Electronic Journal of Statistics
Vol. 7 (2013) 1747–1782
ISSN: 1935-7524
DOI: 10.1214/13-EJS823

Combining predictive distributions*

Tilmann Gneiting

Institute for Applied Mathematics, Heidelberg University, Germany
e-mail: t.gneiting@uni-heidelberg.de

and

Roopesh Ranjan

Mzaya Private Limited, Bangalore, India
e-mail: roopesh.ranjan@gmail.com

Abstract: In probabilistic forecasting combination formulas for the aggregation of predictive distributions need to be estimated based on past experience and training data. We study combination formulas and aggregation methods for predictive cumulative distribution functions from the perspectives of calibration and dispersion, taking an original prediction space approach that applies to discrete, mixed discrete-continuous and continuous predictive distributions alike. The key idea is that aggregation methods ought to be parsimonious, yet sufficiently flexible to accommodate any type of dispersion in the component distributions. Both linear and non-linear aggregation methods are investigated, including generalized, spread-adjusted and beta-transformed linear pools. The effects and techniques are demonstrated theoretically, in simulation examples, and in case studies, where we fit combination formulas for density forecasts of S&P 500 returns and daily maximum temperature at Seattle-Tacoma Airport.

AMS 2000 subject classifications: Primary 62; secondary 91B06.
Keywords and phrases: Beta transform, conditional calibration, density forecast, flexibly dispersive, forecast aggregation, linear pool, probability integral transform, probabilistic calibration.

Received February 2013.

Contents

*This research has been supported by the United States National Science Foundation under Awards ATM-0724721 and DMS-0706745 to the University of Washington, and by the European Union Seventh Framework Programme under grant agreement no. 290976 to Heidelberg University. Part of it was done at the Newton Institute in Cambridge. We are grateful to Cliff Mass and Jeff Baars for providing the temperature data, and to Hajo Holzmann, Don Percival, Adrian Raftery, Michael Scheuerer, Thordis Thorarinsdottir, Alexander Tsyplakov, Bob Winkler and Johanna Ziegel for discussions and preprints. Of course, the usual disclaimer applies.

1. Introduction

Probabilistic forecasts aim to provide calibrated and sharp predictive distributions for future quantities or events of interest. As they admit the assessment of forecast uncertainty and allow for optimal decision making, probabilistic forecasts continue to gain prominence in a wealth of applications, ranging from economics and finance to meteorology and climatology (Gneiting, 2008). The general goal is to maximize the sharpness of the predictive distributions subject to calibration (Murphy and Winkler, 1987; Gneiting et al., 2007). For a real-valued outcome, a probabilistic forecast can be represented in the form of a predictive cumulative distribution function, which might be discrete, mixed discrete-continuous or continuous, with the latter case corresponding to density forecasts.

In many situations, complementary or competing probabilistic forecasts from dependent or independent information sources are available. For example, the individual forecasts might stem from distinct experts, organizations, or statistical models. The prevalent method for aggregating the individual predictive distributions into a single combined forecast is the linear pool (Stone, 1961). While other methods for combining predictive distributions are available (Genest and Zidek, 1986; Clemen and Winkler, 2007), the linear pool is typically the method of choice, with the pioneering work of Winkler (1968) and Zarnowitz (1969), and recent papers by Mitchell and Hall (2004), Wallis (2005), Hall and Mitchell (2007), Jore et al. (2010), Kascha and Ravazzolo (2007) and Garratt et al. (2011) being examples in the case of density forecasts. Similarly, linear pools have been applied successfully to combine discrete predictive distributions; for recent reviews, see Ranjan and Gneiting (2010), Clements and Harvey (2011) and Allard et al. (2012).

In practice, combination formulas need to be estimated based on past experience and training data. To fix the idea, we consider a general real-valued outcome, so that the training data are of the form

$$\{(F_{1j}, \ldots, F_{kj}, y_j) : j = 1, \ldots, J\}, \tag{1}$$

where there are J cases in the training set, with the first k arguments denoting the individual predictive cumulative distribution functions, and the final argument the realizing observation. To aggregate the individual predictive cumulative distribution functions in out-of-sample cases, one specifies an *aggregation method*, that is, a family $\mathcal{G} = \{G_\theta : \theta \in \Theta\}$ of *combination formulas* of the form

$$G_\theta : \mathcal{F}^k = \mathcal{F} \times \cdots \times \mathcal{F} \to \mathcal{F}, \qquad (F_1, \ldots, F_k) \mapsto G_\theta(F_1, \ldots, F_k).$$

where $\mathcal{F}$ is a suitable class of cumulative distribution functions. For example, if $\mathcal{G}$ is the traditional linear pool, we can take $\mathcal{F}$ to be any convex class of cumulative distribution functions, and we may identify the index set Θ with the unit simplex in $\mathbb{R}^k$. The goal then is to estimate an optimal combination formula based on training data of the form (1).

Despite the ubiquitous success of the linear pool in a vast number of applications, for which Krüger (2013) provides an appealing partial explanation, fragmented recent work points at shortcomings and limitations. Hora (2004) and Ranjan and Gneiting (2010) showed in special cases that if each of the individual predictive distributions is calibrated, any nontrivial linear combination is necessarily uncalibrated. As calibration is a critical requirement for a probabilistic forecast to be practically useful (Dawid, 1984; Diebold et al., 1998), these results suggest that linear pooling might be suboptimal, in that nonlinear combination formulas might outperform linear methods, as demonstrated empirically by Ranjan and Gneiting (2010) and Allard et al. (2012).

Our initial goal here is to unify and extend the aforementioned results. Towards this end, we develop novel theoretical approaches to studying combination formulas and aggregation methods. Technically, we operate in terms of cumulative distribution functions, which permits a unified treatment of all real-valued predictands, including the cases of density forecasts, mixed discrete-continuous predictive distributions, probability mass functions for count data, and probability forecasts of a dichotomous event. The extant literature compares combination formulas by examining whether or not they possess certain analytic characteristics, such as the strong setwise function and external Bayes properties (Genest and Zidek, 1986; French and Ríos Insua, 2000). In contrast to the earlier work, we assess combination formulas and aggregation methods from the perspectives of calibration and dispersion.

Section 2 sets the stage by introducing the key tool of a prediction space, which is a probability space tailored to the study of forecasts and combination formulas. In this framework, training data of the form (1) are interpreted as a sample from the underlying joint distribution of the forecasts and the observations. We revisit the work of Gneiting et al. (2007) and Ranjan and Gneiting (2010) in the prediction space setting and show, perhaps surprisingly, that if the outcome is binary, conditional calibration is equivalent to probabilistic calibration. Section 3 is devoted to the study of specific, linear and non-linear combination formulas and aggregation methods. A major result is, roughly, that dispersion tends to increase under linear pooling. This helps explain the success of linear combination formulas in aggregating underdispersed component

distributions, and allows us to show that the traditional linear pool fails to be flexibly dispersive. Parsimonious nonlinear alternatives include generalized linear pools, the spread-adjusted linear pool, which has been used successfully in meteorological applications, and the beta-transformed linear pool proposed by Ranjan and Gneiting (2010), which we demonstrate to be flexibly dispersive. Section 4 turns to a simulation study and data examples, where we fit combination formulas for aggregating density forecasts of S&P 500 returns and daily maximum temperature at Seattle-Tacoma Airport. The paper ends in Section 5, where we discuss our findings and suggest directions for future work.

2. Prediction spaces, combination formulas and aggregation methods

In a seminal paper, Murphy and Winkler (1987) proposed a general framework for the evaluation of point forecasts, which is based on the joint distribution of the forecast and the observation. Dawid et al. (1995) developed and used a related framework in studying multiple probability forecasts for a binary event. Here we respond to the call of Dawid et al. (1995, p. 28) for an extension, and we start with an informal sketch of a fully general approach, in which the observations take values in just any space.

The most general setting considers the joint distribution of multiple probabilistic forecasts and the observation on a probability space $(\Omega, \mathcal{A}, \mathbb{Q})$. More explicitly, we assume that the elements of the sample space Ω can be identified with tuples of the form

$$(P_1, \ldots, P_k, Y),$$

where each of $P_1, \ldots, P_k$ is a probability measure on the outcome space of the observation, Y. For $i = 1, \ldots, k$, we require the random probability measure P_i to be measurable with respect to the sub-σ-algebra $\mathcal{A}_i \subseteq \mathcal{A}$ that encodes the forecast's information set or information basis, consisting of data, expertise, theories and assumptions at hand. The probability measure $\mathbb{Q}$ on $(\Omega, \mathcal{A})$ specifies the joint distribution of the probabilistic forecasts and the observation.

In this setting, the probabilistic forecasts $P_1, \ldots, P_k$ might stem from distinct experts, organizations or statistical models, as commonly encountered in the practice of forecasting. In aggregating them, the ideal strategy is to combine information sets, that is, to issue the conditional distribution of the observation Y given the σ-algebra $\sigma(\mathcal{A}_1, \ldots, \mathcal{A}_k)$ generated by the information sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$. However, as Dawid et al. (1995, p. 264) note,

> "this ideal will almost always be rendered unattainable, by the extent of the data, company confidentiality, or the inability of the experts to identify clearly the empirical basis and background knowledge leading to their intuitive opinions."

The best that we can hope for in practice is to find the conditional distribution of the observation Y given the σ-algebra $\sigma(P_1, \ldots, P_k)$ generated by the random

probability measures $P_1, \ldots, P_k$. Of course, it is always true that

$$\sigma(P_1, \ldots, P_k) \subseteq \sigma(\mathcal{A}_1, \ldots, \mathcal{A}_k),$$

and in most cases of practical interest the left-hand side constitutes a substantially lesser information basis than the right-hand side.

2.1. Prediction spaces

In what follows, we restrict the discussion to the case of a real-valued observation. A probabilistic forecast then corresponds to a Lebesgue-Stieltjes measure on the real line, $\mathbb{R}$, which we identify with the associated right-continuous cumulative distribution function (CDF). We use the symbol $\mathcal{L}$ generically to denote an unconditional or conditional law or distribution and follow standard conventions in identifying the sub-σ-algebras on which we condition. In particular, we write $\sigma(\mathcal{A}_1, \ldots, \mathcal{A}_m)$ and $\sigma(X_1, \ldots, X_n)$ to denote the σ-algebra generated by the families $\mathcal{A}_1, \ldots, \mathcal{A}_m$ of subsets of Ω, and the random variables $X_1, \ldots, X_n$, respectively.

We now introduce the key tool of a prediction space, which is a probability space tailored to the study of combination formulas for real-valued outcomes, though we allow the case $k = 1$ of a single probabilistic forecast.

Definition 2.1. Let $k \geq 1$ be an integer. A *prediction space* is a probability space $(\Omega, \mathcal{A}, \mathbb{Q})$ together with sub-σ-algebras $\mathcal{A}_1, \ldots, \mathcal{A}_k \subseteq \mathcal{A}$, where the elements of the sample space Ω can be identified with tuples $(F_1, \ldots, F_k, Y, V)$ such that

(P1) for $i = 1, \ldots, k$, F_i is a CDF-valued random quantity that is measurable with respect to the sub-σ-algebra $\mathcal{A}_i$,[1]
(P2) Y is a real-valued random variable,
(P3) V is a random variable that is uniformly distributed on the unit interval and independent of $\mathcal{A}_1, \ldots, \mathcal{A}_k$ and Y.

All subsequent definitions and results are within the prediction space setting. Phrases such as *almost surely* or *with positive probability* refer to the probability measure $\mathbb{Q}$ on $(\Omega, \mathcal{A})$ that determines the joint distribution of the probabilistic forecasts and the observations. While (P1) and (P2) formalize the predictive distributions and the observation, assumption (P3) is purely technical, allowing us to define a generalized version of the classical probability integral transform. The sub-σ-algebra $\mathcal{A}_i$ encodes the information set for the CDF-valued random quantity F_i which may, but need not, be ideal in the following sense.[2]

Definition 2.2. The CDF-valued random quantity F_i is *ideal* relative to the sub-σ-algebra $\mathcal{A}_i$ if $F_i = \mathcal{L}(Y \,|\, \mathcal{A}_i)$ almost surely.

[1] That is, $\{F_i(x_j) \in B_j \text{ for } j = 1, \ldots, n\} \in \mathcal{A}$ for all finite collections $x_1, \ldots, x_n$ of real numbers and $B_1, \ldots, B_n$ of Borel sets.

[2] In independent work, Tsyplakov (2011, 2013) proposes the same terminology.

The subsequent examples serve to illustrate the notions of prediction spaces and ideal forecasts. We write $\mathcal{N}(\mu, \sigma^2)$ for the univariate normal distribution with mean μ and variance σ^2, and we use the symbols Φ and ϕ to denote the standard normal cumulative distribution function and density function, respectively.

