

EVALUATING PROBABILITY FORECASTS

BY TZE LEUNG LAI*, DAVID BO SHEN AND SHULAMITH GROSS†

Stanford University and Baruch College/CUNY

Probability forecasts of events are routinely used in climate predictions, in forecasting default probabilities on bank loans, or in estimating the probability of a patient's positive response to treatment. Scoring rules have long been used to assess the efficacy of the forecast probabilities after observing the occurrence, or non-occurrence, of the predicted events. We develop herein a statistical theory for scoring rules and propose an alternative approach to the evaluation of probability forecasts. This approach uses loss functions relating the predicted to the actual probabilities of the events, and applies martingale theory to exploit the temporal structure between the forecast and the subsequent occurrence or non-occurrence of the event.

1. Introduction. Probability forecasts of future events are widely used in diverse fields of application. Oncologists routinely predict the probability of a cancer patient's progression-free survival beyond a certain time horizon (Hari et al., 2009). Economists give the probability forecasts of an economic rebound or a recession by the end of a fiscal year. Banks are required by regulators assessing their capital requirements to predict periodically the risk of default of the loans they make. Engineers are routinely called upon to predict the survival probability of a system or infrastructure beyond five or ten years; this includes bridges, sewer systems and other structures. Finally, lawyers also assess the probability of particular trial outcome (Fox and Birke, 2002) in order to determine whether to go to trial or settle out of court. This list would not be complete without mentioning the field that is most advanced in its daily probability predictions, namely meteorology. In the past 60 years, remarkable advances in forecasting precipitation probabilities, temperatures, and rainfall amounts have been made in terms of breadth and accuracy. Murphy and Winkler (1984) provide an illuminating history of the US National Weather Service's transition from non-probabilistic to probability predictions and its development of reliability and accuracy measures for these probability forecasts. Accuracy assessment is difficult to carry

*Work supported in part by NSF grant DMS 0805879.

†Work supported in part by PSC-CUNY 2008 and 2009 grants and a 2008 Summer Research Support grant from Baruch Zicklin School of Business.

Keywords and phrases: forecasting, loss functions, martingales, scoring rules

out directly because it requires comparing a forecaster’s predicted probabilities with the actual but unknown probabilities of the events under study. Reliability is measured using “scoring rules”, which are empirical distance measures between repeated predicted probabilities of an event, such as having no rain the next day, and indicator variables that take on the value 1 if the predicted event actually occurs, and 0 otherwise; see Gneiting and Raftery (2007) and Gneiting, Balabdaoui and Raftery (2007) for recent reviews.

To be more specific, a scoring rule for a sequence of n probability forecasts \hat{p}_i , $i = 1, \dots, n$, is the average score $n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i)$, where $Y_i = 1$ or 0 according to whether the i th event actually occurs or not. An example is the widely used Brier’s score $L(y, \hat{p}) = (y - \hat{p})^2$ (Brier, 1950). Noting that the Y_i are related to the actual but unknown probability p_i via $Y_i \sim \text{Bernoulli}(p_i)$, Cox (1958) proposed to evaluate how well the \hat{p}_i predict p_i by using the estimates of (β_1, β_2) in the regression model

$$\text{logit}(p_i) = \beta_1 + \beta_2 \text{logit}(\hat{p}_i) \quad (1.1)$$

and developed a test of the null hypothesis $(\beta_1, \beta_2) = (0, 1)$, which corresponds to perfect prediction. Spiegelhalter (1986) subsequently proposed a test of the null hypothesis $H_0 : \hat{p}_i = p_i$ for all $i = 1, \dots, n$, based on a standardized form (under H_0) of Brier’s score. Redelmeier, Bloch and Hickam (1991) extended Spiegelhalter’s idea to develop a test of equality of the predictive probabilities of two forecasters. A serious limitation of the hypothesis testing approach is the unrealistic benchmark of perfect prediction to formulate the null hypothesis, so significant departures from it are expected when n is large and they convey little information on how well the \hat{p}_i predict p_i . Another limitation is the implicit assumption that the \hat{p}_i are independent random variables, which clearly is violated since \hat{p}_i usually involves previous observations and predictions.

