Combining Probability Forecasts

Roopesh Ranjan and Tilmann Gneiting

Technical Report no. 543

Department of Statistics, University of Washington

October 2008

Abstract

Linear pooling is by the far the most popular method for combining probability forecasts. However, any nontrivial weighted average of two or more distinct, calibrated probability forecasts is necessarily uncalibrated and lacks sharpness. In view of this, linear pooling requires recalibration, even in the ideal case in which the individual forecasts are calibrated. Toward this end, we propose a beta transformed linear opinion pool (BLP) for the aggregation of probability forecasts from distinct, calibrated or uncalibrated sources. The BLP method fits an optimal nonlinearly recalibrated forecast combination, by compositing a beta transform and the traditional linear opinion pool. The technique is illustrated in a simulation example and in a case study on statistical and National Weather Service probability of precipitation forecasts.

KEYWORDS: calibration, forecast combination, linear opinion pool, model averaging, probability forecasting, reliability, resolution, sharpness

1 Introduction

Probabilistic forecasts take account of the uncertainty in a prediction, by taking the form of a predictive probability distribution for a future quantity or event. The simplest case is that of a future binary or dichotomous event, such as a recession versus no recession, or rain versus no rain. In the binary case, a predictive probability distribution is simply an ex ante probability for the event to happen. While the roots of probability forecasting can be traced back to the 18th century, the transition to probability of precipitation forecasts by the US National Weather Service in 1965 was perhaps the most influential and important event in their development (Murphy 1998; Winkler and Jose 2008). In economics, the Survey of Professional Forecasters has included probability variables since 1968 (Croushore 1993). Of course, there are many other important applications of probability forecasts, including but not limited to medical diagnosis (Wilson et al. 1998; Pepe 2003), educational

testing, and political and socio-economic foresight (Tetlock 2005). Arguably, a far-reaching transdisciplinary transition to distributional forecasting is well under way (Gneiting 2008).

In many instances, multiple probability forecasts for the same event are available. In surveys, economic experts might provide diverse probability assessments of a future recession. Distinct numerical and/or statistical models might provide a collection of probability of precipitation forecasts, and a group of physicians might assign individual survival probabilities. In this type of situation, there is strong empirical evidence that combined probability forecasts that draw an all the experts' or models' strengths result in improved predictive performance. This is very much in the spirit of model averaging, which has primarily been developed for the purpose of statistical inference (Hoeting et al. 1999).

Various ways of combining probability forecasts into a single aggregated forecast have been proposed. Genest and Zidek (1986), Wallsten et al. (1997), Clemen and Winkler (1999, 2007) and Primo et al. (2008) provide excellent reviews. In practice, most aggregation techniques rely on a weighted linear combination of the individual probability forecasts, which is often referred to as a linear opinion pool. Substantial empirical evidence attests to the benefits of linear opinion pools, with successful applications ranging from meteorology (Sanders 1963; Vislocky and Fritsch 1995; Baars and Mass 2005) to economics (Graham 1996), psychology (Ariely et al. 2000), and medical diagnosis (Winkler and Poses 1993), among other fields.

The goal in probability forecasting is to maximize the sharpness of the forecasts subject to calibration (Murphy and Winkler 1987; Gneiting, Balabdaoui and Raftery 2007). Calibration or reliability measures how close conditional event frequencies are to the forecast probabilities. Sharpness describes how far away the forecasts are from the naive, climatological baseline forecast, that is, the marginal event frequency (Gneiting et al. 2008; Winkler and Jose 2008). The more extreme the forecast probabilities are, that is, the closer to the most confident values of zero or one, the sharper the forecast. Strictly proper scoring rules such as the Brier or quadratic score (Brier 1950; Selten 1998) and the logarithmic score (Good 1952) provide summary measures of predictive performance that address calibration and sharpness simultaneously (Gneiting and Raftery 2007).[1]

It is therefore critical that probability assessments are aggregated in ways that promote calibrated and sharp combined forecasts. In Section 2 we demonstrate a striking result, in that any weighted linear combination of distinct, individually calibrated probability forecasts is necessarily uncalibrated and lacks sharpness. In this light, linear opinion pools are suboptimal, so in Section 3 we propose a nonlinear generalization, the beta-transformed linear opinion pool (BLP). The BLP method fits an optimally recalibrated forecast combination,

[1] A scoring rule assigns a numerical score, $\mathrm{S}(x, y)$, to the probability forecast $x \in [0, 1]$ and the binary event y, where $y = 1$ if the event occurs and $y = 0$ otherwise. We consider scoring rules to be negatively oriented penalties, that is, the smaller the better. A scoring rule is strictly proper if it encourages honest assessments, that is, if

$$x\,\mathrm{S}(x, 1) + (1 - x)\mathrm{S}(x, 0) < x\,\mathrm{S}(x', 1) + (1 - x)\mathrm{S}(x', 0) \qquad \text{for all} \qquad 0 \le x \ne x' \le 1.$$

See Dawid (1986), Winkler (1996) and Gneiting and Raftery (2007) for reviews and discussion.

by compositing a beta transform and the traditional linear opinion pool. Section 4 illustrates the BLP method in a case study on statistical and National Weather Service probability of precipitation forecasts at 29 major cities in the continental US. The BLP combined forecast is calibrated and sharp and outperforms the individual and linearly combined forecasts. The paper closes with a discussion in Section 5.

2 Some shortcomings of linearly combined probability forecasts

The overarching message in this section is that linear opinion pools are generally uncalibrated, even in the ideal case in which each individual source is calibrated. We give a rigorous probabilistic version of this result in Theorem 2.1, which is then illustrated in a simulation study.

2.1 Theoretical results

We work within a probabilistic framework which considers the joint distribution of the random vector

$$(Y, p_1, \ldots, p_k),$$

where $Y \in \{0, 1\}$ is a binary or dichotomous event, and $0 \leq p_1, \ldots, p_k \leq 1$ are probability forecasts that take values in the closed unit interval. This is akin to the setting in DeGroot and Fienberg (1982, 1983) and Murphy and Winkler (1987), but considers an arbitrary number, k, of individual probability forecasts, each of which is a random variable, with full generality in the joint dependence structure. In this framework a probability forecast is any random variable, p, that is measurable with respect to the σ-algebra generated by $p_1, \ldots, p_k$, with the linear opinion pool,

$$p = w_1 p_1 + \cdots + w_k p_k \quad \text{where} \quad w_1, \ldots, w_k \geq 0 \quad \text{and} \quad w_1 + \cdots + w_k = 1, \tag{1}$$

being one such example. The probability forecast p is calibrated for Y if[2]

$$\mathbb{P}(Y = 1|p) = \mathbb{E}(Y|p) = p \qquad \text{almost surely.}$$

From the basic properties of conditional expectations, it is immediate that if p is a calibrated probability forecast then

$$\mathbb{E}p = \mathbb{E}\mathbb{E}(Y|p) = \mathbb{E}Y.$$

[2]This definition is in accordance with the economic, psychological, statistical and meteorological forecasting literature and can be traced to Murphy and Winkler (1987) and Schervish (1989). It differs from the game-theoretic approach to calibration that has been developed in a far-reaching, related strand of literature (Dawid 1982; Foster and Vohra 1998; Lehrer 2001; Sandroni, Smorodinsky and Vohra 2003; Vovk and Shafer 2005; Al-Najjar and Weinstein 2008; Feinberg and Stewart 2008).

This latter property can be thought of as a weak form of calibration, and we refer to it as marginal consistency. It resembles the notion of marginal calibration for probabilistic forecasts of continuous variables (Gneiting, Balabdaoui and Raftery 2007).

We are now in a position to state our key result. The proof is deferred to the Appendix.