Example 2.3 (probability forecasts of a binary event)**.** We consider a slight generalization of the simulation experiment of Ranjan and Gneiting (2010). In this setting, three forecasters issue the probability forecasts

$$p_1 = \Phi\left(\frac{\omega_1}{\sqrt{1+\sigma_2^2}}\right), \quad p_2 = \Phi\left(\frac{\omega_2}{\sqrt{1+\sigma_1^2}}\right) \quad \text{and} \quad p_3 = \Phi(\omega_1 + \omega_2) \tag{2}$$

for a binary event with success probability $\Phi(\omega_1 + \omega_2)$, where ω_1 and ω_2 are independent normal random variables with mean zero and variance σ_1^2 and σ_2^2.

To construct a suitable prediction space, let

$$\Omega = \mathbb{R} \times \mathbb{R} \times \{0, 1\} \times (0, 1),$$

write $\omega = (\omega_1, \omega_2, \omega_3, \omega_4) \in \Omega$ for an elementary event, and let $\mathcal{A}$ be the corresponding Borel-σ-algebra. We define $\mathbb{Q}$ to be the product of $\mathcal{N}(0, \sigma_1^2)$, $\mathcal{N}(0, \sigma_2^2)$ and a standard uniform measure on the first, second and fourth coordinate projections, respectively, and let

$$\mathbb{Q}(B_1 \times B_2 \times \{1\} \times (0,1)) = \frac{1}{\sigma_1 \sigma_2} \int_{B_1} \int_{B_2} \Phi(\omega_1 + \omega_2)\, \phi\left(\frac{\omega_1}{\sigma_1}\right) \phi\left(\frac{\omega_2}{\sigma_2}\right) d\lambda(\omega_1)\, d\lambda(\omega_2)$$

for Borel sets $B_1, B_2 \subseteq \mathbb{R}$, where λ denotes the Lebesgue measure. Having defined the triple $(\Omega, \mathcal{A}, \mathbb{Q})$, we construct the probability forecasts $p_1(\omega)$, $p_2(\omega)$ and $p_3(\omega)$ as in (2), and define the observation $Y(\omega) = \omega_3$ as well as the auxiliary variable $V(\omega) = \omega_4$. The CDF-valued random quantities F_i then are of the form $F_i(y) = (1 - p_i)\, \mathbb{1}(y \geq 0) + p_i\, \mathbb{1}(y \geq 1)$ for $i = 1, 2$ and 3, where F_1 is measurable with respect to the sub-σ-algebra $\mathcal{A}_1 = \sigma(\omega_1)$, F_2 is measurable with respect to $\mathcal{A}_2 = \sigma(\omega_2)$ and F_3 is measurable with respect to $\mathcal{A}_3 = \sigma(\omega_1, \omega_2)$. Moreover, F_1 is ideal relative to $\mathcal{A}_1$, F_2 is ideal relative to $\mathcal{A}_2$ and F_3 is ideal relative to $\mathcal{A}_3$.

Typically, it suffices to consider the joint distribution of the tuple $(F_1, \ldots, F_k, Y)$, without any need to explicitly specify other facets of the prediction space, as illustrated in the following example.

Example 2.4 (density forecasts)**.** To define a prediction space, let

$$Y \mid \mu \sim \mathcal{N}(\mu, 1) \quad \text{where} \quad \mu \sim \mathcal{N}(0, 1).$$

In this simplified version of the simulation example in Gneiting et al. (2007), the perfect forecast $F_1 = \mathcal{N}(\mu, 1)$ is ideal relative to the sub-σ-algebra generated by the random variable μ. The climatological forecast $F_2 = \mathcal{N}(0, 2)$ is ideal relative to the trivial sub-σ-algebra.

Readers interested in further examples might wish to look ahead to Section 4.1, where we describe a regression setting, in which the density forecasts are ideal relative to sub-σ-algebras that represent both public and proprietary information.

2.2. Calibration and dispersion

If F denotes a fixed, non-random predictive cumulative distribution function for an observation Y, the probability integral transform is the random variable $Z_F = F(Y)$. It is well known that if F is continuous and $Y \sim F$ then Z_F is standard uniform (Rosenblatt, 1952). If the more general, randomized version of the probability integral transform studied by Rüschendorf (1981) is used, the uniformity result applies to arbitrary, not necessarily continuous, but still fixed, non-random cumulative distribution functions.

In the prediction space setting, we need the following, further extension that allows for F to be a CDF-valued random quantity.

Definition 2.5. In the prediction space setting, the random variable

$$Z_F = \lim_{y \uparrow Y} F(y) + V\left(F(Y) - \lim_{y \uparrow Y} F(y)\right)$$

is the *probability integral transform* of the CDF-valued random quantity F.

In a nutshell, the probability integral transform is the value that the predictive cumulative distribution function attains at the observation, with suitable adaptations at any points of discontinuity. The probability integral transform takes values in the unit interval, and so the possible values of its variance are constrained to the closed interval $[0, \frac{1}{4}]$. A variance of $\frac{1}{12}$ corresponds to a uniform distribution and continues to be the most desirable, as evidenced by Theorem 2.8 below.

We are now ready to define and study notions of calibration and dispersion. In doing so, we use the terms CDF-valued random quantity and forecast interchangeably.

Definition 2.6. In the prediction space setting, let F and G be CDF-valued random quantities with probability integral transforms Z_F and Z_G.

(a) The forecast F is *marginally calibrated* if $\mathbb{E}_{\mathbb{Q}}[F(y)] = \mathbb{Q}(Y \leq y)$ for all $y \in \mathbb{R}$.
(b) The forecast F is *probabilistically calibrated* if its probability integral transform Z_F is uniformly distributed on the unit interval.
(c) The forecast F is *overdispersed* if $\mathrm{var}(Z_F) < \frac{1}{12}$, *neutrally dispersed* if $\mathrm{var}(Z_F) = \frac{1}{12}$, and *underdispersed* if $\mathrm{var}(Z_F) > \frac{1}{12}$.
(d) The forecast F is *at least as dispersed* as the forecast G if $\mathrm{var}(Z_F) \leq \mathrm{var}(Z_G)$. It is *more dispersed* than G if $\mathrm{var}(Z_F) < \mathrm{var}(Z_G)$.
(e) The forecast F is *regular* if the support of the distribution of Z_F is the unit interval.

Dawid (1984), Diebold et al. (1998), Gneiting et al. (2007) and Czado et al. (2009), among others, have argued powerfully that marginal and probabilistic calibration are critical requirements for a probabilistic forecast to be practically useful. In the defining equality $\mathbb{E}_{\mathbb{Q}}[F(y)] = \mathbb{Q}(Y \le y)$ for marginal calibration, the left-hand side depends on the law of the predictive distribution, whereas the right-hand side depends on the law of the observation. In parts (c) and (d) of Definition 2.6, we define dispersion in terms of the variance of the probability integral transform, thus involving the joint law of the predictive distribution and the observation. In contrast, the spread of the predictive distribution itself is a measure of sharpness that does not consider the observation.

Our current setting of prediction spaces differs from, but relates closely to, the approach of Gneiting et al. (2007), who studied notions of calibration from a prequential perspective. Specifically, if the CDF-valued random quantity F is probabilistically calibrated in the sense of Definition 2.6 and we sample from the joint law of F and Y, the resulting sequence is probabilistically calibrated in the sense of Gneiting et al. (2007). An analogous statement applies to marginal calibration.

Returning to the prediction space setting, the following result is immediate.

Proposition 2.7. *A probabilistically calibrated forecast is neutrally dispersed and regular.*

The converse is not necessarily true, in that a forecast which is neutrally dispersed need not be calibrated nor regular. However, an ideal forecast is always calibrated.

Theorem 2.8. *A forecast that is ideal relative to a σ-algebra is both marginally calibrated and probabilistically calibrated.*

Proof. Suppose that $F = \mathcal{L}(Y \,|\, \mathcal{A}_0)$ is ideal relative to the σ-algebra $\mathcal{A}_0$, so that $F(y) = \mathbb{Q}(Y \le y \,|\, \mathcal{A}_0)$ almost surely for all $y \in \mathbb{R}$. Then

$$\mathbb{E}_{\mathbb{Q}}[F(y)] = \mathbb{E}_{\mathbb{Q}}[\mathbb{Q}(Y \le y \,|\, \mathcal{A}_0)] = \mathbb{E}_{\mathbb{Q}}\, \mathbb{E}_{\mathbb{Q}}\, [\, \mathbb{1}(Y \le y) \,|\, \mathcal{A}_0] = \mathbb{Q}(Y \le y),$$

where $\mathbb{1}$ denotes an indicator function, thereby proving the statement about marginal calibration. Turning to probabilistic calibration, let $\mathbb{Q}_0$ denote the marginal law of Y under $\mathbb{Q}$, so that $Z_F = \mathbb{Q}_0((-\infty, Y) \,|\, \mathcal{A}_0) + V\, \mathbb{Q}_0(\{Y\} \,|\, \mathcal{A}_0)$ and

$$\mathbb{Q}(Z_F \le z) = \mathbb{E}_{\mathbb{Q}}\, \mathbb{E}_{\mathbb{Q}}\, [\, \mathbb{1}(Z_F \le z) \,|\, \mathcal{A}_0] = z$$

for $z \in (0, 1)$, where the final equality uses the uniformity of the traditional randomized probability integral transform, as proved in Lemma 3 of Rüschendorf (1981). □

In the special case of marginal calibration and discrete predictive distributions, the statement of Theorem 2.8 is due to Bröcker (2009, Appendix B). An interesting open question is whether there are any forecasts that are both probabilistically calibrated and marginally calibrated, but are not ideal. While we conjecture that the answer is in the positive, we do not know of any such examples.

TABLE 1
Probabilistic forecasts in Examples 2.4 and 2.9. The observation Y is normal with mean μ and variance 1, where μ is standard normal. The random variable τ attains the values -1 and 1 with probability $\frac{1}{2}$, independently of μ and Y

Forecast	Predictive Distribution	Marginally Calibrated	Probabilistically Calibrated	Ideal
Perfect	$F_1 = \mathcal{N}(\mu, 1)$	Yes	Yes	Yes
Climatological	$F_2 = \mathcal{N}(0, 2)$	Yes	Yes	Yes
Unfocused	$F_3 = \frac{1}{2}(\mathcal{N}(\mu, 1) + \mathcal{N}(\mu + \tau, 1))$	No	Yes	No
Sign-reversed	$F_4 = \mathcal{N}(-\mu, 1)$	Yes	No	No

We now revisit and extend Example 2.4.

Example 2.9. Let

$$Y \mid \mu \sim \mathcal{N}(\mu, 1) \quad \text{where} \quad \mu \sim \mathcal{N}(0, 1),$$

and let τ attain the values 1 and -1 with equal probability, independently of μ and Y. Table 1 places the density forecasts in the simulation example of Gneiting et al. (2007) in this setting. By Example 2.4 the perfect forecast and the climatological forecast are ideal, and so by Theorem 2.8 they are both probabilistically calibrated and marginally calibrated. Arguments nearly identical to those in Gneiting et al. (2007) show that the unfocused forecast is probabilistically calibrated but not marginally calibrated, and that the sign-biased forecast is marginally calibrated but not probabilistically calibrated. Hence, there is no sub-σ-algebra or information set relative to which the unfocused or the sign-biased forecast is ideal.

As noted, probabilistic calibration is a critical requirement for probabilistic forecasts that take the form of predictive cumulative distribution functions, with Theorem 2.8 lending further support to this approach. Indeed, checks for the uniformity of the probability integral transform have formed a cornerstone of density forecast evaluation. In practice, one observes a sample from the joint distribution $\mathbb{Q}$ of the probabilistic forecasts and the observation, and the uniformity of the probability integral transform is assessed empirically. The prevalent way of doing this is by plotting histograms of the probability integral transform values for the various forecasting methods, which show the corresponding frequency distribution over an evaluation or test set. U-shaped histograms correspond to underdispersed predictive distributions with prediction intervals that are too narrow on average, while hump or inverse U-shaped histograms indicate overdispersed predictive distributions.

