952. Rational decisions. *J. R. Stat. Soc. B* 14:107–14
Granger CWJ, Newbold P. 1986. *Forecasting Economic Time Series*. San Diego, CA: Academic. 2nd ed.
Groen JJJ, Paap R, Ravazzolo F. 2013. Real-time inflation forecasting in a changing world. *J. Bus. Econ. Stat.* 31:29–44
Hall P, Racine J, Li Q. 2004. Cross-validation and the estimation of conditional probability densities. *J. Am. Stat. Assoc.* 99:1015–26
Hall P, Wolff RCL, Yao Q. 1999. Methods for estimating a conditional distribution function. *J. Am. Stat. Assoc.* 94:154–63
Hall SJ. 2011. Scientists on trial: at fault? *Nature* 477:264–69
Hamill TM. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Weather Rev.* 129:550–60
Hammond G. 2012. *State of the art of inflation targeting*. Cent. Cent. Bank. Stud. Handb. No. 29, Bank Engl., London, UK. **http://www.bankofengland.co.uk/education/Documents/ccbs/handbooks/pdf/ccbshb29.pdf**
Hendrickson AD, Buehler RJ. 1971. Proper scores for probability forecasters. *Ann. Math. Stat.* 42:1916–21
Hilden J, Gerds TA. 2013. A note on the evaluation of novel biomarkers: Do not rely on IDI and NRI. *Stat. Med.* In press. doi:10.1002/sim.5804
Holzmann H, Eulert M. 2013. *The role of the information set for forecasting—with applications to risk management*. Work. Pap., Dept. Math., Univ. Marburg, Marburg, Ger.
Hood L, Heath JR, Phelps ME, Lin B. 2004. Systems biology and new technologies enable predictive and preventative medicine. *Science* 306:640–43

A comprehensive review of methods for combining predictive distributions.

A comprehensive review article on the evaluation of point forecasts.

A review of proper scoring rules.

Hothorn T, Kneib T, Bühlmann P. 2013. Conditional transformation models. *J. R. Stat. Soc. B.* In press. doi:10.1111/rssb.12017

Hyndman RJ, Bashtannyk DM, Grunwald GK. 1996. Estimating and visualizing conditional densities. *J. Comp. Graph. Stat.* 5:315–36

Hyvärinen A. 2005. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* 6:695–709

Jeon J, Taylor JW. 2012. Using conditional kernel density estimation for wind power density forecasting. *J. Am. Stat. Assoc.* 107:66–79

Jones HE, Spiegelhalter DJ. 2012. Improved probabilistic prediction of healthcare performance indicators using bidirectional smoothing models. *J. R. Stat. Soc. A* 175:729–47

Jordan TH. 2013. The value, protocols, and scientific ethics of earthquake forecasting. *Geophys. Res. Abstr.* 15:EGU2013–12789

Jordan TH, Chen YT, Gasparini P, Madariaga R, Main I, et al. 2011. Operational earthquake forecasting: state of knowledge and guidelines for utilization. *Ann. Geophys.* 54:315–91

Kleiber W, Raftery AE, Baars J, Gneiting T, Mass CF, Grimit EP. 2011a. Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Mon. Weather Rev.* 139:2630–49

Kleiber W, Raftery AE, Gneiting T. 2011b. Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting. *J. Am. Stat. Assoc.* 106:1291–303

Kneib T. 2013. Beyond mean regression. *Stat. Model.* 13:275–303

Knüppel M. 2011. *Evaluating the calibration of multi-step ahead density forecasts using raw moments.* Deutsche Bundesbank Discuss. Pap. Ser. 1: Econ. Stud., Frankfurt, Ger. **http://econstor.eu/bitstream/10419/54982/1/684344750.pdf**

Koenker R. 2005. *Quantile Regression*. Cambridge, UK: Cambridge Univ. Press

Koenker R, Bassett G. 1978. Regression quantiles. *Econometrica* 46:33–50

Krüger F. 2013. *Jensen's inequality and the success of linear prediction pools*. Work. Pap., Dep. Econ., Univ. Konstanz, Konstanz, Ger.