Theorem 2.1. *Suppose that $p_1, \ldots, p_k$ are calibrated for the binary event Y and such that $p_i \neq p_j$ with strictly positive probability for at least one pair $i \neq j$. Consider the linear opinion pool,*

$$p = w_1 p_1 + \cdots + w_k p_k,$$

where $w_1, \ldots, w_k > 0$ and $w_1 + \cdots + w_k = 1$. Let

$$q = \mathbb{P}(Y = 1|p) = \mathbb{E}(Y|p)$$

denote the recalibrated version of p, that is, the conditional probability of Y given p. Then the following holds.

(a) *The linear opinion pool p lacks calibration, in that $q \neq p$ with strictly positive probability.*

(b) *The linear opinion pool p lacks sharpness, in that*

$$\mathbb{E}(p - p_0)^2 < \mathbb{E}(q - p_0)^2 \qquad \textit{where} \qquad p_0 = \mathbb{E}p = \mathbb{E}q = \mathbb{E}Y.$$

In words, both p and q are marginally consistent, but on average p is closer to its expectation, the naive climatological forecast p_0, than its recalibrated version, q.

(c) *The recalibrated forecast q is calibrated, that is, $\mathbb{P}(Y = 1|q) = q$ almost surely, and it outperforms p, in that*

$$\mathbb{E}\mathrm{S}(q, Y) < \mathbb{E}\mathrm{S}(p, Y)$$

for every strictly proper scoring rule S.

The statement about the lack of calibration of the linear opinion pool in part (a) is our main result.[3] Part (b) concerns a lack of sharpness, which we express in terms of the expected deviation from the climatological baseline probability, p_0. For a sharp forecast, the forecast probabilities are close to zero or one, so the larger this deviation the sharper the forecast (Murphy and Winkler 1992; Gneiting et al. 2008; Winkler and Jose 2008).[4] As a result, the linear opinion pool is underconfident. Part (c) demonstrates the superiority of the recalibrated forecast in terms of strictly proper scoring rules (Gneiting and Raftery 2007)

[3] A similar result that applies to the case of multiple density forecasts for a continuous quantity was proved by Hora (2004). This uses a very different mode of calibration, and there is no apparent way of deducing our result from Hora's, or vice versa.

[4] It is readily seen that $\mathbb{E}(p - c)^2 < \mathbb{E}(q - c)^2$ irrespectively of the choice of the baseline probability, c.

and is akin to Theorem 6.3 of Schervish (1989). Proper scoring rules address calibration and sharpness simultaneously, so in view of parts (a) and (b) this is an unsurprising result.

We proceed to discuss related results in the literature. Theorem 2 of Wallsten and Diederich (2001) considers the combination of expert probability judgements, assuming that the assessments are conditionally independent and that each expert's expressed (overt) opinion is a monotone stochastic transform of a hidden (covert) opinion which is calibrated. Then the simple average of the expert opinions becomes increasingly diagnostic of the future event as the number of experts grows to infinity, roughly in the sense that if the average exceeds $\frac{1}{2}$ the true conditional probability of the event converges to 1, and otherwise converges to 0. In other words, the theoretical calibration curve becomes sigmoidal with a fixed point at $\frac{1}{2}$. In contrast to our Theorem 2.1, which is a finite sample result and does not make any assumptions on the dependence structure, Wallsten and Diederich (2001) rely critically on the asymptotic scenario and conditional independence.

Another related result is Theorem 4.1 of Genest and Schervish (1985), which adopts a Bayesian point of view and derives a formula for the posterior opinion of a decision maker. As Wallsten and Diederich (2001) note, the posterior opinion converges to 0 or 1 if the individual judgements lie below or above $\frac{1}{2}$. This result also depends on the conditional independence of the individual probability assessments.

Despite Theorem 2.1 being critical of the linear opinion tool, there is overwhelming empirical evidence that linearly combined probability forecasts outperform individual forecasts. This is not a contradiction and can readily be explained, by noting that linear opinion pools outperform individual forecasts, but are suboptimal themselves, and can potentially be improved upon by using nonlinear recalibration methods.

2.2 Simulation study

We now illustrate our theoretical findings in a simulation study. First we describe a statistical model that gives rise to a joint distribution for the binary event Y and probability forecasts p_1 and p_2, which represent forecasters with access to independent sources of information. Then we define linearly combined forecasts and assess calibration.

Specifically, let

$$p = \Phi(a_1 + a_2),$$

where $a_1 \sim \mathcal{N}(0, 1)$ and $a_2 \sim \mathcal{N}(0, 2)$ are independent random variables and Φ denotes the standard normal cumulative distribution function. Suppose that Y is a Bernoulli random variable with conditional success probability

$$\mathbb{P}(Y = 1 \,|\, p) = \mathbb{E}(Y \,|\, p) = p.$$

Forecaster 1 has access to a_1 only. This assessor's **probability forecast p_1** is the conditional

Table 1: Maximum likelihood estimates of OLP and BLP parameters in the simulation example, with standard errors in brackets.

Method	w_1	w_2	α
OLP	0.246 (0.014)	0.754 (0.014)	
BLP	0.519 (0.005)	0.481 (0.005)	9.55 (0.35)

event probability

$$p_1 = \mathbb{P}(Y = 1|a_1) = \mathbb{E}(Y|a_1) = \mathbb{E}(p|a_1) = \mathbb{E}\left[\Phi(a_1 + a_2)|a_1\right] = \Phi\left(\frac{a_1}{\sqrt{3}}\right). \tag{2}$$

The second forecaster has knowledge of source a_2 only, whence **probability forecast p_2** becomes

$$p_2 = \mathbb{P}(Y = 1|a_2) = \Phi\left(\frac{a_2}{\sqrt{2}}\right). \tag{3}$$

A detailed derivation of the final equality in (2) and (3) is given in the Appendix. Evidently, p, p_1 and p_2 are calibrated.

We take p_1 and p_2 as the individual forecasts from which we form combinations, namely the **equally weighted linear opinion pool (ELP)**, that is, the equally weighted average of p_1 and p_2, and an **optimally weighted linear opinion pool (OLP)**. The OLP weights for p_1 and p_2 are estimated on a training sample of size 10,000, using the maximum likelihood method and the special case of the log likelihood function (9) below, in which $\alpha = \beta = 1$. Table 1 shows the OLP estimates and their standard errors. The more resolved individual forecast, p_2, obtains a substantially higher OLP weight, w_2, of about $\frac{3}{4}$.

In the simulation experiment, we consider an independent test sample of size 10,000 from the joint distribution of Y, p_1 and p_2 and generate the combined ELP and OLP forecasts. Figure 1 shows empirical calibration curves or reliability diagrams (Sanders 1963; Pocernich 2008) for the four types of forecasts, which plot the conditional empirical event frequency versus the forecast probability. The red circles show the conditional empirical frequency; the broken lines give pointwise 95% lower and upper critical values under the null hypothesis of calibration, obtained with the bootstrap technique of Bröcker and Smith (2007). Significant deviations from the diagonal suggest that a forecast is uncalibrated. The inset histograms show the frequency distribution of the forecast probabilities and can be used diagnostically to assess sharpness.

The calibration curves for the individual forecasts, p_1 and p_2, show that they are empirically well calibrated, and the inset histograms confirm that p_2 is the more resolved, sharper forecast, with forecast probabilities that are further away from the climatological event frequency, $p_0 = \frac{1}{2}$. The linearly pooled ELP and OLP forecasts are empirically uncalibrated. The direction of departure is as anticipated, towards underconfidence, and the extent of the lack of calibration is startling, even for the optimally weighted OLP forecast.