Example 2.10. Let $Y = X + \epsilon$, where X and ϵ are independent, standard normal random variables, and consider the Gaussian predictive distribution $F_\sigma = \mathcal{N}(X, \sigma^2)$. A stochastic domination argument, the details of which we give in Appendix A, shows that F_σ is underdispersed if $\sigma < 1$, neutrally dispersed if $\sigma = 1$ and overdispersed if $\sigma > 1$. If $\sigma = 1$ then F_σ is ideal and thus

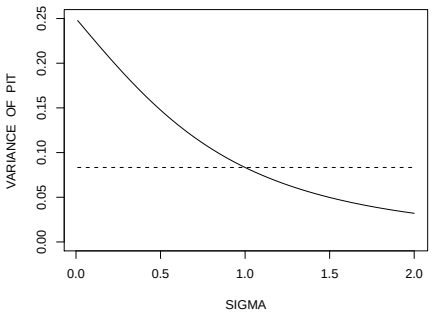

FIG 1. *The variance (3) of the probability integral transform $Z_\sigma = F_\sigma(Y)$ for the predictive distribution F_σ in Example 2.10 as a function of the predictive standard deviation, σ. The dashed horizontal line at $\frac{1}{12}$ indicates a neutrally dispersed forecast.*

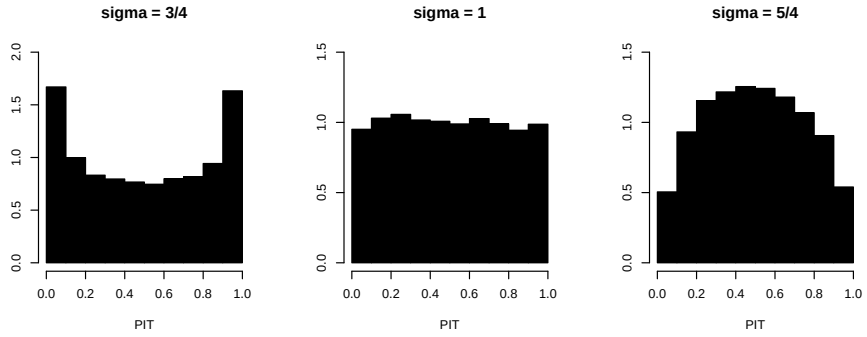

FIG 2. *Probability integral transform histograms for the predictive distribution F_σ in Example 2.10, where $\sigma = \frac{3}{4}$ (underdispersed), $\sigma = 1$ (neutrally dispersed and calibrated) and $\sigma = \frac{5}{4}$ (overdispersed).*

both marginally calibrated and probabilistically calibrated. A more detailed calculation, which is also given in Appendix A, shows that the probability integral transform $Z_\sigma = F_\sigma(Y)$ satisfies

$$\operatorname{var}(Z_\sigma) = 2\int_0^1 z\left(1 - \Phi(\sigma(\Phi^{-1}(z)))\right) \mathrm{d}z - \left(\int_0^1 \left(1 - \Phi(\sigma(\Phi^{-1}(z)))\right) \mathrm{d}z\right)^2. \tag{3}$$

In Figure 1 we plot $\operatorname{var}(Z_\sigma)$ as a function of the predictive standard deviation, σ. Figure 2 shows probability integral transform histograms for a Monte Carlo sample of size $10,000$ from the joint distribution of the observation Y and the forecasts F_σ, where $\sigma = \frac{3}{4}$, 1 and $\frac{5}{4}$. The histograms are U-shaped, uniform, and inverse U-shaped, reflecting underdispersion, neutral dispersion and calibration, and overdispersion, respectively.

In the case of a binary outcome Y, we identify a CDF-valued random quantity $F(y) = (1-p)\,\mathbb{1}(y \geq 0) + p\,\mathbb{1}(y \geq 1)$ with the probability forecast p for a success, that is, $Y = 1$. The extant literature, including Schervish (1989) and Ranjan and Gneiting (2010) and the references therein, calls p calibrated if

$$\mathbb{Q}(Y = 1 \,|\, p) = p \quad \text{almost surely.} \tag{4}$$

Here we refer to this natural property as *conditional calibration*. Perhaps surprisingly, our next result shows that if the outcome is binary, the notions of conditional calibration and probabilistic calibration are equivalent. Thus, the general notion of probabilistic calibration nests the traditional concept of conditional calibration.

Theorem 2.11. *Consider a prediction space $(\Omega, \mathcal{A}, \mathbb{Q})$ with a binary outcome Y, where $Y = 1$ corresponds to a success and $Y = 0$ to a failure, and a CDF-valued random quantity $F(y) = (1-p)\,\mathbb{1}(y \geq 0) + p\,\mathbb{1}(y \geq 1)$, which can be identified with the probability forecast p for a success. Then the following statements are equivalent:*

(i) *The probability forecast p is conditionally calibrated, that is, $\mathbb{Q}(Y = 1 \,|\, p) = p$ almost surely.*
(ii) *The forecast F is probabilistically calibrated, that is, its probability integral transform Z_F is uniformly distributed on the unit interval.*
(iii) *The forecast F is ideal relative to the σ-algebra generated by the probability forecast p.*

Proof. It is clear that (i) and (iii) are equivalent, and by Theorem 2.8 the statement (iii) implies (ii). To conclude the proof, we show that statement (ii) implies (i). To this end, suppose that the forecast F is probabilistically calibrated. By standard properties of conditional expectations, there exists a measurable function $q : [0,1] \to [0,1]$ such that $\mathbb{Q}(Y = 1 \,|\, p) = q(p)$ almost surely. Let H denote the marginal law of p under $\mathbb{Q}$. If H has a point mass at 0 or 1, it is readily seen that $q(0) = 0$ or $q(1) = 1$, respectively.

A version of the conditional density $u(z|x)$ of the probability integral transform Z_F given that $p = x \in [0,1]$ satisfies $u(z|x) = (1 - q(x))/(1-x)$ for $z \in [0, 1-x]$ and $u(z|x) = q(x)/x$ for $z \in (1-x, 1]$. The marginal density u of Z_F is standard uniform, so that

$$\begin{aligned} u(z+\delta) - u(z) &= \int_{[0,1]} \Big(u(z+\delta|x) - u(z|x) \Big)\, \mathrm{d}H(x) \\ &= \int_{(1-z-\delta, 1-z]} \frac{q(x) - x}{x(1-x)}\, \mathrm{d}H(x) = 0 \end{aligned}$$

for Lebesgue-almost all $z \in (0,1)$ and $\delta \in (0, 1-y)$. Let $0 < a < b < 1$, and consider the signed measure defined by

$$\mu(A) = \int_A \frac{q(x) - x}{x(1-x)}\, \mathrm{d}H(x)$$

TABLE 2
Example of a probabilistically calibrated, but not auto-calibrated CDF-valued random quantity F for a ternary outcome Y

$\mathbb{Q}$-probability	$F(x)$				$\mathbb{Q}(Y = i \mid F)$		
	$x < 0$	$0 \le x < 1$	$1 \le x < 2$	$x \ge 2$	$i = 0$	$i = 1$	$i = 2$
$\frac{1}{2}$	0	$\frac{1}{2}$	1	1	$\frac{3}{4}$	$\frac{1}{4}$	0
$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$

for Borel sets $A \subseteq [a, b]$. We have just shown that $\mu(A) = 0$ for all intervals $(c, d] \subseteq [a, b]$, except possibly for c or d in a Lebesgue null set. Since the family of such intervals generates the Borel-σ-algebra, this is only possible if μ is the null measure, so that $\mu(A) = 0$ for all Borel sets A in $[a, b]$. In particular, $\mu(A) = 0$ for $A = \{x \in [a, b] : q(x) > x\}$ and $A = \{x \in [a, b] : q(x) < x\}$, which implies that $q(x) = x$ almost surely with respect to the restriction of H to $[a, b]$. To summarize, we have shown that $q(x) = x$ almost surely with respect to H, whence $\mathbb{Q}(Y = 1 \mid p) = p$ almost surely with respect to $\mathbb{Q}$, as desired. □

Theorem 2.11 draws a connection from the probability integral transform histogram to the *reliability diagram* or *calibration curve*, which is the key diagnostic tool for assessing the calibration of probability forecasts for a binary event (Dawid, 1986; Murphy and Winkler, 1992; Ranjan and Gneiting, 2010). A reliability diagram plots conditional event frequencies against binned forecast probabilities, with deviations from the diagonal indicating violations of the conditional calibration condition (4).

For non-binary outcomes Y and the natural generalization of the conditional calibration criterion, namely the *auto-calibration* property

$$\mathcal{L}(Y \mid F) = F \quad \mathbb{Q}\text{-almost surely}$$

introduced by Tsyplakov (2011, 2013), the equivalence to probabilistic calibration fails, as demonstrated in Table 2 for a ternary outcome. For general real-valued outcomes, auto-calibration implies both probabilistic and marginal calibration, while probabilistic calibration and marginal calibration are logically independent of each other, as illustrated in Table 1. Empirical tests of auto-calibration are unlikely to be feasible, except for very special circumstances, when forecasters constrain themselves to providing a small number of distinct predictive distributions only, or when attention focuses on certain distributional features, with some of these facets having been explored by Hamill (2001), Mason et al. (2007) and Held et al. (2010).

Generally, probabilistic calibration continues to be the most practically useful and most practically relevant notion of calibration. It is possible for a forecast to be probabilistically calibrated but not marginally calibrated, as we have seen, and probabilistic calibration may be sufficient in many situations. In other settings, such as climate prediction, ideas closely related to marginal calibration play crucial roles, as recently emphasized by DelSole and Shukla (2010), Arnold et al. (2013) and Fricker et al. (2013).

TABLE 3
Some classes of fixed, non-random cumulative distribution functions, where the subscript refers to an interval $\mathrm{I} \subseteq \mathbb{R}$. *In the case of Bernoulli measures, we identify a success with 1 and a non-success with 0, so that the corresponding cumulative distribution function has jump discontinuities at these values, and otherwise is constant*

Class	Characterization of the Members
$\mathcal{F}_{\mathrm{I}}$	support in I
$\mathcal{F}_{\mathrm{I}}^{+}$	support in I; strictly increasing on I
$\mathcal{C}_{\mathrm{I}}$	support in I; continuous
$\mathcal{C}_{\mathrm{I}}^{+}$	support in I; continuous; strictly increasing on I
$\mathcal{D}_{\mathrm{I}}$	support in I; admits Lebesgue density
$\mathcal{D}_{\mathrm{I}}^{+}$	support in I; admits Lebesgue density; strictly increasing on I
$\mathcal{B}$	Bernoulli measure
$\mathcal{B}^{+}$	Bernoulli measure with nondegenerate success probability

2.3. Combination formulas and aggregation methods

As noted, in aggregating predictive cumulative distribution functions, the ideal strategy is to combine information sets, that is, to issue the conditional distribution of the observation Y given the σ-algebra $\sigma(\mathcal{A}_1, \ldots, \mathcal{A}_k)$ generated by the information sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$. However, information aggregation often is not feasible in practice, when individual sources of expertise reveal predictive distributions, rather than information sets. What we can realistically aim at is to model the conditional distribution of the observation Y given the σ-algebra generated by the predictive cumulative distribution functions, namely

$$G = \mathcal{L}\,(\,Y \,|\, F_1, \ldots, F_k\,),$$

where we define

$$\mathcal{L}\,(\,Y \,|\, F_1, \ldots, F_k\,) = \mathcal{L}\,(\,Y \,|\, F_i(x) : i = 1, \ldots, k,\, x \in \mathbb{Q}\,),$$

with $\mathbb{Q}$ being the set of the rational numbers.

In practice, one resorts to parametric families of combination formulas, which are then fitted on the basis of past experience and training data. Specifically, let $\mathcal{F}$ be a class of fixed, non-random cumulative distribution functions such that $F_1, \ldots, F_k \in \mathcal{F}$ almost surely. For example, if we are concerned with density forecasts on the real line $\mathbb{R}$, we consider the class $\mathcal{D}_{\mathbb{R}}$ of the cumulative distribution functions that admit a Lebesgue density. Further classes $\mathcal{F}$ of interest are listed in Table 3. A *combination formula* then is a mapping of the form

$$G : \mathcal{F}^k = \underbrace{\mathcal{F} \times \cdots \times \mathcal{F}}_{k \text{ times}} \to \mathcal{F}, \quad (F_1, \ldots, F_k) \mapsto G(F_1, \ldots, F_k). \tag{5}$$

Following French and Ríos Insua (2000, p. 113) we say that the combination formula G is *anonymous* if

$$G(F_{\pi(1)}, \ldots, F_{\pi(k)}) = G(F_1, \ldots, F_k)$$

for all $F_1, \ldots, F_k \in \mathcal{F}$ and all permutations π on k elements. For example, the only linear anonymous combination formula is the equally weighted sum. We allow the case $k = 1$, where the mapping G may provide calibration and dispersion adjustments for a single predictive distribution, as discussed in Section 5.

An *aggregation method* is a family

$$\mathcal{G} = \{G_\theta : \theta \in \Theta\}$$

of combination formulas G_θ of the form (5) that share a common value of k and a common class $\mathcal{F}$ of fixed, non-random cumulative distribution functions. For example, if $\mathcal{G}$ is the traditional linear pool, we can take $\mathcal{F}$ to be any convex class of cumulative distribution functions, and we may identify the index set Θ with the unit simplex in $\mathbb{R}^k$.