In this paper we develop a new statistical methodology for evaluating probability forecasts via the average loss $L_n = n^{-1} \sum_{i=1}^n L(p_i, \hat{p}_i)$. When L is linear in p_i , $L(Y_i, \hat{p}_i)$ is an unbiased estimate of $L(p_i, \hat{p}_i)$ since $E(Y_i | \hat{p}_i) = p_i$. We show in Section 2, where an overview of loss functions and scoring rules is also given, that even for L that is nonlinear in p_i there is a “linear equivalent” which carries the same information as L for comparing different forecasts. In Section 3 we make use of this insight to construct inferential procedures, such as confidence intervals, for the average loss L_n under certain assumptions and for comparing the average losses of different forecasters. Section 4 gives a simulation study of the performance of the proposed methodology, and Section 5 provides an illustrative application to the evaluation of some MOS

(model output statistics) probability forecasts of the US National Weather Service. Some concluding remarks and discussion are given in Section 7.

2. Scoring Rules and Associated Loss Functions. Instead of defining a scoring rule via L (which associates better forecasts with smaller values of L), Gneiting and Raftery (2007) and others assign higher scores to better forecasts; this is tantamount to using $-L$ instead of L in defining a scoring rule. More generally, considering p and its forecast \hat{p} as probability measures, they call a scoring rule S *proper* relative to a class \mathcal{P} of probability measures if $E_p S(Z, p) \geq E_p S(Z, \hat{p})$ for all p and \hat{p} belonging to \mathcal{P} , where Z is an observed random vector (generated from p) on which scoring is based. For the development in the subsequent sections, we find it more convenient to work with L instead of $-L$ and restrict to $Z = (Y_1, \dots, Y_n)$ so that $S(Z, (\hat{p}_1, \dots, \hat{p}_n)) = -n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i)$.

The function L in the scoring rule $n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i)$ measures the closeness of the probability forecast \hat{p}_i of event i before the indicator variable Y_i of the event is observed. We can also use L as a loss function in measuring the accuracy of \hat{p}_i as an estimate of the probability p_i of event i . Besides the squared error loss $L(p, \hat{p}) = (p - \hat{p})^2$ used in Brier's score, another widely used loss function is the Kullback–Leibler divergence

$$L(p, \hat{p}) = p \log(p/\hat{p}) + (1-p) \log[(1-p)/(1-\hat{p})], \quad (2.1)$$

which is closely related to the log score introduced by Good (1952), as shown below.

We call a loss function $\tilde{L}(p, \hat{p})$ a *linear equivalent* of the loss function $L(p, \hat{p})$ if $\tilde{L}(p, \hat{p})$ is a linear function of p and

$$L(p, \hat{p}) - \tilde{L}(p, \hat{p}) \text{ does not depend on } \hat{p}. \quad (2.2)$$

For example, $\tilde{L}(p, \hat{p}) = -2p\hat{p} + \hat{p}^2$ is a linear equivalent of the squared error loss $(p - \hat{p})^2$ used by Brier's score. A linear equivalent \tilde{L} of the Kullback–Leibler divergence (2.1) is given by $-\tilde{L}(p, \hat{p}) = p \log(\hat{p}) + (1-p) \log(1-\hat{p})$. This is the conditional expected value (given \hat{p}) of $Y \log(\hat{p}) + (1-Y) \log(1-\hat{p})$, which is Good's log score. Since the probability \hat{p}_i is determined before the Bernoulli random variable Y_i is observed,

$$E \{L(Y_i, \hat{p}_i) | \hat{p}_i\} = p_i L(1, \hat{p}_i) + (1-p_i) L(0, \hat{p}_i). \quad (2.3)$$