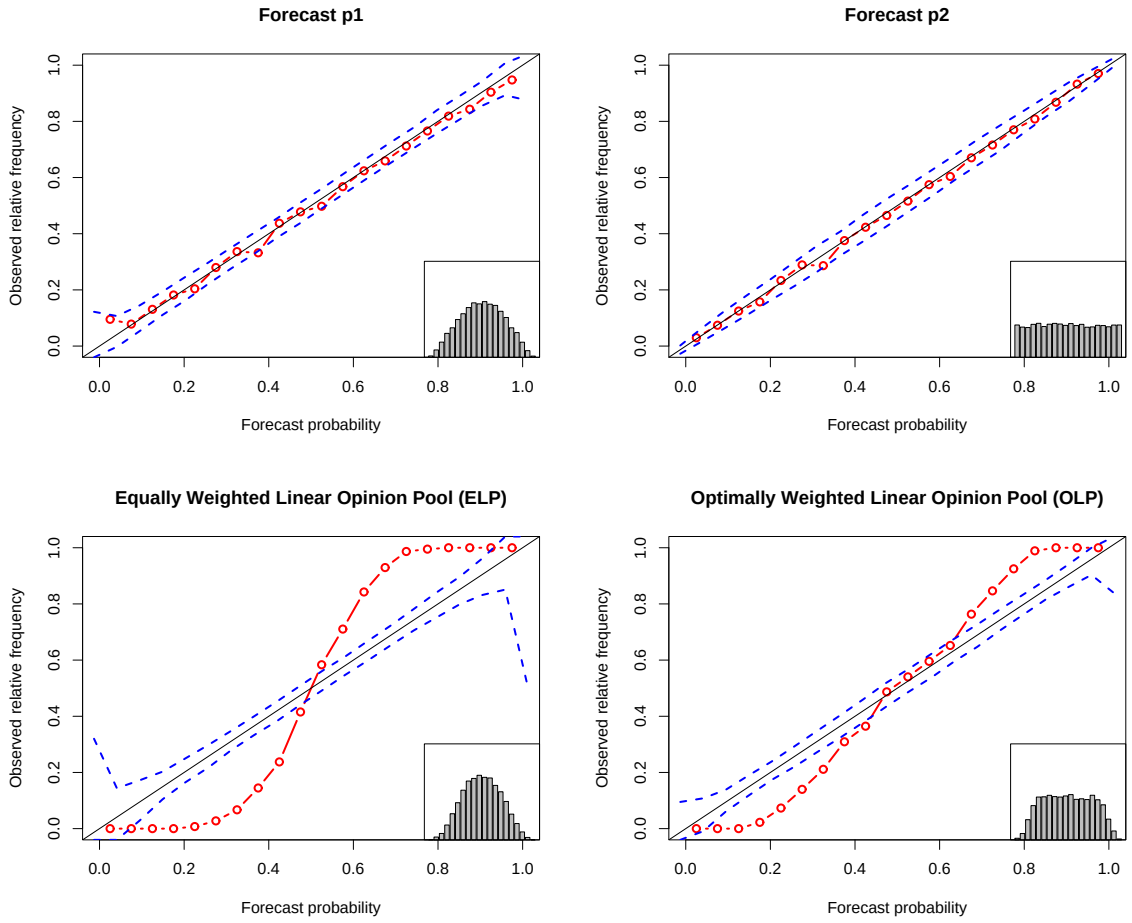

Figure 1: Calibration curves and 95% bootstrap intervals under the null hypothesis of calibration for the individual and linearly combined forecasts in the simulation example. The histograms show the empirical distribution of the forecast values over the unit interval.

3 Recalibration

We have seen that the linear opinion pool yields a suboptimal combined probability forecast, in that it is uncalibrated even in the ideal case in which the individual sources are calibrated. If the individual forecasts are uncalibrated, the need for recalibration typically is even more pronounced. Before proposing a method that addresses these issues, by applying a recalibration transform to the linear opinion pool, we digress to discuss a theoretically optimal approach to forecast aggregation.

We have chosen to work in a probabilistic setting that considers the joint distribution of the binary event and the individual probability forecasts.[5] In this framework, the theoretically

[5]Bayesian approaches may suggest other types of recalibration techniques. For example, Lindley (1982)

optimal combined forecast, $\hat{p}$, is the conditional probability (CP), or conditional expectation of the binary event Y, given the individual forecasts $p_1, \ldots, p_k$, that is,

$$\hat{p} = \mathbb{P}(Y = 1|p_1, \ldots, p_k) = \mathbb{E}(Y|p_1, \ldots, p_k). \tag{4}$$

By definition, this is the best approximation of the binary random variable Y in terms of the individual probability forecasts, $p_1, \ldots, p_k$, in the sense that $\mathbb{E}(\hat{p} - Y)^2 \leq \mathbb{E}(p - Y)^2$ for all functions p that are measurable with respect to the σ-algebra generated by $p_1, \ldots, p_k$. Hence, $\hat{p}$ minimizes the expected Brier score and, indeed, the expectation of any proper scoring rule S, in that

$$\begin{aligned}
\mathbb{E}\mathrm{S}(\hat{p}, Y) &= \mathbb{E}\mathbb{E}[\mathrm{S}(\hat{p}, Y)|p_1, \ldots, p_k] \\
&= \mathbb{E}[\hat{p}\mathrm{S}(\hat{p}, 1) + (1 - \hat{p})\mathrm{S}(\hat{p}, 0)] \\
&\leq \mathbb{E}[p\mathrm{S}(\hat{p}, 1) + (1 - p)\mathrm{S}(\hat{p}, 0)] \\
&= \mathbb{E}\mathbb{E}[\mathrm{S}(p, Y)|p_1, \ldots, p_k] \\
&= \mathbb{E}\mathrm{S}(p, Y),
\end{aligned}$$

with equality if and only if $p = \hat{p}$ almost surely. Under the conditions of Theorem 2.1, the conditional probability $\hat{p}$ is a necessarily nonlinear function of the individual forecasts. For example, in the simulation study in Section 2.2 there are two individual forecasts, p_1 and p_2, and the conditional probability (4) equals

$$\begin{aligned}
\hat{p} &= \mathbb{P}(Y = 1|p_1, p_2) \\
&= \mathbb{P}(Y = 1|a_1, a_2) \\
&= \Phi(a_1 + a_2) \\
&= \Phi(\sqrt{3}\,\Phi^{-1}(p_1) + \sqrt{2}\,\Phi^{-1}(p_2)).
\end{aligned} \tag{5}$$

3.1 The beta-transformed linear opinion pool (BLP)

In the practice of forecasting, the functional form of the conditional probability (4) is unknown and needs to be estimated from training data. Nonparametric approaches are feasible; however, we prefer parsimonious, yet flexible parametric approximations. Specifically, our preferred approach to aggregating individual probability forecasts, $p_1, \ldots, p_k$, is to first form a linear opinion pool, and then to apply a beta transform to achieve calibration. We call this the beta-transformed linear opinion pool (BLP), which takes the form

$$p = H_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i p_i\right), \tag{6}$$

suggested a way of recalibrating probability judgements in a Bayesian setting. Clemen and Winkler (1987) applied Lindley's method to National Weather Service probability of precipitation forecasts and noted little improvement over the individual sources.

where $w_1, \ldots, w_k \geq 0$ and $w_1 + \cdots + w_k = 1$, and

$$H_{\alpha,\beta}(x) = B(\alpha,\beta)^{-1} \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t \qquad \text{for} \qquad x \in [0,1]$$

is the cumulative distribution function of the beta density with shape parameters $\alpha > 0$ and $\beta > 0$. Note that the BLP model nests the traditional linear opinion pool that arises in the special case when $\alpha = \beta = 1$. If furthermore $w_1 = \cdots = w_k = \frac{1}{k}$ we recover the equally weighted linear opinion pool (ELP). While the use of the beta transform for the purpose of calibration dates back at least to Graham (1996), the statistical model (6) that merges the linear opinion pool with a parametric recalibration transformation appears to be new. It applies very generally and can be used to aggregate calibrated as well as uncalibrated sources.