The extant literature studies individual combination formulas by examining whether or not they possess certain analytic characteristics, such as the strong setwise function and external Bayes properties (McConway, 1981; Genest, 1984a,b; Genest and Zidek, 1986; Genest et al., 1986; French and Ríos Insua, 2000). In contrast, we share the recent perspective of Hora (2010) and put the focus on calibration and dispersion. In particular, we study aggregation methods in terms of the behavior of the probability integral transform under the corresponding family $\mathcal{G} = \{G_\theta : \theta \in \Theta\}$ of combination formulas. The probability integral takes values in the unit interval, and so the possible values of its variance lie between 0 and $\frac{1}{4}$. The value $\frac{1}{12}$ corresponds to neutral dispersion and is the most desirable.

We are now ready to define notions of flexibility for aggregation methods.

Definition 2.12. Consider a family $\mathcal{G} = \{G_\theta : \theta \in \Theta\}$ of combination formulas of the form (5) that share a common $k \geq 1$ and a common class $\mathcal{F}$ of fixed, non-random cumulative distribution functions.

(a) The aggregation method $\mathcal{G}$ is *flexibly dispersive* relative to the class $\mathcal{F}$ if for all $F_0 \in \mathcal{F}$ and $F_1, \ldots, F_k \in \mathcal{F}$ there exists a parameter value $\theta \in \Theta$ such that if $\mathcal{L}(Y) = F_0$ then $G_\theta(F_1, \ldots, F_k)$ is a neutrally dispersed forecast for Y.

(b) The aggregation method $\mathcal{G}$ is *exchangeably flexibly dispersive* relative to the class $\mathcal{F}$ if for all $F_0 \in \mathcal{F}$ and $F_1, \ldots, F_k \in \mathcal{F}$ there exists a parameter value $\theta \in \Theta$ such that G_θ is anonymous and if $\mathcal{L}(Y) = F_0$ then $G_\theta(F_1, \ldots, F_k)$ is a neutrally dispersed forecast for Y.

The applied relevance of the definitions is appreciated as follows. Suppose that the aggregation method $\mathcal{G}$ is flexibly dispersive relative to $\mathcal{F}$. Then, given any marginal law $F_0 \in \mathcal{F}$ for the observation Y and any collection $F_1, \ldots, F_k \in \mathcal{F}$ of probabilistic forecasts for Y, we can find a combination formula $G_\theta \in \mathcal{G}$ such that the aggregated predictive distribution, namely $G_\theta(F_1, \ldots, F_k)$, is neutrally dispersed. If $\mathcal{G}$ is exchangeably flexibly dispersive, we can do so while treating $F_1, \ldots, F_k$ exchangeably, which is a frequent requirement in the practice of the combination of expert judgements (Jouini and Clemen, 1996).

In a nutshell, aggregation methods ought to be sufficiently flexible to accommodate situations typically encountered in practice. Evidently, a positive statement about flexible dispersivity is the stronger, the larger the class $\mathcal{F}$. Conversely, a statement about the lack of flexible dispersivity is the stronger, the smaller the class $\mathcal{F}$.

3. Linear and nonlinear aggregation methods

In this section we study specific combination formulas and aggregation methods from the perspectives of calibration and dispersion. First, we consider the traditional linear pool, then we move on to discuss non-linear ramifications, namely generalized linear pools, the spread-adjusted linear pool, and the beta-transformed linear pool, along with optimum score techniques for the estimation of combination formulas.

3.1. Linear and generalized linear pools

We proceed to state and prove a simple but powerful result about linear combination formulas that generalizes earlier findings by Hora (2004) and Ranjan and Gneiting (2010). The gist of the statement is that dispersion tends to increase under linear aggregation.

Theorem 3.1. *In the prediction space setting, suppose that $k \geq 2$ and consider any linearly combined probabilistic forecast $F = \sum_{i=1}^{k} w_i F_i$ with weights $w_1, \ldots, w_k$ that are strictly positive and sum to 1. For $i \neq j$, suppose that $F_i \neq F_j$ with positive probability. Then the following holds:*

(a) *The linearly combined forecast F is at least as dispersed as the least dispersed of the components $F_1, \ldots, F_k$.*
(b) *If the component forecasts $F_1, \ldots, F_k$ are regular, then F is more dispersed than the least dispersed of the components.*
(c) *If the components are neutrally dispersed and regular, then F is overdispersed. In particular, if the components are probabilistically calibrated, then F is overdispersed.*

Proof. For $i = 1, \ldots, k$, let Z_i denote the probability integral transform of F_i. The probability integral transform of $F = \sum_{i=1}^{k} w_i F_i$ is $Z = \sum_{i=1}^{k} w_i Z_i$, whence

$$\begin{aligned} \operatorname{var}(Z) &= \sum_{i=1}^{k} \sum_{j=1}^{k} w_i w_j \operatorname{cov}(Z_i, Z_j) \\ &\leq \sum_{i=1}^{k} w_i \sum_{j=1}^{k} w_j \left(\max_{1 \leq i \leq k} \operatorname{var}(Z_i) \right) = \max_{1 \leq i \leq k} \operatorname{var}(Z_i), \end{aligned}$$

which demonstrates part (a). To prove part (b) suppose, for a contradiction, that $F_1, \ldots, F_k$ are regular and $\operatorname{var}(Z) = \max_{1 \leq i \leq k} \operatorname{var}(Z_i)$. Then Z_i and Z_j

are perfectly correlated for $i, j = 1, \ldots, k$, and we conclude that there exist constants $a_{ij} > 0$ and $b_{ij} \in \mathbb{R}$ such that $Z_i = a_{ij} Z_j + b_{ij}$ almost surely. By the assumption of regularity, Z_i and Z_j are supported on the unit interval, whence $a_{ij} = 1$ and $b_{ij} = 0$. Therefore, $Z_i = Z_j$ almost surely, contrary to the assumption that $F_i \neq F_j$ with positive probability. Part (c) concerns the special case of part (b) in which $\mathrm{var}(Z_i) = \frac{1}{12}$ for $i = 1, \ldots, k$. □

As noted, Theorem 3.1 yields various extant results as corollaries. For instance, Hora (2004) applied Fourier analytic tools to show that if two distinct density forecasts are probabilistically calibrated, then any nontrivial linear combination is uncalibrated, which is an immediate consequence of part (c). However, the first statement in part (c) is considerably stronger, in that it substitutes the weaker condition of neutral dispersion and regularity for the assumption of probabilistic calibration, allows for any number $k \geq 2$ of components, allows for cumulative distribution functions rather than the special case of densities, and exposes the direction of the deviation, in that the linearly combined forecast is overdispersed. Each of the four facets is useful in practice. For instance, there are real data situations where density forecasts are approximately neutrally dispersed and regular, but clearly not calibrated. Discrete and mixed discrete-continuous predictive cumulative distribution functions also occur frequently in practice, such as in quantitative precipitation forecasting (Sloughter et al., 2007) and for count data (Czado et al., 2009). Finally, the tendency to increase dispersion helps explain the success of linear pooling in applications, where the component distributions are frequently underdispersed. For a prominent example, see Table 10 of Hoeting et al. (1999).

Thus far, we have considered individual linear combination formulas. The following result views the traditional linear pool as an aggregation method $\mathcal{G} = \{G_\theta : \theta \in \Theta\}$, where we may identify the parameter space $\Theta = \Delta_{k-1}$ with the unit simplex in $\mathbb{R}^k$. We state the theorem relative to the full class $\mathcal{F}_\mathbb{R}$, even though it remains valid relative to much smaller classes.

Theorem 3.2. *The linear pool fails to be flexibly dispersive relative to the class* $\mathcal{F}_\mathbb{R}$.

Proof. In view of part (a) of Theorem 3.1 it suffices to find an F_0 and distinct $F_1, \ldots, F_k$, each of which is an overdispersed forecast for an observation Y with $\mathcal{L}(Y) = F_0$. For example, we can take F_0 to be standard normal and F_i to be normal with mean zero and variance $i + 1$ for $i = 1, \ldots, k$. □

Dawid et al. (1995) introduced and studied *generalized linear* combination formulas for combining probability forecasts of a binary event. Here we apply the approach to cumulative distribution functions to obtain combination formulas of the form

$$G(y) = h^{-1}\left(\sum_{i=1}^{k} w_i\, h(F_i(y))\right) \quad \text{or} \quad h(G(y)) = \sum_{i=1}^{k} w_i\, h(F_i(y)), \tag{6}$$

TABLE 4
Specifics of the generalized linear combination formula in equation (6). The table states assumptions on the weights, $w_1, \ldots, w_k$, and instances of classes $\mathcal{F}$, such that the combination formula maps $\mathcal{F}^k$ into $\mathcal{F}$. The conditions depend on the domain and the range of the link function, h, which we assume to be continuous and strictly monotone

Type	Domain	Range	Weights	Class $\mathcal{F}$	Example
A	$[0,1]$	any	$w_i \geq 0$; $\sum_{i=1}^k w_i = 1$	$\mathcal{F}_{\mathbb{R}}$	$h(x) = x$
B	$(0,1)$	$(1,\infty)$	$w_i \geq 0$; $\sum_{i=1}^k w_i = 1$	$\mathcal{F}_{\mathrm{I}}^+$ or $\mathcal{B}^+$	$h(x) = 1/x$
C	$(0,1)$	$(-\infty,0)$	$w_i \geq 0$; $\sum_{i=1}^k w_i > 0$	$\mathcal{F}_{\mathrm{I}}^+$ or $\mathcal{B}^+$	$h(x) = \log x$
D	$(0,1)$	$\mathbb{R}$	$w_i \geq 0$; $\sum_{i=1}^k w_i > 0$	$\mathcal{F}_{\mathrm{I}}^+$ or $\mathcal{B}^+$	$h(x) = \Phi^{-1}(x)$

where h is a continuous and strictly monotone link function. Table 4 shows conditions on the weights, $w_1, \ldots, w_k$, along with instances of classes $\mathcal{F}$, so that the generalized linear combination formula (6) maps $\mathcal{F}^k$ into $\mathcal{F}$. Link functions of the first type are defined on the closed unit interval, and the combination formula operates on the full class $\mathcal{F}_{\mathbb{R}}$, with the traditional linear pool, for which $h(x) = x$ is the identity function, being the most prominent example. Link functions of the other types are defined on the open unit interval only, and we need to restrict attention to $\mathcal{F}_{\mathrm{I}}^+$ or $\mathcal{B}^+$, with the harmonic pool and the geometric pool being examples, occuring when $h(x) = 1/x$ and $h(x) = \log x$, respectively. While not being exhaustive, the listing in the table is comprehensive, in that most link functions can be adapted to fit one of the types considered. The defining equation (6) implies that

$$\operatorname{var}(h(G(Y))) \leq \left(\sum_{i=1}^k w_i \right)^2 \max_{1 \leq i \leq k} \operatorname{var}(h(F_i(Y))). \tag{7}$$

In the case of the identity link the inequality yields the results in Theorems 3.1 and 3.2. In the general case it suggests that generalized linear pools with link functions of types A and B, for which the first factor on the right-hand side of (7) reduces to the constant 1, may fail to be flexibly dispersive.

3.2. Spread-adjusted linear pool

The aforementioned limitations of linear and generalized linear pools suggest that we consider more flexible, nonlinear aggregation methods. In this section, we focus on the class $\mathcal{D}_{\mathbb{R}}^+$, so that we may identify the cumulative distribution functions $F_1, \ldots, F_k$ with the corresponding Lebesgue densities $f_1, \ldots, f_k$.

In the context of probabilistic weather forecasts and approximately neutrally dispersed Gaussian components $f_1, \ldots, f_k$, Berrocal et al. (2007), Glahn et al. (2009) and Kleiber et al. (2011) observed empirically that linearly combined predictive distributions are overdispersed, as confirmed by Theorem 3.1. In an ad hoc approach, they proposed a nonlinear aggregation method which we now generalize and refer to as the *spread-adjusted linear pool* (SLP).

To describe this technique, it is convenient to write $F_i(y) = F_i^0(y - \mu_i)$ and $f_i(y) = f_i^0(y - \mu_i)$, where μ_i is the unique median of $F_i \in \mathcal{D}_{\mathbb{R}}^+$, for $i = 1, \ldots, k$. The SLP combined predictive distribution then has cumulative distribution function and Lebesgue density

$$G_c(y) = \sum_{i=1}^{k} w_i F_i^0\left(\frac{y - \mu_i}{c}\right) \quad \text{and} \quad g_c(y) = \frac{1}{c}\sum_{i=1}^{k} w_i f_i^0\left(\frac{y - \mu_i}{c}\right), \tag{8}$$

respectively, where $w_1, \ldots, w_k$ are nonnegative weights that sum to 1, and c is a strictly positive spread adjustment parameter. For neutrally dispersed or overdispersed components values of $c < 1$ are appropriate; for example, Table 2 of Berrocal et al. (2007) reports estimates ranging from 0.65 to 1.03. Underdispersed components may suggest values of $c \geq 1$, and the traditional linear pool arises when $c = 1$.

The SLP method performs well in the aforementioned applications, and the following result serves to quantify its flexibility.