Krzysztofowicz R. 2001. The case for probabilistic forecasting in hydrology. *J. Hydrol.* 249:2–9

Lambert NS, Pennock DM, Shoham Y. 2008. Eliciting properties of probability distributions. *Proc. 9th ACM Conf. Electron. Commer.*, *Chicago*, July 8–12, pp. 129–38. New York: ACM

Lawless JF. 1987. Negative binomial and mixed Poisson regression. *Can. J. Stat.* 15:209–25

Leutbecher M, Palmer TN. 2008. Ensemble forecasting. *J. Comp. Phys.* 227:3515–39

Lozano P, Wang H, Foreman KJ, Rajaratnam JK, Naghavi M, et al. 2011. Progress towards Millennium Development Goals 4 and 5 on maternal and child mortality: an updated systematic analysis. *Lancet* 378:1139–65

Masarotto G, Varin C. 2012. Gaussian copula marginal regression. *Electron. J. Stat.* 6:1517–49

Matheson JE, Winkler RL. 1976. Scoring rules for continuous probability distributions. *Manag. Sci.* 22:1087–96

McCullagh P, Nelder J. 1989. *Generalized Linear Models*. Boca Raton, FL: Chapman and Hall/CRC. 2nd edition

Möller A, Lenkoski A, Thorarinsdottir TL. 2013. Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Q. J. R. Meteorol. Soc.* 193:982–91

Molteni F, Buizza R, Palmer TN, Petroliagis T. 1996. The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.* 122:73–119

Montgomery JM, Hollenbach FM, Ward MD. 2012. Ensemble predictions of the 2012 US presidential election. *PS: Polit. Sci. Polit.* 45:651–54

Murphy AH, Winkler RL. 1987. A general framework for forecast verification. *Mon. Weather Rev.* 115:1330–38

Murphy AH, Winkler RL. 1992. Diagnostic verification of probability forecasts. *Int. J. Forecast.* 7:435–55

Nat. Publ. Group. 2012. Shock and law. *Nature* 490:446

Natl. Cent. Atmos. Res. Res. Appl. Program. 2010. *verification: Forecast Verification Utilities.* R package version 1.31. **http://CRAN.R-project.org/package=verification**

A seminal paper arguing for the consideration of the joint distribution of the forecast and the observation in evaluating predictive performance.

Newey WK, Powell JL. 1987. Asymmetric least squares estimation and testing. *Econometrica* 55:819–47

Osband KH. 1985. *Providing incentives for better cost forecasting*. PhD Thesis. Univ. Calif., Berkeley

Ovcharov E. 2013. *Multivariate local proper scoring rules*. Work. Pap., Inst. Appl. Math., Univ. Heidelberg, Heidelberg, Ger.

Palmer TN. 2002. The economic value of ensemble forecasts as a tool for risk assessment: from days to decades. *Q. J. R. Meteorol. Soc.* 128:747–74

Palmer TN. 2012. Towards the probabilistic Earth-system simulator: a vision for the future of climate and weather prediction. *Q. J. R. Meteorol. Soc.* 138:841–61

Parry M, Dawid AP, Lauritzen S. 2012. Proper local scoring rules. *Ann. Stat.* 40:561–92

Pers TH, Albrechtsen A, Holst C, Sørensen TIA, Gerds TA. 2009. The validation and assessment of machine learning: a game of prediction from high-dimensional data. *PLoS ONE* 4:e6287

Pfanzagl J. 1969. On the measurability and consistency of minimum contrast estimates. *Metrika* 14:249–72

Pinson P. 2012a. Adaptive calibration of (u, v)-wind ensemble forecasts. *Q. J. R. Meteorol. Soc.* 138:1273–84

Pinson P. 2012b. Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions. *J. R. Stat. Soc. C* 61:555–76

Pinson P. 2013. Wind energy: forecasting challenges for its operational management. *Stat. Sci.* In press

Pinson P, Girard R. 2012. Evaluating the quality of scenarios of short-term wind power generation. *Appl. Energy* 96:12–20

Pinson P, Madsen H, Nielsen HA, Papaefthymiou G, Klöckl B. 2009. From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy* 12:51–62

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* 133:1155–74

Raftery AE, Li N, Sevcíková H, Gerland P, Heilig GK. 2012. Bayesian probabilistic population projections for all countries. *Proc. Natl. Acad. Sci. USA* 109:13915–21

Ranjan R, Gneiting T. 2010. Combining probability forecasts. *J. R. Stat. Soc. B* 72:71–91

Richardson DS. 2000. Skill and economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* 126:649–68

Rockafellar RT. 1970. *Convex Analysis*. Princeton, NJ: Princeton Univ. Press

Rockafellar RT, Uryasev S. 2002. Conditional value-at-risk for general loss distributions. *J. Bank. Financ.* 26:1443–71