In many cases, full generality in (6) may not be needed or desirable. For instance, it is often useful to assume that the recalibration transform, $H_{\alpha,\beta}$, satisfies

$$H_{\alpha,\beta}(x) \leq x \quad \text{for} \quad x \leq x_0 \qquad \text{and} \qquad H_{\alpha,\beta}(x) \geq x \quad \text{for} \quad x \geq x_0 \tag{7}$$

for some $x_0 \in (0,1)$. This can be enforced by putting conditions on α and β. For example, if the individual forecasts are calibrated, Theorem 2.1 suggests that the linear opinion pool is underconfident, in the sense that its calibration curve lies under the diagonal for small forecast probabilities, and above the diagonal for high probabilities, with a fixed point at some $x_0 \in (0,1)$. The theoretical results of Wallsten and Diederich (2001) support the choice of $x_0 = \frac{1}{2}$, under which (7) can be enforced by requiring that

$$\alpha = \beta \geq 1. \tag{8}$$

If we aim to address the hard-easy effect that has been described in the psychological literature (Lichtenstein, Fischhoff and Phillips 1982; Kynn 2008, p. 253) the fixed point in (7) can be taken to be $x_0 = \frac{3}{4}$.

We now describe how we go about parameter estimation for the BLP model in (6). Suppose that $y_1, \ldots, y_n$ are binary observations in the training set. Let $p_{i1}, \ldots, p_{in}$ denote the respective individual probability forecasts, for sources $i = 1, \ldots, k$. The aggregated BLP forecast then takes the form

$$p_t = H_{\alpha,\beta}\left(\sum_{i=1}^k w_i p_{it}\right) \qquad \text{for} \qquad t = 1, \ldots, n,$$

where the index ranges over the instances in the training set. Assuming independence, the log likelihood function for the BLP model (6) can be expressed as

$$\begin{aligned}\ell(w_1, \ldots, w_k; \alpha, \beta) \;&=\; \sum_{t=1}^n \left(y_t \log p_t + (1-y_t)\log(1-p_t)\right) \\ &=\; \sum_{t=1}^n y_t \log H_{\alpha,\beta}\left(\sum_{i=1}^k w_i p_{it}\right) + \sum_{t=1}^n (1-y_t)\log\left(1 - H_{\alpha,\beta}\left(\sum_{i=1}^k w_i p_{it}\right)\right). \end{aligned} \tag{9}$$

We obtain maximum likelihood estimates of the weights $w_1, \ldots, w_k$ and the recalibration parameters α and β by numerically optimizing the log likelihood function (9) under the constraints that $w_1, \ldots, w_k \geq 0$, $w_1 + \cdots + w_k = 1$, $\alpha > 0$ and $\beta > 0$. As noted above, it is often appropriate to enforce further constraints, with (8) being one such example. The traditional, non-transformed linear opinion pool arises when $\alpha = \beta = 1$. Estimated standard errors can be obtained in the usual way, by inverting a numerical approximation to the Hessian of the log likelihood function at the maximum likelihood estimates. The estimates can also be interpreted as optimum score estimates based on the logarithmic scoring rule, in the sense described by Gneiting and Raftery (2007, p. 375) and Elliott and Timmermann (2008, p. 20). This latter interpretation does not rely on any assumption of independence.

3.2 Simulation study revisited

We return to the simulation study in Section 2.2 and fit the **beta-transformed linear opinion pool (BLP)** to the individual probability forecasts, p_1 and p_2. Then we compare to the theoretically optimal forecast, the **conditional probability (CP)** forecast (4) which here has the closed form solution (5).

Recall that both p_1 and p_2 are calibrated, so we estimate the BLP model (6) under the constraint in (8), that is, we assume that $\alpha = \beta \geq 1$. Table 1 shows maximum likelihood estimates for the BLP parameters and compares to the respective OLP estimates. The individual forecasts, p_1 and p_2, get approximately equal weights, much in contrast to the OLP model. The estimate for the BLP recalibration parameter, α, is far from the identity transform that arises when $\alpha = 1$, reflecting the striking lack of calibration of the traditional linear opinion pool.

Have we succeeded in our goal of approximating the theoretically optimal CP forecast (5) by the estimated, nonlinearly aggregated BLP model (6)? The empirical calibration curve for the BLP forecast in Figure 2 does not show any systematic departure from the diagonal, and the inset histogram shows that it is much sharper than any of the individual or linearly combined forecasts. A more detailed analysis reveals that if $0 < p_1 = p_2 < 1$ the maximal difference between the CP forecast and the fitted BLP model is 0.0215.

Table 2 shows the mean Brier or quadratic score and its reliability, resolution and uncertainty components for the various forecasts (Murphy 1973; Dawid 1986). Suppose that the probability forecasts p_t for the binary event y_t, where $t = 1, \ldots, n$, take discrete values $f_i \in [0, 1]$, where $i = 1, \ldots, I$. Let n_i be the number of times that the forecast value f_i occurs, so that $n = n_1 + \cdots + n_I$, and let q_i be the respective empirical conditional event frequency, that is, the ex post recalibrated forecast. Let

$$\bar{q} = \frac{1}{n} \sum_{i=1}^{I} n_i q_i = \frac{1}{n} \sum_{t=1}^{n} y_t$$

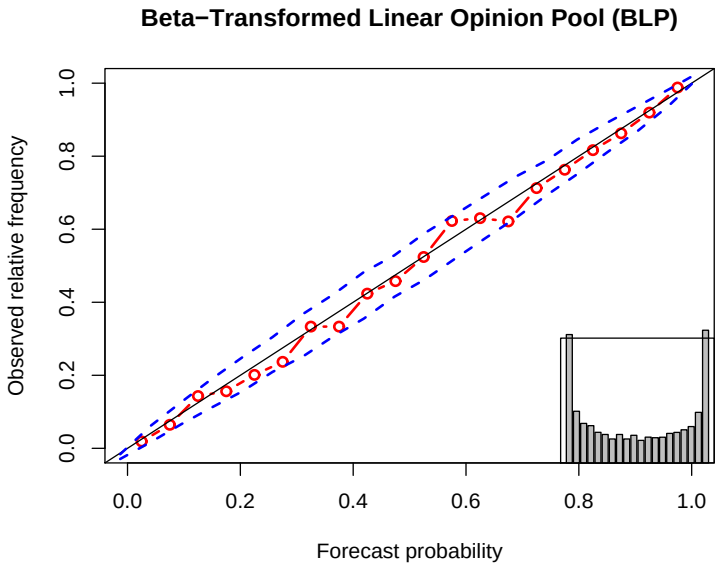

Figure 2: Calibration curve and 95% bootstrap intervals under the null hypothesis of calibration for the BLP forecast in the simulation example. The histogram shows the empirical distribution of the forecast values over the unit interval.

denote the marginal event frequency. Then the mean Brier score,

$$\mathrm{BS} = \frac{1}{n}\sum_{t=1}^{n}(p_t - y_t)^2,$$

decomposes as

$$\mathrm{BS} = \underbrace{\frac{1}{n}\sum_{i=1}^{I} n_i(f_i - q_i)^2}_{\mathrm{REL}} - \underbrace{\frac{1}{n}\sum_{i=1}^{I} n_i(q_i - \bar{q})^2}_{\mathrm{RES}} + \underbrace{\bar{q}(1-\bar{q})}_{\mathrm{UNC}}.$$

The reliability term (REL) quantifies calibration and is negatively oriented, that is, the smaller the better. The resolution component (RES) equals the variance of the ex post recalibrated forecast and is positively oriented. For a calibrated forecast, it quantifies sharpness; for an uncalibrated forecast, it measures potential sharpness. As noted above, we generally seek a forecast which is as sharp as possible subject to it being calibrated (Murphy and Winkler 1987; Gneiting, Balabdaoui and Raftery 2007). The uncertainty term (UNC) is computed from the observations alone and independent of the forecast.[6]

[6]If the probability forecast is a continuous variable, the decomposition depends on a binning of the forecast values and is approximate only. It can be made exact by considering two additional components in the decomposition, as recently proposed by Stephenson, Coelho and Jolliffe (2008). In our case, the extra terms make very little difference, and we consider the classical decomposition only.