Proposition 3.3. *Suppose that $\mathcal{L}(Y) = F_0 \in \mathcal{D}_{\mathbb{R}}^+$ and that $F_1, \ldots, F_k \in \mathcal{D}_{\mathbb{R}}^+$ have medians $\mu_1 \leq \cdots \leq \mu_k$. Let $Z_c = G_c(Y)$ denote the probability integral transform of the SLP aggregated predictive cumulative distribution function. Let $v_0 = 0$ and $p_0 = F_0(\mu_1)$, let $v_i = \sum_{j=1}^{i} w_j$ and $p_i = F_0(\mu_{i+1}) - F_0(\mu_i)$ for $i = 1, \ldots, k-1$, and let $v_k = 1$ and $p_k = 1 - F_0(\mu_k)$. Then as the spread adjustment parameter $c > 0$ varies, the variance of Z_c attains any positive value less than*

$$\sum_{i=0}^{k} p_i \left(v_i - \sum_{j=0}^{k} p_j v_j\right)^2. \tag{9}$$

Proof. As $c \to 0$, the function G_c converges to the cumulative distribution function of the discrete probability measure with mass $w_1, \ldots, w_k$ at $\mu_1, \ldots, \mu_k$, respectively. Since the distribution of Y is F_0, the law of $Z_c = G_c(Y)$ converges weakly to the discrete probability measure with mass $p_0 = F_0(\mu_1)$ at $v_0 = 0$, mass $p_i = F_0(\mu_{i+1}) - F_0(\mu_i)$ at $v_i = \sum_{j=1}^{i} w_j$ for $i = 1, \ldots, k-1$, and mass $p_k = 1 - F_0(\mu_k)$ at $v_k = 1$. Hence as $c \to 0$ the variance of Z_c converges to (9). As $c \to \infty$, the law of $Z_c = G_c(Y)$ converges weakly to the Dirac measure in $\frac{1}{2}$. In view of the variance of Z_c being a continuous function of the spread adjustment parameter $c > 0$, this proves the claim. □

Our next result views the spread-adjusted linear pool as an aggregation method with parameter space $\Theta = \Delta_{k-1} \times \mathbb{R}_+$. While the SLP approach is sufficiently rich in typical applications, where the individual predictive distributions are neutrally dispersed or underdispersed, its flexibility is limited.

Theorem 3.4. *The spread-adjusted linear pool fails to be flexibly dispersive relative to the class $\mathcal{D}_{\mathbb{R}}^+$.*

Proof. Let F_0 be standard normal, and for $i = 1, \ldots, k$ let F_i be normal with mean $m + \frac{i}{m}$ and variance 1. As $m \to \infty$, the probability integral transform

of the SLP combined forecast G_c attains values less than $\frac{1}{2}$ with probability tending to one, irrespectively of the values of the SLP weights $w_1, \ldots, w_k$ and the spread adjustment parameter c. Thus, if m is sufficiently large, the variance of the PIT remains below the critical value of $\frac{1}{12}$ that corresponds to neutral dispersion. □

The SLP combination formula (8) can be generalized to allow for distinct spread adjustment parameters for the individual components. However, such an extension does not allow for flexible dispersivity either, and tends not to be beneficial in applications, unless the component densities have drastically varying degrees of dispersion. The assumption of a common spread adjustment parameter yields a more parsimonious model and stabilizes the estimation.

3.3. Beta-transformed linear pool

The *beta-transformed linear pool* (BLP) composites the traditional linear pool with a beta transform. Introduced by Ranjan and Gneiting (2010) in the context of probability forecasts for a binary event, it generalizes readily to the full class $\mathcal{F}_{\mathbb{R}}$ of the cumulative distribution functions on $\mathbb{R}$. Specifically, the BLP combination formula maps $F_1, \ldots, F_k \in \mathcal{F}_{\mathbb{R}}$ to $G_{\alpha,\beta} \in \mathcal{F}_{\mathbb{R}}$, where

$$G_{\alpha,\beta}(y) = B_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i F_i(y)\right) \tag{10}$$

for $y \in \mathbb{R}$. Here, $w_1, \ldots, w_k$ are nonnegative weights that sum to 1, and $B_{\alpha,\beta}$ denotes the cumulative distribution function of the beta density with parameters $\alpha > 0$ and $\beta > 0$. In contrast to the spread-adjusted linear pool, the value of the BLP aggregated predictive cumulative distribution function $G_{\alpha,\beta}$ at $y \in \mathbb{R}$ depends on $F_1, \ldots, F_k$ only through the values $F_1(y), \ldots, F_k(y)$, in a locality characteristic that resembles the strong setwise function property of McConway (1981). If F_i has Lebesgue density f_i for $i = 1, \ldots, k$, the aggregated cumulative distribution function $G_{\alpha,\beta}$ is absolutely continuous with Lebesgue density

$$g_{\alpha,\beta}(y) = \left(\sum_{i=1}^{k} w_i f_i(y)\right) b_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i F_i(y)\right),$$

where $b_{\alpha,\beta}$ denotes the beta density with parameters $\alpha > 0$ and $\beta > 0$. This nests the traditional linear pool that arises when $\alpha = \beta = 1$.

The following result concerns the flexibility of the BLP combination formula (10) when the cumulative distribution functions $F_0 \in \mathcal{C}_{\mathbb{R}}$ and $F_1, \ldots, F_k \in \mathcal{C}_{\mathbb{R}}$ are continuous and the weights $w_1, \ldots, w_k \geq 0$ are fixed, while the transformation parameters vary.

Proposition 3.5. *Let Y have distribution $F_0 \in \mathcal{C}_{\mathbb{R}}$ and suppose that $F_1, \ldots, F_k \in \mathcal{C}_{\mathbb{R}}$ are such that*

$$\operatorname{supp}(F_1) \cup \cdots \cup \operatorname{supp}(F_k) = \operatorname{supp}(F_0). \tag{11}$$

Let $Z_{\alpha,\beta} = G_{\alpha,\beta}(Y)$ denote the probability integral transform of the BLP aggregated predictice cumulative distribution function, where the weights are fixed at strictly positive values that sum to 1. Then as the transformation parameters $\alpha > 0$ and $\beta > 0$ vary, the variance of $Z_{\alpha,\beta}$ attains any value in the open interval $(0, \frac{1}{4})$.

Proof. The variance of $Z_{\alpha,\beta}$ depends continuously on the transformation parameters $\alpha > 0$ and $\beta > 0$, with $Z_{\alpha,\alpha}$ converging weakly to the Dirac measure in $\frac{1}{2}$ as $\alpha \to \infty$, so that $\mathrm{var}(Z_{\alpha,\alpha}) \to 0$ as $\alpha \to \infty$. If we can demonstrate the existence of a sequence $(\alpha, \beta(\alpha)) \to (0, 0)$ such that $G_{\alpha,\beta(\alpha)}(y_0) = \frac{1}{2}$, where y_0 is any median of F_0, the proof is complete, as the corresponding probability integral transform $Z_{\alpha,\beta(\alpha)}$ converges weakly to the Bernoulli measure with success probability $\frac{1}{2}$, so that $\mathrm{var}(Z_{\alpha,\beta(\alpha)}) \to \frac{1}{4}$ as $\alpha \to 0$.

We thus strive to find a sequence $(\alpha, \beta(\alpha)) \to (0, 0)$ such that

$$G_{\alpha,\beta(\alpha)}(y_0) = B_{\alpha,\beta(\alpha)}(u_0) = \frac{1}{2},$$

where $u_0 = \sum_{i=1}^{k} w_i F_i(y_0) \in (0, 1)$ by the support condition (11). First we show that for every $\alpha > 0$ there exists a unique $\beta(\alpha) > 0$ such that $B_{\alpha,\beta(\alpha)}(u_0) = \frac{1}{2}$; then we prove that $\beta(\alpha) \to 0$ as $\alpha \to 0$. As regards the first claim, three cases are to be distinguished. If $u_0 < \frac{1}{2}$ then $B_{\alpha,\alpha}(u_0) < \frac{1}{2}$ and $B_{\alpha,\beta}(u_0) \to 1$ as $\beta \to \infty$, and continuity and monotonicity with respect to β imply the existence of a unique $\beta(\alpha) > \alpha$ such that $B_{\alpha,\beta(\alpha)}(u_0) = \frac{1}{2}$. If $u_0 = \frac{1}{2}$ the choice $\beta(\alpha) = \alpha$ is unique. If $u_0 > \frac{1}{2}$ then $B_{\alpha,\alpha}(u_0) > \frac{1}{2}$ and $B_{\alpha,\beta}(u_0) \to 0$ monotonically as $\beta \to 0$, and thus there exists a unique $\beta(\alpha) < \alpha$ such that $B_{\alpha,\beta(\alpha)}(u_0) = \frac{1}{2}$. To prove the second claim, suppose that $\beta(\alpha) > \beta_0 > 0$ for a sequence $\alpha \to 0$. Then as $\alpha \to 0$ the beta distribution with parameters α and $\beta(\alpha)$ has mean $\alpha/(\alpha + \beta(\alpha)) \to 0$, whereas its median u_0 remains fixed and strictly positive, for the desired contradiction. □

The next result views the beta-transformed linear pool as an aggregation method with parameter space $\Theta = \Delta_{k-1} \times \mathbb{R}^2_+$.

Theorem 3.6. *The beta-transformed linear pool is exchangeably flexibly dispersive relative to the class $\mathcal{C}^+_{\mathrm{I}}$, for every interval $I \subseteq \mathbb{R}$.*

Proof. If $F_0 \in \mathcal{C}^+_{\mathrm{I}}$ and $F_1, \ldots, F_k \in \mathcal{C}^+_{\mathrm{I}}$, the support condition (11) is satisfied. We may thus apply Proposition 3.5 with weights $w_1 = \cdots = w_k = \frac{1}{k}$. □

In practices, the BLP weights $w_1, \ldots, w_k$ and transformation parameters $\alpha, \beta > 0$ need to be estimated from training data $\{(F_{1j}, \ldots, F_{kj}, y_j) : j = 1, \ldots, J\}$ of the form (1). If the predictive cumulative distribution functions $F_{1j}, \ldots, F_{kj}$ are absolutely continuous with Lebesgue densities $f_{1j}, \ldots, f_{kj}$ for $j = 1, \ldots, J$, the aggregated predictive distributions are also absolutely continuous, and our preferred estimation technique is to maximize the mean (or sum) of the logarithmic score (Gneiting and Raftery, 2007) over the training data,

namely

$$\begin{aligned}\ell(w_1,\ldots,w_k;\alpha,\beta) &= \sum_{j=1}^{J}\log(g_{\alpha,\beta}(y_j)) \qquad (12)\\ &= \sum_{j=1}^{J}\log\left(\sum_{i=1}^{k}w_i f_{ij}(y_j)\right) + \sum_{j=1}^{J}\log\left(b_{\alpha,\beta}\left(\sum_{i=1}^{k}w_i F_{ij}(y_j)\right)\right)\\ &= \sum_{j=1}^{J}\left((\alpha-1)\log\left(\sum_{i=1}^{k}w_i F_{ij}(y_j)\right)+(\beta-1)\log\left(1-\sum_{i=1}^{k}w_i F_{ij}(y_j)\right)\right)\\ &\quad + \sum_{j=1}^{J}\log\left(\sum_{i=1}^{k}w_i f_{ij}(y_j)\right) - J\log \mathrm{B}(\alpha,\beta),\end{aligned}$$

where B denotes the classical beta function. The logarithmic score is simply the logarithm of the value that the density forecast attains at the realizing observation. It is positively oriented, that is, the higher the score, the better, and it is proper, in the sense that truth telling is an expectation maximizing strategy.

The optimization can be carried out numerically using the method of scoring, for which we give details in Appendix B. Approximate standard errors for the estimates can be obtained in the usual way, by evaluating and inverting the Hessian matrix for the mean logarithmic score or log likelihood function. However, the estimates of the weights $w_1,\ldots,w_k$ need to be nonnegative. Thus, if unconstrained optimization results in negative weights, we turn to the active barrier algorithm implemented in the constrained optimization routine constrOptim in R (R Development Core Team, 2011). Similarly, linear, generalized linear and spread-adjusted linear combination formulas can be fitted by maximizing the mean logarithmic score over training data. The corresponding optimum score estimates can be viewed as maximum likelihood estimates under the assumption of independence between the training cases, and for reasons of simplicity and tradition, we refer to them as maximum likelihood estimates.

4. Simulation and data examples

We now illustrate and complement our theoretical results in simulation and data examples on density forecasts. This corresponds to the prediction space setting, where the CDF-valued random quantities $F_1,\ldots,F_k$ are absolutely continuous almost surely, and thus can be identified with random Lebesgue densities $f_1,\ldots,f_k$. Throughout the section, we fit combination formulas by maximizing the mean logarithmic score over training data, in the ways described in Section 3.3 and Appendix B. To lighten the notation, we use the acronyms PIT, TLP, SLP and BLP to refer to the probability integral transform and the traditional, spread-adjusted and beta-transformed linear pool, respectively.