Therefore the conditional expected loss of a scoring rule $L(Y, \hat{p})$ yields a loss function

$$\tilde{L}(p, \hat{p}) = \{L(1, \hat{p}) - L(0, \hat{p})\} p + L(0, \hat{p}) \quad (2.4)$$

that is linear in p . For example, the absolute value scoring rule $L(Y, \hat{p}) = |Y - \hat{p}|$ is associated with $\tilde{L}(p, \hat{p}) = p(1 - \hat{p}) + (1 - p)\hat{p}$ that is linear in each argument. Using the notation (2.4), the scoring rule $L(Y, \hat{p})$ is proper if $\tilde{L}(p, p) \leq \tilde{L}(p, \hat{p})$ for all $p, \hat{p} \in [0, 1]$, and is strictly proper if $\min_{0 \leq \hat{p} \leq 1} \tilde{L}(p, \hat{p})$ is uniquely attained at $p = \hat{p}$. The scoring rule $|Y - \hat{p}|$, therefore, is not proper; moreover, $|p - \hat{p}|$ does not have a linear equivalent.

3. A New Approach to Evaluation of Probability Forecasts. In this section we first consider the evaluation of a sequence of probability forecasts $\hat{p}_1, \dots, \hat{p}_n$ based on the corresponding sequence of indicator variables Y_1, \dots, Y_n that denote whether the events actually occur or not. Whereas the traditional approach to evaluating $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_n)$ uses the scoring rule $n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i)$, we propose to evaluate $\hat{\mathbf{p}}$ via

$$L_n = n^{-1} \sum_{i=1}^n L(p_i, \hat{p}_i), \quad (3.1)$$

where L is a loss function and p_i is the actual probability of the occurrence of the i th event. Allowing the forecast \hat{p}_k to depend on an information set \mathcal{F}_{k-1} that consists of the event and forecast histories and other covariates before Y_k is observed, the conditional distribution of Y_i given \mathcal{F}_{i-1} is Bernoulli(p_i), and therefore

$$P(Y_i = 1 | \mathcal{F}_{i-1}) = p_i. \quad (3.2)$$

In view of (3.2), an obvious estimate of the unknown p_i is Y_i . Even though it is based on a sample of size 1, averaging these estimates in $n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i)$ provides a consistent and asymptotically normal estimate of (3.1) when $L(p, \hat{p})$ is linear in p . This case is considered in Section 3.1. Extensions to more general loss functions are given in Section 3.2, which considers the problem of evaluating the comparative performance of different forecasts, and in Section 3.3 that assumes some additional structure on how the unknown p_i are generated.

3.1. Linear case. Suppose $L(p, \hat{p})$ is linear in p , as in the case of linear equivalents of general loss functions. Combining this linearity property with (3.2) yields

$$E \{L(Y_i, \hat{p}_i) | \mathcal{F}_{i-1}\} = L(p_i, \hat{p}_i), \quad (3.3)$$

and therefore $L(Y_i, \hat{p}_i) - L(p_i, \hat{p}_i)$ is a martingale difference sequence with respect to $\{\mathcal{F}_i\}$. In the Appendix we apply martingale theory to prove the following.

THEOREM 1. *Suppose $L(p, \hat{p})$ is linear in p . Let $\hat{L}_n = n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i)$ and define L_n by (3.1). Letting*

$$\sigma_n^2 = n^{-1} \sum_{i=1}^n \{L(1, \hat{p}_i) - L(0, \hat{p}_i)\}^2 p_i(1 - p_i), \quad (3.4)$$

assume that $\sigma_n^2 = O(1)$ with probability 1. Then $\hat{L}_n - L_n$ converges to 0 with probability 1. If σ_n^2 converges in probability to some non-random positive constant, then $\sqrt{n}(\hat{L}_n - L_n)/\sigma_n$ has a limiting standard normal distribution.

To apply Theorem 1 to construct confidence intervals for L_n , one needs to estimate the unknown p_i in (3.4). Whereas substituting p_i by Y_i in $L(p_i, \hat{p}_i)$ leads to a consistent estimate of L_n when L is linear, such substitution gives 0 as an overly optimistic estimate of $p_i(1 - p_i) = \text{var}(Y_i | \mathcal{F}_{i-1})$. A conservative confidence interval for L_n can be obtained by replacing $p_i(1 - p_i)$ in (3.4) by its upper bound 1/4. In Section 3.3, we consider estimation of σ_n^2 and of $n^{-1} \sum_{i=1}^n L(p_i, \hat{p}_i)$ when L is nonlinear in p_i , under additional assumptions on how the p_i are generated.