Table 2: Mean Brier score (BS) and its reliability (REL), resolution (RES) and uncertainty (UNC) components for the probability forecasts in the simulation example.

Forecast	BS	REL	RES	UNC
p_1	0.2113	0.0003	0.0389	0.2500
p_2	0.1685	0.0002	0.0816	0.2500
ELP	0.1590	0.0382	0.1291	0.2500
OLP	0.1563	0.0111	0.1048	0.2500
BLP	0.1199	0.0004	0.1305	0.2500
CP	0.1186	0.0004	0.1318	0.2500

From Table 2 we see that the linearly combined ELP and OLP forecasts have lower Brier score than any of the individual forecasts. In both cases, the improvement stems from the resolution component, which is high, because the ex post recalibrated forecast is sharp, even though the forecast itself is uncalibrated and lacks sharpness, as reflected in Figure 1. The BLP forecast is much better calibrated, and simultaneously more resolved, than the ELP and OLP forecasts, resulting in a hugely improved Brier score. As anticipated, the theoretically optimal CP forecast shows the lowest Brier score. However, the BLP forecast is a very close competitor; it is equally well calibrated and nearly as sharp as the CP forecast.

4 Case study: Probability of precipitation forecasts

We turn to a data example on statistical and National Weather Service probability of precipitation forecasts in the continental US. With some one-third of the US economy being weather sensitive, and severe weather causing billions of dollars in damage and hundreds of deaths annually, there is a critical need for calibrated and sharp probabilistic weather forecasts, to allow for optimal decision making under inherent uncertainty (Dutton 2002; Regnier 2008).

Baars and Mass (2005) consider probability of precipitation forecasts for 29 meteorological stations at major urban centers spread across the continental US. They compare the performance of individual and linearly combined model output statistics (MOS) and National Weather Service (NWS) forecasts, and conclude that a linear opinion pool of the machine generated MOS forecasts is competitive or superior to the NWS forecast at nearly all locations. Here we consider the aggregate performance of individual and combined forecasts at all 29 stations, based on the automated **GMOS**, **EMOS** and **NMOS** forecasts, and the human generated, operational **NWS** forecast. The MOS probability forecasts are statistical forecasts that apply logistic regression techniques to the output of a numerical weather prediction model and recent weather observations (Glahn and Lowry 1972; Wilks 2006).

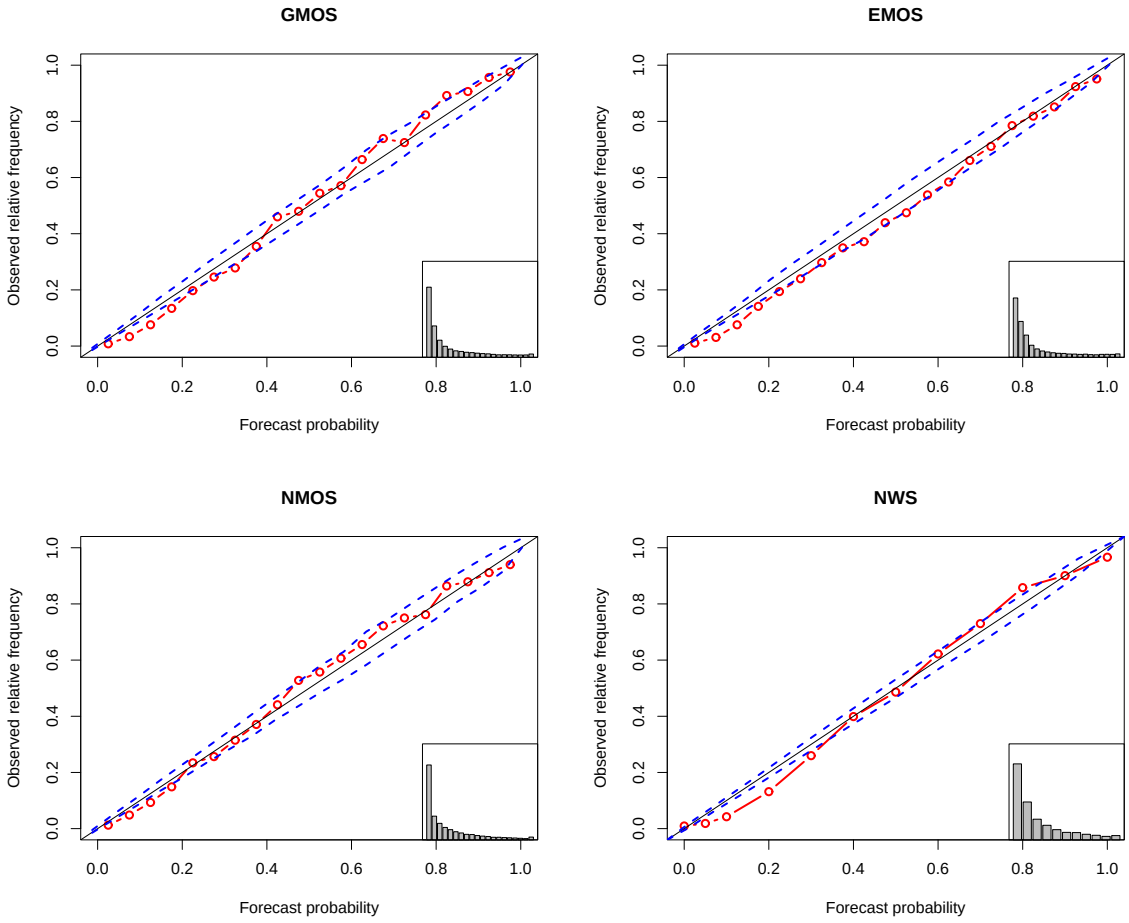

Figure 3: Calibration curves and 95% bootstrap intervals under the null hypothesis of calibration for the four individual probability of precipitation forecasts in the test period. The histograms show the empirical distribution of the forecast values over the unit interval.

The MOS forecasts are recorded in multiples of a hundredth; the NWS forecasts come in multiples of a tenth, except that a forecast probability of 0.05 is issued occasionally.

We consider 2-days ahead probability of precipitation forecasts for the 12 hour term denoted period 2 by Baars and Mass (2005), ranging from July 1, 2003 to March 3, 2008. This includes but is not limited to the one year record studied by Baars and Mass (2005). We use the first two years (July 1, 2003 to June 30, 2005) as training data, on which we fit OLP and BLP models that apply at all stations simultaneously.[7] The balance of the record (July 1, 2005 to March 3, 2008) is used as test data on which we evaluate the forecasts. All results are aggregated over the test period and the 29 stations.

[7]Station-specific BLP fits might well lead to improved predictive performance. A detailed investigation is beyond the scope of the paper and left for future research.

Figure 3 shows calibration curves for the four individual forecasts over the test period. We are in the desirable situation in which the calibration curves show only minor deviations from the diagonal, and so we fit the BLP model (6) under the constraint (8). Hence, the BLP model has a single additional recalibration parameter, $\alpha \geq 1$, when compared to the traditional linear opinion pool.

4.1 Combining statistical forecasts

Following Baars and Mass (2005), we consider combined probability forecasts that use the three statistical probability forecasts, namely the **GMOS**, **EMOS** and **NMOS** forecasts. As previously, the **equally weighted linear opinion pool (ELP)** is obtained as the simple average of the three forecasts. Table 3 shows maximum likelihood (ML) estimates for the **optimally weighted linear opinion pool (OLP)** and the **beta transformed linear opinion pool (BLP)**, which we fit on the training data. For both methods, the GMOS and EMOS weights are about equal and nearly reach $\frac{1}{2}$, with the NMOS weight being much smaller. The ML estimate of the BLP recalibration parameter, α, is 1.48.

Reliability diagrams for the combined forecasts are shown in Figure 4. The calibration curve for the OLP forecast deviates significantly from the diagonal; the effect is stronger than for any of the individual forecasts, and the direction of the departure agrees with our theoretical results, in that the linearly combined forecast is underconfident. The calibration curve for the ELP forecast is very similar and so it is not shown here. The nonlinearly recalibrated BLP forecast is empirically well calibrated and sharper than the OLP forecast.