The recent work of Ranjan and Gneiting (2010), Clements and Harvey (2011) and Allard et al. (2012) contains a wealth of simulation and data examples on

the combination of probability forecasts for a binary event. In the concluding Section 5 we summarize these experiences and relate them to the findings in the case studies hereinafter.

4.1. Simulation example

In this simulation example, the data generating process for the observation, Y, is the regression model

$$Y = X_0 + a_1X_1 + a_2X_2 + a_3X_3 + \epsilon, \tag{13}$$

where a_1, a_2 and a_3 are real constants, and X_0, X_1, X_2, X_3 and ϵ are independent, standard normal random variables. The individual predictive distributions rest on partial knowledge of the data generating process, in that density forecast f_1 has access to the covariates X_0 and X_1, but not to X_2 or X_3, and similarly for f_2 and f_3. Thus, we seek to combine the density forecasts

$$f_1 = \mathcal{N}(X_0 + a_1X_1, 1 + a_2^2 + a_3^2),$$

$$f_2 = \mathcal{N}(X_0 + a_2X_2, 1 + a_1^2 + a_3^2) \quad \text{and} \quad f_3 = \mathcal{N}(X_0 + a_3X_3, 1 + a_1^2 + a_2^2),$$

where X_0 stands for shared, public information, while X_1, X_2 and X_3 represent proprietary information sets. The density forecasts represent the true conditional distributions under the regression model (13), given the corresponding partial information, as represented by the σ-algebras $\mathcal{A}_1 = \sigma(X_0, X_1)$, $\mathcal{A}_2 = \sigma(X_0, X_2)$ and $\mathcal{A}_3 = \sigma(X_0, X_3)$, respectively. Hence, the forecasts are ideal in the sense of Definition 2.2, and by Theorem 2.8 they are both probabilistically calibrated and marginally calibrated.

We estimate the TLP, SLP and BLP combination formulas on a simple random sample $\{(f_{1j}, f_{2j}, f_{3j}, Y_j) : j = 1, \ldots, J\}$ of size $J = 500$ from the joint distribution of the forecasts and the observation, and evaluate on an independent test sample of the same size. The regression coefficients in the data generating model (13) are taken to be $a_1 = a_2 = 1$ and $a_3 = 1.1$, so that f_3 is a more concentrated, sharper density forecast than f_1 and f_2.

Table 5 shows maximum likelihood estimates, along with approximate standard errors, for TLP, SLP and BLP combination formulas. For all three methods, the weight estimate is highest for f_3, whereas the estimates for f_1 and f_2 are smaller and not significantly different from each other. The SLP spread adjustment parameter c is estimated at 0.78, and the BLP transformation parameters α and β at 1.49 and 1.44, respectively.

The PIT histograms for the various types of density forecasts over the test set are displayed in Figure 3, with complementary results shown in Table 6. In addition to the variance of the PIT, which is our standard measure of dispersion, the table quantifies sharpness in terms of the root mean variance (RMV), that is, the square root of the average of the variance of the predictive density over the evaluation set. The component forecasts f_1, f_2 and f_3 are probabilistically calibrated and thus show uniform empirical PIT histograms, up to sample

TABLE 5
Maximum likelihood estimates with approximate standard errors (in brackets) for the parameters of the combined density forecasts in the simulation example

	w_1	w_2	w_3	c	α	β
TLP	0.212 (0.083)	0.254 (0.084)	0.534 (0.080)	—	—	—
SLP	0.257 (0.060)	0.283 (0.061)	0.460 (0.059)	0.783 (0.030)	—	—
BLP	0.256 (0.057)	0.293 (0.057)	0.451 (0.054)	—	1.492 (0.062)	1.440 (0.059)

TABLE 6
Variance of the PIT (dispersion) and root mean variance of the density forecast (sharpness) in the simulation example, for the test set. A value of $\frac{1}{12}$ *or about* 0.083 *for the variance of the PIT indicates neutral dispersion*

	var(PIT)	RMV
f_1	0.081	1.79
f_2	0.086	1.79
f_3	0.085	1.73
TLP	0.066	1.94
SLP	0.081	1.62
BLP	0.084	1.57

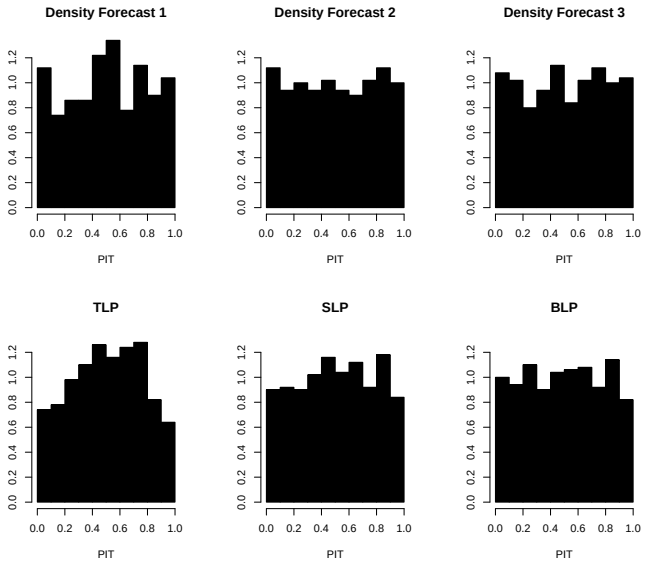

FIG 3. *PIT histograms for the individual and combined density forecasts in the simulation example, for the test set.*

TABLE 7
Mean logarithmic score for the individual and combined density forecasts in the simulation example, for the training set and the test set

	Training	Test
f_1	−2.025	−2.018
f_2	−2.017	−2.022
f_3	−1.956	−1.992
TLP	−1.907	−1.922
SLP	−1.871	−1.892
BLP	−1.865	−1.886

fluctuations. As mandated by Theorem 3.1, the linearly combined TLP density forecast is overdispersed and lacks sharpness. The SLP and BLP aggregated density forecasts show nearly uniform PIT histograms; they are approximately neutrally dispersed and much sharper than their competitors.

Table 7 shows the mean logarithmic score for the various types of density forecasts. The best individual density forecast is f_3, because it is sharper than f_1 and f_2. The linearly combined density forecast outperforms the individual density forecasts, even though it is overdispersed. The nonlinearly aggregated SLP and BLP density forecasts show higher scores than any of the individual or linearly combined forecasts, both for the training data, where this is trivially true, as the nonlinear methods nest the traditional linear pool, and for the test data, where such cannot be guaranteed.

4.2. Density forecasts for daily maximum temperature at Seattle-Tacoma Airport

With estimates of some one-third of the economy, as well as much of human activity in general, being weather sensitive (Dutton, 2002), there is a critical need for calibrated and sharp probabilistic weather forecasts, to allow for optimal decision making under inherent environmental uncertainty.

In practice, probabilistic weather forecasts rely on ensemble prediction systems. An ensemble system comprises multiple runs of a numerical weather prediction model, with the runs differing in the initial conditions and/or the details of the mathematical representation of the atmosphere (Gneiting and Raftery, 2005). Here we consider two-days ahead forecasts of daily maximum temperature at Seattle-Tacoma Airport, based on the University of Washington Mesoscale Ensemble (Eckel and Mass, 2005), which employs a regional numerical weather prediction model over the Pacific Northwest, with initial and lateral boundary conditions supplied by eight distinct weather centers. A description of the ensemble members is given in Table 8.

Our training period ranges from January 1, 2006 to August 12, 2007, with a few days missing in the data record, for a total of 500 training cases. The test period extends from August 13, 2007 to June 30, 2009, for a total of 559 cases. Each ensemble member is a point forecast, which can be viewed as the most extreme form of an underdispersed density forecast. To address the un-

derdispersion and obtain approximately neutrally dispersed components, we use the maximum likelihood method on the training data to fit, for each ensemble member $i = 1, \ldots, 8$ individually, a Gaussian predictive density of the form

$$f_i = \mathcal{N}(a_i + b_i x_{ij}, \sigma_i^2).$$

Here x_{ij} is the point forecast from the ith ensemble member on day j, a_i and b_i are member specific linear bias correction parameters, and σ_i is the member specific predictive standard deviation. From Table 9 we see that the estimates for $\sigma_1, \ldots, \sigma_8$ range from 1.958 to 2.214.

Next we combine the eight individual density forecasts. Table 10 shows maximum likelihood estimates for TLP, SLP and BLP combination formulas. For all three methods, the GFS member, f_1, obtains the highest weight and the ETA member, f_3, the lowest weight. This can readily be explained, in that both members have a common institutional origin, and thus are highly correlated, whence the more competitive GFS member subsumes the weight of the ETA member. The SLP spread adjustment parameter is estimated at 0.768, and the BLP transformation parameters both at 1.467.

Figure 4 illustrates the various density forecasts for June 28, 2008, an unusually hot day at Seattle-Tacoma Airport with a verifying maximum temperature of 32.8 degrees Celsius or 91 degrees Fahrenheit. The member specific individual density forecasts are shown by the dotted lines, and the linearly combined TLP forecast by the dash-dotted line. The nonlinearly aggregated SLP and BLP density forecasts, which are shown by the solid and dashed line, respectively, are sharper than the TLP density.

PIT histograms for the test period are shown in Figure 5, along with summary measures of dispersion and sharpness in Table 11. The individual, member specific density forecasts tend to be a bit overdispersed. The linearly aggregated TLP density forecast is much more severely overdispersed, as reflected by an inverse U-shaped and skewed PIT histogram. Of course, the overdispersion is not surprising, as it is a direct consequence of Theorem 3.1. The SLP and BLP aggregated density forecasts show somewhat rough and skewed, yet more nearly uniform PIT histograms.

These results are corroborated by Table 12, which shows the mean logarithmic score for the various types of density forecasts, both for the training period and the test period. The linearly combined TLP forecast shows a higher score than any of the individual density forecasts, which attests to the benefits of aggregation. Nevertheless, the linearly combined density forecast is suboptimal, because it is overdispersed and lacks sharpness, and thus it is outperformed by the nonlinearly aggregated SLP and BLP density forecasts.

Finally, we compare to the Bayesian model averaging (BMA) technique (Raftery et al., 2005), which is a state of the art approach to generating density forecasts from forecast ensembles. The BMA density forecast for day j is of the form

$$g = \sum_{i=1}^{8} w_i \, \mathcal{N}(a_i + b_i x_{ij}, \, \sigma^2), \tag{14}$$

TABLE 8

Composition of the eight-member University of Washington Mesoscale Ensemble (Eckel and Mass, 2005), with member acronyms and organizational sources for initial and lateral boundary conditions. The United States National Centers for Environmental Prediction supply two distinct sets of initial and lateral boundary conditions, namely, from its Global Forecast System (GFS) and Limited-Area Mesoscale Model (ETA)

Index	Acronym	Source of Initial and Lateral Boundary Conditions
1	GFS	National Centers for Environmental Prediction
2	CMCG	Canadian Meteorological Centre
3	ETA	National Centers for Environmental Prediction
4	GASP	Australian Bureau of Meteorology
5	JMA	Japanese Meteorological Agency
6	NGPS	Fleet Numerical Meteorology and Oceanography Center
7	TCWB	Taiwan Central Weather Bureau
8	UKMO	United Kingdom Met Office

TABLE 9

Maximum likelihood estimates for the predictive standard deviation, σ_i, for the individual, member specific density forecasts in the temperature example

σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8
1.966	2.051	2.119	2.214	1.958	2.055	2.084	1.995

TABLE 10

Maximum likelihood estimates for the parameters of the combined density forecasts in the temperature example, including the Bayesian model averaging (BMA) approach of Raftery et al. (2005)

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	c	α	β	σ
TLP	0.394	0.005	0.000	0.000	0.317	0.030	0.144	0.109	—	—	—	—
SLP	0.304	0.080	0.000	0.085	0.216	0.051	0.172	0.090	0.768	—	—	—
BLP	0.295	0.079	0.000	0.083	0.230	0.062	0.173	0.076	—	1.467	1.467	—
BMA	0.305	0.075	0.000	0.081	0.216	0.056	0.170	0.098	—	—	—	1.566

TABLE 11

Variance of the PIT (dispersion) and root mean variance of the density forecast (sharpness) in the temperature example, for the test period. A value of $\frac{1}{12}$ or about 0.083 for the variance of the PIT indicates neutral dispersion

	var(PIT)	RMV
f_1	0.070	1.97
f_2	0.067	2.05
f_3	0.069	2.12
f_4	0.068	2.21
f_5	0.070	1.96
f_6	0.073	2.06
f_7	0.074	2.08
f_8	0.069	2.00
TLP	0.057	2.15
SLP	0.070	1.79
BLP	0.072	1.77
BMA	0.070	1.80