3.2. Application to comparison of probability forecasts. Consider two sequences of probability forecasts $\hat{\boldsymbol{p}}' = (\hat{p}'_1, \dots, \hat{p}'_n)$ and $\hat{\boldsymbol{p}}'' = (\hat{p}''_1, \dots, \hat{p}''_n)$ of $\boldsymbol{p} = (p_1, \dots, p_n)$. Suppose a loss function $L(p, q)$ is used to evaluate each forecast, and let $\tilde{L}(p, q)$ be its linear equivalent. Since $L(p, q) - \tilde{L}(p, q)$ does not depend on q in view of (2.2), it is a function only of p , which we denote by $d(p)$. Hence

$$\begin{aligned} L(p_i, \hat{p}'_i) - L(p_i, \hat{p}''_i) &= \{\tilde{L}(p_i, \hat{p}'_i) + d(p_i)\} - \{\tilde{L}(p_i, \hat{p}''_i) + d(p_i)\} \\ &= \tilde{L}(p_i, \hat{p}'_i) - \tilde{L}(p_i, \hat{p}''_i) \end{aligned}$$

is a linear function of p_i , and therefore we can estimate $\Delta_n = n^{-1} \sum_{i=1}^n \{L(p_i, \hat{p}'_i) - L(p_i, \hat{p}''_i)\}$ by the difference $n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}'_i) - n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}''_i)$ of scores of the two forecasts. Application of Theorem 1 then yields the following theorem, whose part (ii) is related to (2.4).

THEOREM 2. *Let $\hat{\Delta}_n = n^{-1} \sum_{i=1}^n \{L(Y_i, \hat{p}'_i) - L(Y_i, \hat{p}''_i)\}$ and*

$$\begin{aligned} \delta_i &= \{L(1, \hat{p}'_i) - L(0, \hat{p}'_i)\} - \{L(1, \hat{p}''_i) - L(0, \hat{p}''_i)\}, \\ s_n^2 &= n^{-1} \sum_{i=1}^n \delta_i^2 p_i(1 - p_i). \end{aligned} \quad (3.5)$$

- (i) Suppose L has a linear equivalent. Letting $\Delta_n = n^{-1} \sum_{i=1}^n \{L(p_i, \hat{p}'_i) - L(p_i, \hat{p}''_i)\}$, assume that $s_n^2 = O(1)$ with probability 1. Then $\hat{\Delta}_n - \Delta_n$ converges to 0 with probability 1. If furthermore s_n converges in probability to some non-random positive constant, then $\sqrt{n}(\hat{\Delta}_n - \Delta_n)/s_n$ has a limiting standard normal distribution.
- (ii) Without assuming that L has a linear equivalent, the same conclusion as in (i) still holds with $\Delta_n = n^{-1} \sum_{i=1}^n \{\delta_i p_i + L(0, \hat{p}'_i) - L(0, \hat{p}''_i)\}$.

3.3. *Risk buckets and quadratic loss functions.* Both (3.4) and (3.5) involve $p_i(1 - p_i)$, which is the variance of the Bernoulli random variable Y_i . It is not possible to estimate this variance based on a single observation unless there is some statistical structure on the p_i to make (3.4) or (3.5) estimable, and a conservative approach in the absence of such structure is to use the upper bound 1/4 for $p_i(1 - p_i)$ in (3.4) or (3.5), as noted in Section 3.1. One such structure is that the p_i can be grouped into buckets within which they have the same value, as in risk assessment of a bank's retail loans (e.g., mortgages, automobile loans and personal loans), for which the obligors are grouped into risk buckets within which they can be regarded as having the same risk (or more precisely, the same probability of default on their loans). According to the Basel Committee on Banking Supervision (2006, p. 91), each bank has to use at least seven risk buckets for borrowers who have not defaulted and at least one for those who have defaulted previously at the time of loan application.