Table 4 shows the Brier score and its reliability, resolution and uncertainty components for the individual and combined forecasts. The BLP forecast performs the best, both in terms of the Brier score, the reliability or calibration component, and the resolution component. The improvement of the nonlinear BLP method over the linear OLP forecast is about the same as that of the OLP forecast over the best individual forecast, the GMOS forecast.

4.2 Combining statistical and National Weather Service forecasts

We turn to combined probability forecasts that are based on all four individual sources, now including the NWS forecast, in addition to the GMOS, EMOS and NMOS forecasts. This possibility was not explored by Baars and Mass (2005), who aimed to compare the automated MOS forecasts to the subjective, human generated NWS forecast.

Table 5 shows ML estimates for the OLP and BLP models, which we fit on the training data. For both methods, the GMOS and EMOS forecasts receive weights that are nearly equal, at about 0.37. The NWS forecast receives weights of 0.27 and 0.22, respectively; the weights for the NMOS forecast are negligible. The ML estimate of the BLP recalibration parameter, α, is 1.49.

Table 3: Combined probability forecasts in the precipitation example, using the statistical forecasts only. Maximum likelihood estimates for the OLP and BLP parameters from the training period with standard errors in brackets.

Method	GMOS	EMOS	NMOS	α
OLP	0.485 (0.026)	0.465 (0.027)	0.050 (0.020)	
BLP	0.462 (0.022)	0.447 (0.022)	0.091 (0.021)	1.48 (0.03)

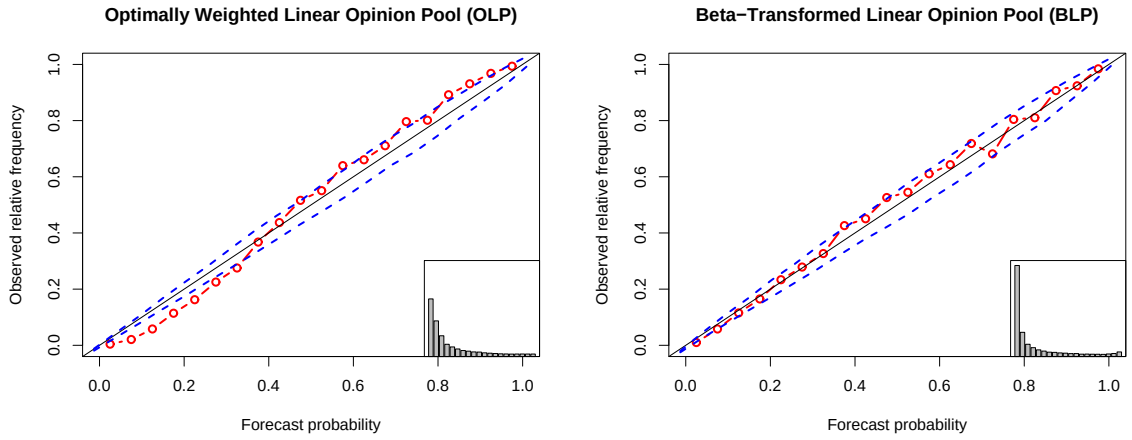

Figure 4: Calibration curves and 95% bootstrap intervals under the null hypothesis of calibration for the OLP and BLP probability of precipitation forecasts in the test period, using the statistical forecasts only. The histograms show the empirical distribution of the forecast values over the unit interval.

Table 4: Mean Brier score (BS) and its reliability (REL), resolution (RES) and uncertainty (UNC) components for individual and combined probability of precipitation forecasts in the test period, using the statistical forecasts only.

Forecast	BS	REL	RES	UNC
GMOS	0.0816	0.0011	0.0735	0.1540
EMOS	0.0866	0.0011	0.0685	0.1540
NMOS	0.0932	0.0005	0.0612	0.1540
ELP	0.0803	0.0022	0.0759	0.1540
OLP	0.0799	0.0021	0.0764	0.1540
BLP	0.0781	0.0004	0.0764	0.1540

Table 5: Same as Table 3 but now using all four individual forecasts, including the NWS forecast.

Method	GMOS	EMOS	NMOS	NWS	α
OLP	0.362 (0.031)	0.368 (0.030)	0.000 (0.026)	0.270 (0.032)	
BLP	0.371 (0.024)	0.377 (0.023)	0.032 (0.022)	0.220 (0.024)	1.49 (0.03)

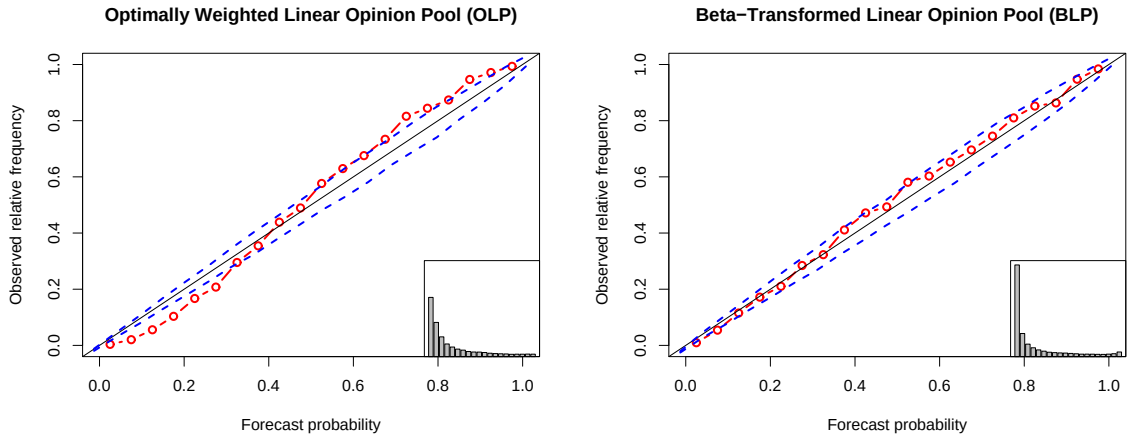

Figure 5: Same as Figure 4 but now using all four individual forecasts, including the NWS forecast.

Table 6: Same as Table 4 but now using all four individual forecasts, including the NWS forecast.

Forecast	BS	REL	RES	UNC
GMOS	0.0816	0.0011	0.0735	0.1540
EMOS	0.0866	0.0011	0.0685	0.1540
NMOS	0.0932	0.0005	0.0612	0.1540
NWS	0.0829	0.0009	0.0721	0.1540
ELP	0.0800	0.0026	0.0770	0.1540
OLP	0.0789	0.0024	0.0778	0.1540
BLP	0.0770	0.0004	0.0777	0.1540

Calibration curves for the OLP and BLP forecasts are shown in Figure 5. We see the now familiar results, in that the linearly combined OLP forecast lacks calibration. The BLP forecast is empirically well calibrated, and considerably sharper than the OLP forecast. The Brier scores in Table 6 echo these results. The BLP forecast outperforms the OLP and ELP forecasts, which perform better than any of the individual forecasts. If we compare to Table 4, we see that the combined probability forecasts benefit from the inclusion of the human generated NWS forecast, with the improvement due to an increase in resolution.

5 Discussion

Our aim in this paper is to provide theoretical and applied guidance in combining probability forecasts from distinct, calibrated or uncalibrated sources. Historically, the linear opinion pool has been the preferred method for doing this. Indeed, there is overwhelming empirical evidence that linearly combined probability forecasts perform better than individual forecasts, and and our results make no exception. That said, our paper demonstrates theoretically and empirically that the linear opinion pool is suboptimal, lacking both calibration and sharpness. To address these shortcomings, we propose the use of the nonlinearly recalibrated, beta transformed linear opinion pool (BLP) that nests the traditional, linearly combined probability forecast.