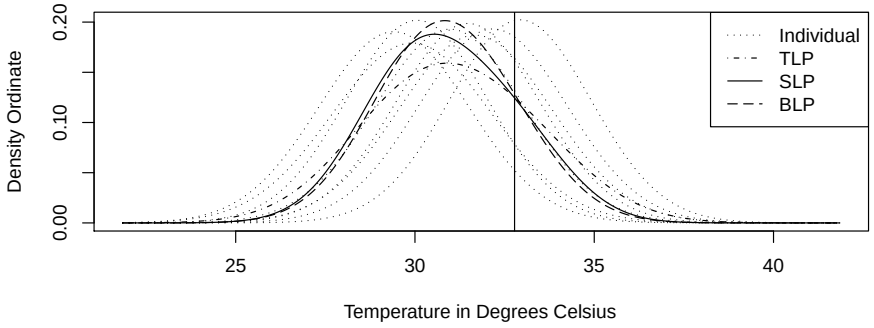

FIG 4. *Two-day ahead density forecasts for the maximum temperature at Seattle-Tacoma Airport on June 28, 2008. The vertical line is at the verifying realization, at 32.8 degrees Celsius or 91 degrees Fahrenheit.*

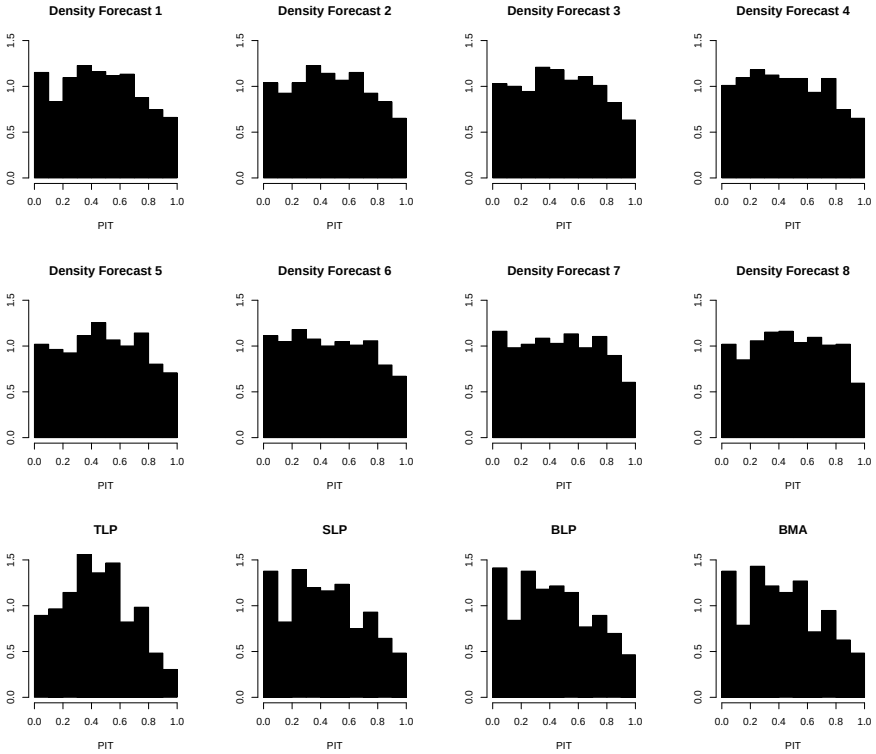

FIG 5. *PIT histograms for the individual and combined density forecasts in the temperature example, for the test period.*

TABLE 12
Mean logarithmic score for the individual and combined density forecasts in the temperature example, for the training period and the test period

	Training	Test
f_1	−2.091	−2.088
f_2	−2.134	−2.071
f_3	−2.167	−2.093
f_4	−2.211	−2.172
f_5	−2.088	−2.043
f_6	−2.136	−2.143
f_7	−2.150	−2.131
f_8	−2.107	−2.041
TLP	−2.027	−2.010
SLP	−1.990	−1.961
BLP	−1.988	−1.960
BMA	−1.992	−1.963

with BMA weights, $w_1, \ldots, w_8$, that are nonnegative and sum to 1, member specific bias parameters a_i and b_i for $i = 1, \ldots, 8$, and a common variance parameter, σ^2. In view of our individual density forecasts being Gaussian, the TLP and BMA densities are of the same functional form. However, there is a conceptual difference, in that the TLP weights are fitted conditionally on the individual density forecasts. Thus, a two-stage procedure is used, in which the member specific component densities are estimated first, and only then the weights, with the components held fixed. In contrast, the BMA method estimates the weights, $w_1, \ldots, w_8$, and the common spread parameter, σ, for the component forecasts in the Gaussian mixture model (14) simultaneously. While the BMA method can be employed with member specific spread parameters, the assumption of a common spread parameter stabilizes the estimation algorithm and does not appreciably deteriorate the predictive performance (Raftery et al., 2005).

Table 10 shows maximum likelihood estimates for the BMA parameters, obtained with the R package ENSEMBLEBMA (Fraley et al., 2011). The BMA weights echo the SLP weights. The BMA spread parameter σ is estimated at 1.566 and differs from the predictive standard deviations for the member specific density forecasts in Table 9 by factors ranging from 0.707 to 0.800, much in line with our estimate of 0.768 for the SLP spread adjustment parameter, c. Thus, the SLP and BMA density forecasts are very much alike, which is confirmed by the PIT histograms in Figure 5, the summary measures in Table 11 and the logarithmic scores in Table 12. In Figure 4 the graphs for the SLP and BMA density forecasts are nearly identical and lie essentially on top of each other, and so we refrain from plotting the BMA density.

4.3. Density forecasts for S&P 500 returns

In this final data example, we follow Diebold et al. (1998) in considering S&P 500 log returns for the period of July 3, 1962 to December 29, 1995. The data record through December 1978 is used as training set, for a total of 4,133 training

TABLE 13
Maximum likelihood estimates of the parameters for the combined density forecasts in the S&P 500 example

	w_1	w_2	c	α	β
TLP	0.821	0.179	—	—	—
SLP	0.756	0.244	0.940	—	—
BLP	0.758	0.242	—	1.100	1.081

TABLE 14
Mean logarithmic score for the individual and combined density forecasts in the S&P 500 example, for the training period and the test period

	Training	Test
f_1	3.606	3.458
f_2	3.492	3.247
TLP	3.612	3.469
SLP	3.614	3.470
BLP	3.614	3.470

cases. All estimates reported are maximum likelihood fits on the training period obtained with the R package FGARCH (Wuertz, 2007). The balance of the record is used as test period, for a total of 4,298 one-day ahead density forecasts.

The first component forecast, f_1, is based on a generalized autoregressive conditional heteroscedasticity (GARCH) specification Bollerslev (1986) for the variance structure. With r_t denoting the log return on day t, our GARCH(1,1) model assumes that $r_t = \sigma_t \epsilon_t$, where ϵ_t is Student-t distributed with ν degrees of freedom and variance 1, while σ_t evolves dynamically as

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2.$$

The maximum likelihood estimates for the GARCH parameters are $\omega = 0.000$, $\alpha = 0.089$, $\beta = 0.903$ and $\nu = 9.25$.

The second component forecast, f_2, is based on a standard moving average (MA) model for the mean dynamics, which assumes that $r_t = Z_t + \theta Z_{t-1}$, where $\{Z_t\}$ is a Gaussian white noise process with mean zero and variance σ^2. The maximum likelihood estimates for the MA parameters are $\theta = 0.252$ and $\sigma = 0.00736$.

Our goal now is to combine the density forecasts f_1 and f_2. Table 13 shows maximum likelihood estimates for TLP, SLP and BLP combination formulas. For all three methods, the conditionally heteroscedastic density forecast f_1 obtains a much higher weight than the simplistic density forecast f_2. The SLP spread adjustment parameter is estimated at 0.940, and the BLP transformation parameters α and β at 1.100 and 1.081.

Table 14 shows the mean logarithmic score for the various types of probabilistic forecasts. The TLP density forecast performs slightly better than the individual component f_1, with a score that is very slightly lower than for the nonlinearly aggregated SLP and BLP density forecasts, both for the training and the test period. As also observed by Geweke and Amisano (2011), there is

little reward for using more elaborate, less parsimonious aggregation methods for density forecasts of S&P 500 returns.[3]

Finally, we consider the predictive performance of a more comprehensive predictive model, which addresses both the first and the second order dynamics, in that $r_t = \mu_t + \epsilon_t$ where $\{\mu_t\}$ and $\{\epsilon_t\}$ are MA(1) and Student-t GARCH(1,1) processes, respectively. The maximum likelihood estimates in this mixed specification are $\theta = 0.269$ and $\sigma = 0.00736$ for the MA parameters, and $\omega = 0.000$, $\alpha = 0.098$, $\beta = 0.892$ and $\nu = 8.284$ for the GARCH parameters. The resulting density forecast can be thought of as combining information sets with respect to the first and second order dynamics, as opposed to combining the corresponding component forecasts f_1 and f_2. It outperforms the other types of density forecasts and achieves a mean logarithmic score of 3.638 for the training period and 3.473 for the test period.

5. Discussion

We have studied methods for combining predictive distributions. From a theoretical perspective, our approach departs from previous work in major ways. Technically, we operate in terms of prediction spaces and cumulative distribution functions, which allows for a unified treatment of all real-valued predictands including, for example, density forecasts for continuous variables and probability forecasts of a binary event. In this latter context, Theorem 2.11 is an analytic result of independent interest, in that the general notion of probabilistic calibration embeds the traditional concept of conditional calibration.

Conceptually, our work is motivated by applications in probabilistic forecasting, and thus we assess combination formulas and aggregation methods from the perspective of calibration and dispersion, the key idea being that aggregation methods ought to be parsimonious, yet sufficiently flexible to allow for neutrally dispersed combined forecasts. In typical practice, underdispersed or approximately neutrally dispersed predictive distributions are to be aggregated. In the case of underdispersed components, the tendency of linear combination formulas to increase dispersion can be beneficial, and helps explain the success of linear pooling in applications (Madigan and Raftery, 1994). However, if the components are neutrally dispersed, the failure of the traditional linear pool to be flexibly dispersive is a serious limitation. Berrocal et al. (2007), Glahn et al. (2009) and Kleiber et al. (2011) observed this empirically in the context of probabilistic weather forecasts, and proposed a special case of the spread-adjusted linear pool as an ad hoc remedy. Our theoretical results document the increased flexibility of the spread-adjusted linear pool, and demonstrate that the beta-transformed linear pool is exchangeably flexibly dispersive.

Not surprisingly, the parsimonity principle and the bias-variance tradeoff apply in the practice of the combination of predictive distributions. Thus, in data

[3]The logarithmic scores reported by Geweke and Amisano (2011) are summed, rather than averaged, and apply to percent log returns, rather than log returns. Adjusted for these differences, they are comparable to the scores in Table 14.

poor settings, where training data are scarce, the parsimonious traditional linear pool might be the method of choice, despite its theoretical shortcomings, as demonstrated persuasively in the recent simulation study of Clements and Harvey (2011). In data rich settings, where predictive models can reliably be estimated, linear aggregation tends to be suboptimal. Hence, we have studied parsimonious nonlinear alternatives, including the spread-adjusted linear pool (SLP) and the beta-transformed linear pool (BLP). Further options include generalized linear pools, consensus methods (Winkler, 1968) and nonparametric approaches, including but not limited to isotonic recursive partitioning (Luss et al., 2012). As Winkler (1986, p. 139) noted, "different combining rules are suitable for different situations".

The SLP and BLP approaches can also be used to provide calibration and dispersion corrections to a single predictive distribution, similar to the methods described by Cox (1958), Platt (1999), Zadrozny and Elkan (2002) and Primo et al. (2009) in the context of probability forecasts of a binary event. An interesting question then is whether dispersion adjustments ought to be applied to the individual components prior to the aggregation. In situations in which the components show substantially differing degrees of dispersion, or are uniformly under- or overdispersed, we indeed see potential benefits in doing this, with (here unreported) simulation experiments providing partial support to this view. In our temperature example, the components derive from point forecasts, which is the most extreme form of underdispersion, and prior to aggregating the components we apply a simple Gaussian technique that obtains approximately neutrally dispersed individual density forecasts.