A bucket model for risk assessment involves multivariate forecasts for subjects/locations k at a given time t . Thus, identifying the index i with (t, k) , one has a vector of probability forecasts $(\hat{p}_{t,1}, \dots, \hat{p}_{t,K})$ at time $t - 1$ for the occurrences of K events at time t . The information set \mathcal{F}_{i-1} can then be expressed as \mathcal{G}_{t-1} that consists of event and forecast histories up to time $t - 1$, and therefore conditional on \mathcal{G}_{t-1} , the events at time t can be regarded as the outcomes of K independent Bernoulli trials with respective probabilities $p_{t,1}, \dots, p_{t,K}$. The bucket model assumes that, conditional on \mathcal{G}_{t-1} , events in the same bucket at time t have the same probability of occurrence. Let J_t be the number of buckets at time t and $n_{j,t}$ be the size of the j th bucket, $1 \leq j \leq J_t$, so that $n = \sum_{t=1}^T \sum_{j=1}^{J_t} n_{j,t}$. Then the common p_i of the j th bucket at time t , denoted by $p_t(j)$, can be estimated by the relative frequency $\bar{Y}_t(j) = n_{j,t}^{-1} \sum_{i \in I_{j,t}} Y_i$, where $I_{j,t}$ denotes the index set for the bucket. This in turn yields an unbiased estimate

$$\hat{v}_t(j) = n_{j,t} \bar{Y}_t(j) (1 - \bar{Y}_t(j)) / (n_{j,t} - 1) \quad (3.6)$$

of $p_i(1 - p_i)$ for $i \in I_{j,t}$, and we can replace $p_i(1 - p_i)$ in (3.4) or (3.5) by

$\hat{v}_t(j)$ for $i \in I_{j,t}$ so that the results of Theorem 1 or Theorem 2 still hold with these estimates of the asymptotic variance, as shown in the following.

LEMMA 1. *Using the same notation as in the preceding paragraph, suppose $n_{j,t} \geq 2$ for $1 \leq j \leq J_t$ and define $\hat{v}_t(j)$ by (3.6).*

(i) *Under the same assumptions as in Theorem 1, define*

$$\hat{\sigma}_n^2 = n^{-1} \sum_{t=1}^T \sum_{j=1}^{J_t} \sum_{i \in I_{j,t}} \{L(1, \hat{p}_i) - L(0, \hat{p}_i)\}^2 \hat{v}_t(j).$$

Then $\hat{\sigma}_n^2 - \sigma_n^2$ converges to 0 with probability 1.

(ii) *Under the same assumptions as in Theorem 2, $\hat{s}_n^2 - s_n^2$ converges to 0 with probability 1, where $\hat{s}_n^2 = n^{-1} \sum_{t=1}^T \sum_{j=1}^{J_t} \sum_{i \in I_{j,t}} \delta_i^2 \hat{v}_t(j)$.*

The proof of Lemma 1, which is given in the Appendix, also shows that for the squared error loss $L(p, \hat{p}) = (p - \hat{p})^2$, we can estimate (3.1) in the bucket model by the *adjusted Brier score*

$$\hat{L}_n - n^{-1} \sum_{t=1}^T \sum_{j=1}^{J_t} n_{j,t} \hat{v}_t(j), \quad (3.7)$$

since $\hat{L}_n = n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i)$ is a consistent estimate of the linear equivalent $n^{-1} \sum_{i=1}^n (\hat{p}_i^2 - 2p_i \hat{p}_i + p_i)$ and $n^{-1} \sum_{t=1}^T \sum_{j=1}^{J_t} n_{j,t} \hat{v}_t(j)$ is a consistent estimate of $n^{-1} \sum_{i=1}^n p_i(1 - p_i)$. Consistency of an estimate \hat{l}_n of l_n means that $\hat{l}_n - l_n$ converges to 0 in probability as $n \rightarrow \infty$. Moreover, the following lemma shows that $\sqrt{n}(\hat{L}_n - n^{-1} \sum_{t=1}^T \sum_{j=1}^{J_t} n_{j,t} \hat{v}_t(j) - L_n)$ has a limiting normal distribution in the bucket model and can be studentized to give a limiting standard normal distribution. Its proof is given in the Appendix.