Theorem 2.1 is our analytic key result; it shows that the linear opinion pool is uncalibrated, even in the desirable case in which the individual probability forecasts are calibrated. This is a finite sample result that does not make any restrictive assumptions about the joint dependence structure of the individual forecasts, and complements the asymptotic results of Wallsten and Diederich (2001) that rely on an assumption of conditional independence. It would be of great interest to bridge the finite sample and asymptotic scenarios, and to establish a more general result, roughly to the extent that linearly combined probability forecasts are uncalibrated and underconfident, resulting in probability statements that are closer to the naive climatological forecast than necessary. A result of this type could perhaps be formulated for a general class of averaging operators and under a minimal assumption of marginal consistency, in lieu of calibration.

Empirically, the shortcomings of the linear opinion pool have been well documented in an interdisciplinary strand of literature that includes the works of Clemen and Winkler (1987), Winkler and Poses (1993), Vislocky and Fritsch (1995), Ariely et al. (2000), Wallsten and Diederich (2001) and Johnson et al. (2001). Despite their ubiquity, these issues have frequently been overlooked, with some of our own work (Sloughter et al. 2007) being one such example.[8]

With a view toward applied forecasting problems, we recommend a transition from the traditional linear opinion pool to the nonlinearly recalibrated, beta-transformed linear opinion

[8]Figure 7 of Sloughter et al. (2007) shows the typical S-shaped calibration curve for a linearly combined probability forecast, even though the effect is comparably small.

pool (BLP). The BLP model (6) has at most two, and typically only one, additional parameters when compared to the linear opinion pool, and it is easy to fit, using the maximum likelihood method or related optimum score techniques. More general and more complex, parametric or nonparametric approaches to the aggregation of probability forecasts, can easily be envisioned, including but not limited to copula models (Nelsen 2006), and might provide effective approximations to the hypothetical, ideally combined forecast, namely the conditional probability (CP) forecast (4). However, more complex statistical models bear the danger of overfitting, and the resulting gains in predictive performance, if any, are likely to be incremental.

Appendix: Mathematical details

Proof of Theorem 2.1

From the basic properties of Bernoulli random variables and conditional expectations,

$$\mathbb{P}(Y=1) = \mathbb{E}Y^2 = \mathbb{E}Y = \mathbb{E}\mathbb{E}[Y|p] = \mathbb{E}p = \mathbb{E}q,$$

which will be used repeatedly in what follows. We first prove part (a). For a contradiction, suppose that p is calibrated, that is, $p = q$ almost surely. Then we can condition on p to see that

$$\mathbb{E}(Y-p)^2 = \mathbb{E}[p(1-p)]. \tag{10}$$

We proceed to show that under the conditions of the theorem equality in (10) is violated. Toward this end, note that

$$\begin{aligned}
\mathbb{E}\,(Y-p)^2 &= \mathbb{E}\left(\sum_{i=1}^k w_i(Y-p_i)\right)^2 \\
&= \sum_{i=1}^k\sum_{j=1}^k w_i w_j \mathbb{E}[(Y-p_i)(Y-p_j)] \\
&= \sum_{i=1}^k\sum_{j=1}^k w_i w_j \mathbb{E}[Y - p_iY - p_jY + p_ip_j] \\
&= \sum_{i=1}^k\sum_{j=1}^k w_i w_j \mathbb{E}[\mathbb{E}(Y|p_i) - \mathbb{E}(p_iY|p_i) - \mathbb{E}(p_jY|p_j) + p_ip_j] \\
&= \sum_{i=1}^k\sum_{j=1}^k w_i w_j \mathbb{E}[p_i - p_i^2 - p_j^2 + p_ip_j] \\
&= \sum_{i=1}^k\sum_{j=1}^k w_i w_j \mathbb{E}[p_i(1-p_j)] - \sum_{i=1}^k\sum_{j=1}^k w_i w_j \mathbb{E}(p_i-p_j)^2
\end{aligned}$$

and

$$\mathbb{E}\,[p(1-p)] = \sum_{i=1}^k\sum_{j=1}^k w_i w_j\, \mathbb{E}\,[p_i(1-p_j)],$$

so that

$$\mathbb{E}(Y-p)^2 = \mathbb{E}[p(1-p)] - \sum_{i=1}^k\sum_{j=1}^k w_i w_j \mathbb{E}(p_i-p_j)^2. \tag{11}$$

The double sum on the right-hand side of (11) is strictly positive, whence (10) is violated, for the desired contradiction.

We now prove part (b). From (11) we see that $\mathbb{E}(Y-p)^2 < \mathbb{E}\,[\,p(1-p)]$. A straightforward conditioning argument shows that

$$\mathbb{E}(Y-p)^2 = \mathbb{E}(Y-q)^2 + \mathbb{E}(q-p)^2 > \mathbb{E}[q(1-q)].$$

Hence,

$$\mathbb{E}[q(1-q)] < \mathbb{E}(Y-p)^2 < \mathbb{E}[p(1-p)],$$

which implies that $\mathbb{E}p^2 < \mathbb{E}q^2$. From this, part (b) follows.

As for part (c),

$$\begin{aligned}
\mathbb{E}\mathrm{S}(q,Y) &= \mathbb{E}\mathbb{E}[\mathrm{S}(q,Y)|p] \\
&= \mathbb{E}[q\mathrm{S}(q,1) + (1-q)\mathrm{S}(q,0)] \\
&< \mathbb{E}[q\mathrm{S}(p,1) + (1-q)\mathrm{S}(p,0)] \\
&= \mathbb{E}\mathbb{E}[\mathrm{S}(p,Y)|p] \\
&= \mathbb{E}\mathrm{S}(p,Y)
\end{aligned}$$

with the inequality being strict, because S is a negatively oriented strictly proper scoring rule and $q = \mathbb{E}[Y|p] \neq p$ with positive probability. □

Details for (2) and (3)

The final equality in (2) stems from the fact that if $X \sim \mathcal{N}(\mu, \sigma^2)$ then

$$\begin{aligned}
\mathbb{E}\Phi(X) &= \int_{-\infty}^{\infty} \Phi(x)\,\frac{1}{\sigma}\,\phi\!\left(\frac{x-\mu}{\sigma}\right)\mathrm{d}x \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{x}\phi(y)\,\mathrm{d}y\right)\frac{1}{\sigma}\,\phi\!\left(\frac{x-\mu}{\sigma}\right)\mathrm{d}x \\
&= \int_{y\le x}\phi(y)\,\frac{1}{\sigma}\,\phi\!\left(\frac{x-\mu}{\sigma}\right)\mathrm{d}x\,\mathrm{d}y \\
&= \mathbb{P}(Y \le X) \;=\; \mathbb{P}(Y - X \le 0) \;=\; \Phi\!\left(\frac{\mu}{\sqrt{\sigma^2+1}}\right),
\end{aligned}$$

where ϕ denotes the standard normal density function and Y is standard normal and independent of X, so that $Y - X \sim \mathcal{N}(-\mu, \sigma^2+1)$. The conditional distribution of $a_1 + a_2$ given a_1 is normal with mean a_1 and variance 2, whence

$$\mathbb{E}\,[\Phi(a_1+a_2)|a_1] = \Phi\!\left(\frac{a_1}{\sqrt{3}}\right).$$

An almost identical calculation applies to (3).

Acknowledgements

This research was supported by the National Science Foundation under Awards ATM-0724721 and DMS-0706745, and by the Joint Ensemble Forecasting System (JEFS) under subcontract S06-47225 from the University Corporation for Atmospheric Research (UCAR). Jeff Baars and Cliff Mass (University of Washington Department of Atmospheric Sciences) kindly provided the precipitation forecasts and verification data. We are grateful to Matt Pocernich (National Center for Atmospheric Research, Boulder, Colorado) for providing R code, and to Theo Eicher, Don Percival, Adrian Raftery and seminar attendees at the University of Washington for comments.