Appendix A: Details for Example 2.10

Let $Z_\sigma = F_\sigma(Y)$ denote the probability integral transform of the CDF-valued random quantity F_σ. Then the random variable Z_σ has expectation $\frac{1}{2}$ and its cumulative distribution function is $H_\sigma(z) = \Phi(\sigma\,\Phi^{-1}(z))$. In particular, Z_1 is uniformly distributed. If $\sigma < 1$ then $|Z_\sigma - \frac{1}{2}|$ is stochastically larger than $|Z_1 - \frac{1}{2}|$ and therefore

$$\mathrm{var}(Z_\sigma) = \mathbb{E}(Z_\sigma - \mathbb{E}[Z_\sigma])^2 = \mathbb{E}|Z_\sigma - \tfrac{1}{2}|^2 > \mathbb{E}|Z_1 - \tfrac{1}{2}|^2 = \frac{1}{12}.$$

An analogous argument applies when $\sigma > 1$. To prove the variance formula (3), we use the fact that $\mathrm{var}(Z_\sigma) = \mathbb{E}[Z_\sigma^2] - (\mathbb{E}[Z_\sigma])^2$ and invoke the well-known expectation equality $\mathbb{E}[Z^r] = r\int_0^\infty z^{r-1}(1 - H(z))\,\mathrm{d}z$ for a nonnegative random variable Z with cumulative distribution function H, where $r > 0$.

Appendix B: Method of scoring

Here we give details for the method of scoring (see, for example, Ferguson (1986)) for numerically maximizing the mean logarithmic score or log likelihood function (12) of the BLP model as a function of the nonnegative weights $w_1, \ldots, w_k$ that

sum to 1, and transformation parameters $\alpha, \beta > 0$. Let Y denote a random variable that has a beta distribution with parameters α and β. Then

$$\begin{aligned}\frac{\partial \ell}{\partial \alpha} &= \sum_{j=1}^{J} \log\left(\sum_{i=1}^{k} w_i F_{ij}(y_j)\right) - J\,\mathbb{E}[\log Y],\\ \frac{\partial \ell}{\partial \beta} &= \sum_{j=1}^{J} \log\left(1 - \sum_{i=1}^{k} w_i F_{ij}(y_j)\right) - J\,\mathbb{E}[\log(1-Y)]\end{aligned}$$

and

$$\frac{\partial \ell}{\partial w_i} = \sum_{j=1}^{J}\left(\frac{(\alpha-1)(F_{ij}(y_j) - F_{kj}(y_j))}{\sum_{l=1}^{k} w_l F_{lj}(y_j)} - \frac{(\beta-1)(F_{ij}(y_j) - F_{kj}(y_j))}{1 - \sum_{l=1}^{k} w_l F_{lj}(y_j)} + \frac{f_{ij}(y_j) - f_{kj}(y_j)}{\sum_{l=1}^{k} w_l f_{lj}(y_j)}\right)$$

for $i = 1, \ldots, k-1$. The second derivatives are

$$\frac{\partial^2 \ell}{\partial \alpha^2} = -J\,\mathrm{var}(\log(Y)), \quad \frac{\partial^2 \ell}{\partial \beta^2} = -J\,\mathrm{var}(\log(1-Y)),$$

$$\frac{\partial^2 \ell}{\partial \alpha\,\partial \beta} = -J\,\mathrm{cov}(\log(Y), \log(1-Y))$$

and

$$\frac{\partial^2 \ell}{\partial \alpha\,\partial w_i} = \sum_{j=1}^{J} \frac{F_{ij}(y_j) - F_{kj}(y_j)}{\sum_{l=1}^{k} w_l F_{lj}(y_j)}, \qquad \frac{\partial^2 \ell}{\partial \beta\,\partial w_i} = \sum_{j=1}^{J} \frac{F_{kj}(y_j) - F_{ij}(y_j)}{1 - \sum_{l=1}^{k} w_l F_{lj}(y_j)}$$

for $i = 1, \ldots, k-1$, while

$$\begin{aligned}\frac{\partial^2 \ell}{\partial w_{i_1}\partial w_{i_2}} &= -\sum_{j=1}^{J} \frac{(f_{i_1 j}(y_j) - f_{kj}(y_j))(f_{i_2 j}(y_j) - f_{kj}(y_j))}{(\sum_{l=1}^{k} w_l f_{lj}(y_j))^2}\\ &\quad -\sum_{j=1}^{J}\left(\frac{\alpha-1}{(\sum_{l=1}^{k} w_l F_{lj}(y_j))^2} + \frac{\beta-1}{(1-\sum_{l=1}^{k} w_l F_{lj}(y_j))^2}\right)\\ &\quad \times (F_{i_1 j}(y_j) - F_{kj}(y_j))\,(F_{i_2 j}(y_j) - F_{kj}(y_j))\end{aligned}$$

for $i_1 = 1, \ldots, k-1$ and $i_2 = 1, \ldots, k-1$. The method of scoring now applies Newton's algorithm to optimize the likelihood as a function of the parameter vector.

References

Allard, A., Comunian, A. and Renard, P. (2012). Probability aggregation methods in geoscience. *Mathematical Geosciences*, **44**, 545–581. MR2947804

Arnold, H. M., Moroz, I. M. and Palmer, T. N. (2013). Stochastic parameterizations and model uncertainty in the Lorenz '96 system. *Philiosophical Transactions of the Royal Society Ser. A*, **371**, 20110479.

Berrocal, V. J., Raftery, A. E. and Gneiting, T. (2007). Combining spatial statistical and ensemble information for probabilistic weather forecasting. *Monthly Weather Review*, **135**, 1386–1402.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity, *Journal of Econometrics*, **31**, 307–327. MR0853051

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, **135**, 1512–1519.

Clemen, R. T. and Winkler, R. L. (2007). Aggregating probability distributions. In Ward, E., Miles, R. F. and von Winterfeldt, D. (eds.), *Advances in Decision Analysis: From Foundations to Applications*, Cambridge University Press, pp. 154–176.

Clements, M. P. and Harvey, D. I. (2011). Combining probability forecasts. *International Journal of Forecasting*, **27**, 208–223.

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, **45**, 562–565.

Czado, C., Gneiting, T. and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, **65**, 1254–1261. MR2756513

Dawid, A. P. (1984). Statistical theory: The prequential approach (with discussion and rejoinder). *Journal of the Royal Statistical Society Ser. A*, **147**, 278–290. MR0763811

Dawid, A. P. (1986). Probability forecasting, in Kotz, S., Johnson, N. L. and Read, C. B. (eds.), *Encyclopedia of Statistical Sciences*, Vol. 7, Wiley, New York, pp. 210–218. MR0892738

Dawid, A. P., DeGroot, M. H. and Mortera, J. (1995). Coherent combination of experts' opinions (with discussion and rejoinder). *Test*, **4**, 263–313. MR1379793

DelSole, T. and Shukla, J. (2010). Model fidelity versus skill in seasonal forecasting. *Journal of Climate*, **23**, 4794–4806.

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

Dutton, J. A. (2002). Opportunities and priorities in a new era for weather and climate services. *Bulletin of the American Meteorological Society*, **83**, 1303–1311.

Eckel, F. A. and Mass, C. F. (2005). Aspects of effective short-range ensemble forecasting. *Weather and Forecasting*, **20**, 328–350.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman and Hall. MR1699953

Fraley, C., Raftery, A. E., Gneiting, T., Sloughter, J. M. and Berrocal, V. J. (2011). Probabilistic weather forecasting in R. *R Journal*, **3(1)**, 55–63.

French, S. and Ríos Insua, D. (2000). *Statistical Decision Theory*. Arnold, London.

Fricker, T. E., Ferro, C. A. T. and Stephenson, D. B. (2013). Three recommendations for evaluating climate predictions. *Meteorological Applications*, **20**, 246–255.

Garratt, A., Mitchell, J., Vahey, S. P. and Wakerly, E. C. (2011). Real-time inflation forecast densities from ensemble Phillips curves. *North American Journal of Economics and Finance*, **22**, 77–87.

Genest, C. (1984a). A characterization theorem for externally Bayesian groups. *Annals of Statistics*, **12**, 1100–1105. MR0751297

Genest, C. (1984b). A conflict between two axioms for combining subjective distributions. *Journal of the Royal Statistical Society Ser. B*, **46**, 403–405. MR0790625

Genest, C. and Zidek, J. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, **1**, 114–135. MR0833278

Genest, C., McConway, K. J. and Schervish, M. J. (1986). Characterization of externally Bayesian pooling operators. *Annals of Statistics*, **14**, 487–501. MR0840510

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, **164**, 130–141. MR2821798

Glahn, B., Peroutka, M., Wiedenfeld, J., Wagner, J., Zylstra, G., Schuknecht, B. and Jackson, B. (2009). MOS uncertainty estimates in an ensemble framework. *Monthly Weather Review*, **137**, 246–268.

Gneiting, T. (2008). Editorial: Probabilistic forecasting. *Journal of the Royal Statistical Society Ser. A*, **171**, 319–321. MR2427336

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

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378. MR2345548

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Ser. B*, **69**, 243–268. MR2325275

Hall, S. G. and Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, **23**, 1–13.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, **129**, 550–560.

Held, L., Rufibach, K. and Balabdaoui, F. (2010). A score regression approach to assess calibration of continuous probabilistic predictions. *Biometrics*, **66**, 1295–1305. MR2758518

Hora, S. C. (2004). Probability judgements for continuous quantities: Linear combinations and calibration. *Management Science*, **50**, 597–604.

Hora, S. C. (2010). An analytic method for evaluating the performance of aggregation rules for probability densities. *Operations Research*, **58**, 1440–1449. MR2560546

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: A tutorial (with discussion). *Statistal Science*, **4**, 382–417. MR1765176

Jore, A. S., Mitchell, J. and Vahey, S. P. (2010). Combining forecast densities from VARs with uncertain stabilities. *Journal of Applied Econometrics*, **25**, 621–634. MR2758076

Jouini, M. N. and Clemen, R. T. (1996). Copula models for aggregating expert opinions. *Operations Research*, **44**, 444–457.

Kascha, C. and Ravazzolo, F. (2010). Combining inflation density forecasts. *Journal of Forecasting*, **29**, 231–250. MR2752012

Kleiber, W., Raftery, A. E., Baars, J., Gneiting, T., Mass, C. F. and Grimit, E. (2011). Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Monthly Weather Review*, **139**, 2630–2649.

Krüger, F. (2013). Jensen's inequality and the success of linear prediction pools. Discussion paper, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2080010.

Luss, R., Rosset, S. and Shahar, M. (2012). Efficient regularized isotonic regression with application to gene-gene interaction search. *Annals of Applied Statistics*, **6**, 253–283. MR2951537

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Mason, S. J., Galpin, S., Goddard, L., Graham, N. E. and Rajaratnam, B. (2007). Conditional exceedance probabilities. *Monthly Weather Review*, **135**, 363–372.

McConway, K. J. (1981). Marginalization and linear opinion pools. *Journal of the American Statistical Association*, **76**, 410–414. MR0624342

Mitchell, J. and Hall, S. G. (2005). Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR 'fan' charts of inflation. *Oxford Bulletin of Economics and Statistics*, **67**, 995–1033.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330–1338.

Murphy, A. H. and Winkler, R. L. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting*, **7**, 435–455.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Smola, A., Bartlett, P., Schölkopf, B. and Schuurmans, D., eds., *Advances in Large Margins Classifiers*, MIT Press, pp. 61–74.

Primo, C., Ferro, C. A. T., Jolliffe, I. T. and Stephenson, D. B. (2009). Calibration of probabilistic forecasts of binary events. *Monthly Weather Review*, **137**, 1142–1149.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–1174.

Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society Ser. B*, **72**, 71–91. MR2751244

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, **23**, 470–472. MR0049525

Rüschendorf, L. (1981). Stochastically ordered distributions and monotonicity of the OC-function of sequential probability ratio tests. *Mathematische Operationsforschung Ser. Statistics*, **12**, 327–338. MR0640553

Schervish, M. J. (1989). A general method for comparing probability assessors. *Annals of Statistics*, **17**, 1856–1879. MR1026316

Sloughter, J. M., Raftery, A. E., Gneiting, T. and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, **135**, 3209–3220.

Stone, M. (1961). The linear pool. *Annals of Mathematical Statistics*, **32**, 1339–1342. MR0135190

Tsyplakov, A. (2011). Evaluating density forecasts: a comment. MPRA paper no. 31233, http://mpra.ub.uni-muenchen.de/32728.

Tsyplakov, A. (2013). Evaluation of probabilistic forecasts: proper scoring rules and moments. Preprint, http://dx.doi.org/10.2139/ssrn.2236605.

Wallis, K. F. (2005). Combining density and interval forecasts: A modest proposal. *Oxford Bulletin of Economics and Statistics*, **67**, 983–994.

Winkler, R. L. (1968). The consensus of subjective probability distributions. *Management Science*, **15**, B61–B75.

Winkler, R. L. (1986). Comment on "Combining probability distributions: A critique and an annotated bibliography". *Statistical Science*, **1**, 138–140. MR0833278

Wuertz, D. and Rmetrics Core Team (2007). The fGarch Package. Reference manual, available at http://www.mirrorservice.org/sites/lib.stat.cmu.edu/R/CRAN/doc/packages/f Garch.pdf.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 694–699.

Zarnowitz, V. (1969). The new ASA-NBER survey of forecasts by economic statisticians. *American Statistician*, **23**, 12–16.