Roulin E, Vannitsem S. 2012. Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Mon. Weather Rev.* 140:874–88

Rüschendorf L. 2009. On the distributional transform, Sklar's theorem, and the empirical copula process. *J. Stat. Plan. Inference* 139:3921–27

Savage LJ. 1971. Elicitation of personal probabilities and expectations. *J. Am. Stat. Assoc.* 66:783–801

Schefzik R, Thorarinsdottir TL, Gneiting T. 2013. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.* In press

Schervish MJ. 1989. A general method for comparing probability assessors. *Ann. Stat.* 17:1856–79

Scheuerer M. 2013. Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Q. J. R. Meteorol. Soc.* In press. doi:10.1002/qj.2183

Schuhen N, Thorarinsdottir TL, Gneiting T. 2012. Ensemble model output statistics for wind vectors. *Mon. Weather Rev.* 140:3204–19

Sklar A. 1959. Fonctions de répartition à n dimensions et leur marges. *Publ. Inst. Stat. Univ. Paris* 8:229–31

Sloughter JM, Gneiting T, Raftery AE. 2010. Probabilistic wind forecasting using ensembles and Bayesian model averaging. *J. Am. Stat. Assoc.* 105:25–35

Sloughter JM, Raftery AE, Gneiting T, Fraley C. 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Weather Rev.* 135:3209–20

Sobotka F, Kneib T. 2012. Geoadditive expectile regression. *Comput. Stat. Data Anal.* 56:755–67

Spiegelhalter DJ, Pearson M, Short I. 2011. Visualizing uncertainty about the future. *Science* 333:1393–400

Stigler SM. 1975. The transition from point to distribution estimation. *Bull. Int. Stat. Inst.* 46:332–40

Stone CJ. 1977. Consistent nonparametric regression. *Ann. Math. Stat.* 32:1339–42

Stone M. 1961. The linear pool. *Ann. Math. Stat.* 32:1339–42

Well ahead of its time, this pioneering PhD thesis studied the notions of consistency and elicitability.

A well-argued plea for fully probabilistic weather and climate forecasts.

Derived the Bregman representation for scoring functions that are consistent for the expectation functional.

Thomson W. 1979. Eliciting production possibilities from a well-informed manager. *J. Econ. Theory* 20:360–80
Thorarinsdottir TL, Gneiting T. 2010. Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *J. R. Stat. Soc. A* 173:371–88
Timmermann A. 2000. Density forecasting in economics and finance. *J. Forecast.* 19:231–34
Traiteur JJ, Callicutt DJ, Smith M, Roy SB. 2012. A short-term ensemble wind speed forecasting system for wind power applications. *J. Appl. Meteorol. Climatol.* 51:1763–74
van Stiphout T, Wiemer S, Marzocchi W. 2010. Are short-term evacuations warranted? Case of the 2009 L'Aquila earthquake. *Geophys. Res. Lett.* 37:L06306
Wald A. 1949. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.* 20:595–601
Wallis KF. 2003. Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts. *Int. J. Forecast.* 19:165–75
Wilks DS. 2009. Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorol. Appl.* 16:361–68
Wilks DS, Hamill TM. 2007. Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Weather Rev.* 135:2379–90
Zhu X, Genton MG. 2012. Short-term wind speed forecasting for power system operations. *Int. Stat. Rev.* 80:2–23
Ziegel JF. 2013. *Coherence and elicitability*. Work. Pap., Dep. Math. Stat. Actuar. Sci., Univ. Bern, Bern, Switzerland. **http://arxiv.org/pdf/1303.1690v2.pdf**

RELATED RESOURCES

L'Aquila trial documentation: **http://processoaquila.wordpress.com/**

Mass C, Joslyn S, Pyle J, Tewson P, Gneiting T, et al. 2009. PROBCAST: A web-based portal to mesoscale probabilistic forecasts. *Bull. Am. Meteorol. Soc.* 90:1009–14

Probcast website: **http://www.probcast.com**

Silver N. 2012. *The Signal and the Noise: The Art and Science of Prediction*. New York: Penguin
In a recent *New York Times* best seller, Silver provides an exceptionally well-written popular account of forecasting across all realms of science and society, emphasizing the need for probabilistic thinking.
Tetlock PE. 2005. *Political Expert Judgement*. Princeton, NJ: Princeton Univ. Press
In a well-acclaimed and important study, Tetlock evaluates thousands of expert probability forecasts for future economic, political, and societal events.