References

Al-Najjar, N. I. and Weinstein, J. (2008): "Comparative Testing of Experts," *Econometrica*, 76, 541–559.

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H. B., Wallsten, T. S. and Zauberman, G. (2000): "The Effects of Averaging Subjective Probability Estimates Between and Within Judges," *Journal of Experimental Psychology: Applied*, 6, 130–147.

Baars, J. A. and Mass, C. F. (2005): "Performance of National Weather Service Forecasts Compared to Operational, Consensus, and Weighted Model Output Statistics," *Weather and Forecasting*, 20, 1034–1047.

Brier, G. W. (1950): "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78, 1–3.

Bröcker, J. and Smith, L. A. (2007): "Increasing the Reliability of Reliability Diagrams," *Weather and Forecasting*, 22, 651–661.

Clemen, R. T. and Winkler, R. L. (1987): "Calibrating and Combining Precipitation Probability Forecasts," in Viertl, R. (ed.), *Probability and Bayesian Statistics*, Plenum, New York, pp. 97–110.

——— (1999): "Combining Probability Distributions From Experts in Risk Analysis," *Risk Analysis*, 19, 187–203.

——— (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.

Croushore, D. (1993): "Introducing: The Survey of Professional Forecasters," *Federal Reserve Bank of Philadelphia Business Review*, November/December 3–13.

Dawid, A. P. (1982): "The Well-Calibrated Bayesian," *Journal of the American Statistical Association*, 77, 605–610.

——— (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.

DeGroot, M. H. and Fienberg, S. E. (1982): "Assessing Probability Assessors: Calibration and Refinement," in Gupta, S. S. and Berger, J. O. (eds.), *Statistical Decision Theory and Related Topics III*, Vol. 1, Academic Press, New York, pp. 291–314.

——— (1983): "The Comparison and Evaluation of Forecasters," *Statistician*, 32, 12–22.

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.

Elliott, G. and Timmermann, A. (2008): "Economic Forecasting," *Journal of Economic Literature*, 46, 3–56.

Feinberg, Y. and Stewart, C. (2008): "Testing Multiple Forecasters," *Econometrica*, 76, 561–582.

Foster, D. P. and Vohra, R. V. (1998): "Asymptotic Calibration," *Biometrika*, 85, 379–390.

Genest, C. and Schervish, M. J. (1985): "Modeling Expert Judgements for Bayesian Updating," *Annals of Statistics*, 13, 1198–1212.

Genest, C. and Zidek, J. (1986): "Combining Probability Distributions: A Critique and an Annotated Bibliography," *Statistical Science*, 1, 114–135.

Glahn, H. R. and Lowry, D. A. (1972): "The Use of Model Output Statistics (MOS) in Objective Weather Forecasting," *Journal of Applied Meteorology*, 11, 1203–1211.

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

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

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007): "Probabilistic Forecasts, Calibration and Sharpness," *Journal of the Royal Statistical Society Series B*, 69, 243–268.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. and Johnson, N. A. (2008): "Rejoinder on: Assessing Probabilistic Forecasts of Multivariate Quantities, With an Application to Ensemble Predictions of Surface Winds", *Test*, 17, 256–264.

Good, I. J. (1952): "Rational Decisions," *Journal of the Royal Statistical Society Series B*, 14, 107–114.

Graham, J. R. (1996): "Is a Group of Economists Better Than One? Than None?," *Journal of Business*, 69, 193–232.

Hoeting, J. A., Madigan, D. M., Raftery, A. E. and Volinsky, C. T. (1999): "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–401.

Hora, S. C. (2004): "Probability Judgements for Continuous Quantities: Linear Combinations and Calibration," *Management Science*, 50, 597–604.

Johnson, T. R., Budescu, D. V. and Wallsten, T. S. (2001): "Averaging Proba-

bility Judgements: Monte Carlo Analyses of Asymptotic Diagnostic Value," *Journal of Behavioral Decision Making*, 14, 123–140.

Kynn, M. (2008): "The 'Heuristics and Biases' Bias in Expert Elicitation," *Journal of the Royal Statistical Society Series A*, 171, 239–264.

Lehrer, E. (2001): "Every Inspection is Manipulaple," *Econometrica*, 69, 1333–1347.

Lichtenstein, S., Fischhoff, B. and Phillips, L. D. (1982): "Calibration of Probabilities: The State of the Art to 1980," in Kahnemann, D., Slovic, P. and Tversky, A. (eds.), *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge University Press, pp. 306–334.

Lindley, D. V. (1982): "The Improvement of Probability Judgements," *Journal of the Royal Statistical Society Series A*, 145, 117–126.

Murphy, A. H. (1973): "A New Vector Partition of the Probability Score," *Journal of Applied Meteorology*, 12, 595–600.

——— (1998): "The Early History of Probability Forecasts: Some Extensions and Clarifications," *Weather and Forecasting*, 13, 5–15.

Murphy, A. H., and Winkler, R. L. (1987): "A General Framework for Forecast Verification," *Monthly Weather Review*, 115, 1330–1338.

——— (1992): "Diagnostic Verification of Probability Forecasts," *International Journal of Forecasting*, 7, 435–455.

Nelsen, R. B. (2006): *An Introduction to Copulas*, 2nd edition. Springer, New York.

Pepe, M. S. (2003): *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Pocernich, M. (2008): Verification: Forecast Verification Utilities. R Package Version 1.26.

Primo, C., Ferro, C. A. T., Jolliffe, I. T. and Stephenson, D. B. (2008): "Combination and Calibration Methods for Probabilistic Forecasts of Binary Events," Working Paper.

Regnier, E. (2008): "Doing Something About the Weather," *Omega*, 36, 22–32.

Sanders, F. (1963): "On Subjective Probability Forecasting," *Journal of Applied Meteorology*, 2, 191–201.

Sandroni, A., Smorodinsky, R. and Vohra, R. V. (2003): "Calibration With Many Checking Rules," *Mathematics of Operations Research*, 28, 141–153.

Schervish, M. J. (1989): "A General Method for Comparing Probability Assessors," *Annals of Statistics*, 17, 1856–1879.

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

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.

Stephenson, D. B., Coelho, C. and Jolliffe, I. T. (2008): "Two Extra Components in the Brier Score Decomposition," *Weather and Forecasting*, 23, 752–757.

Tetlock, P. E. (2005): *Expert Political Judgement. How Good is It? How Can we Know?* Princeton University Press.

Vislocky, R. L. and Fritsch, J. M. (1995): "Improved Model Output Statistics Forecasts Through Model Consensus," *Bulletin of the American Meteorological Society*, 76, 1157–1164.

Vovk, V. and Shafer, G. (2005): "Good Randomized Sequential Probability Forecasting is Always Possible," *Journal of the Royal Statistical Society Series B*, 67, 747–763.

Wallsten, T. S. and Diederich, A. (2001): "Understanding Pooled Subjective Probability Estimates," *Mathematical Social Sciences*, 18, 1–18.

Wallsten, T. S., Budescu, D. V., Erev, I. and Diederich, A. (1997): "Evaluating and Combining Subjective Probability Estimates," *Journal of Behavioral Decision Making*, 10, 243–268.

Wilks, D. S. (2006): *Statistical Methods in the Atmospheric Sciences*, 2nd edition. Academic Press.

Wilson, P. W. F., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H. and Kannel, W. B. (1998): "Prediction of Coronary Heart Disease Using Risk Factor Categories", *Circulation*, 97, 1837–1847.

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

Winkler, R. L. and Jose, V. R. R. (2008): "Comments on: Assessing Probabilistic Forecasts of Multivariate Quantities, With an Application to Ensemble Predictions of Surface Winds", *Test*, 17, 251–255.

Winkler, R. L. and Poses, R. M. (1993): "Evaluating and Combining Physicians' Probabilities of Survival in an Intensive Care Unit", *Management Science*, 39, 1526–1543.