LEMMA 2. *Suppose $n_{j,t} \geq 2$ for $1 \leq j \leq J_t$. Letting $L(p, \hat{p}) = (p - \hat{p})^2$, define L_n by (3.1) and the adjusted Brier score by (3.7). Let $v_t(j) = p_t(j)(1 - p_t(j))$,*

$$\beta_n^2 = n^{-1} \sum_{t=1}^T \sum_{j=1}^{J_t} \left\{ v_t(j) \sum_{i \in I_{j,t}} (1 - 2\hat{p}_i)^2 - 2v_t(j) (1 - 2p_t(j)) \sum_{i \in I_{j,t}} (1 - 2\hat{p}_i) + n_{j,t} v_t(j) (1 - 4v_t(j)) + 2n_{j,t} v_t^2(j) / (n_{j,t} - 1) \right\}. \quad (3.8)$$

If β_n converges in probability to some non-random positive constant, then $\sqrt{n}(\hat{L}_n - n^{-1} \sum_{t=1}^T \sum_{j=1}^{J_t} n_{j,t} \hat{v}_t(j) - L_n) / \beta_n$ has a limiting standard normal

distribution. Moreover, if $n_j \geq 3$ for all $1 \leq j \leq J_t$, then $\hat{\beta}_n - \beta_n$ converges to 0 with probability 1, where

$$\begin{aligned} \hat{\beta}_n^2 = & \frac{1}{n} \sum_{t=1}^T \sum_{j=1}^{J_t} \left\{ \hat{v}_t(j) \sum_{i \in I_{j,t}} (1 - 2\hat{p}_i)^2 \right. \\ & - \frac{2n_{j,t}^2}{(n_{j,t} - 1)^3} \left[\sum_{i \in I_{j,t}} (1 - 2\hat{p}_i) \right] \left[\sum_{i \in I_{j,t}} (Y_i - \bar{Y}_t(j))^3 \right] \\ & \left. + \frac{4n_{j,t}(n_{j,t} - 1)}{(n_{j,t} - 2)^2} \sum_{i \in I_{j,t}} \left[\frac{1}{2(n_{j,t} - 1)} \sum_{k \in I_{j,t}, k \neq i} (Y_i - Y_k)^2 - \hat{v}_t(j) \right]^2 \right\}. \end{aligned} \quad (3.9)$$

4. Simulation Studies. The risk buckets in Section 3.3 and the forecasts are usually based on covariates. In this section we consider $T = 2$ in the case of discrete covariates so that there are J_t buckets of various sizes for $n = \sum_{t=1}^2 \sum_{j=1}^{J_t} n_{j,t} = 300$ probability forecasts prior to observing the indicator variables Y_1, \dots, Y_n of the events. We use the Brier score and its associated loss function $L(p, \hat{p}) = (p - \hat{p})^2$ to evaluate the probability forecasts and study by simulations the adequacy of the estimates $\hat{\beta}_n^2$ and \hat{s}_n^2 and their use in the normal approximations. The simulation study covers four scenarios and involves 1000 simulation runs for each scenario. Scenario 1 considers the Brier score of a forecasting rule, while Scenarios 2–4 consider the difference of Brier scores of two forecasts. The bucket sizes and how the p_i and \hat{p}_i are generated in each scenario are described as follows.

SCENARIO 1. There are ten buckets of size 15 each for each period. The common values $p_t(j)$ in the buckets are .1, .25, .3, .35, .4, .5, .65, .7, .75, and .8, respectively, for $t = 1, 2$. The probability forecast $\hat{p}_{t,k}$, $1 \leq k \leq 150$, made at time $t - 1$, uses covariate information to identify the bucket j associated with the k th event at time t and predicts that it occurs with probability $\bar{Y}_{t-1}(j)$, assuming that 150 indicator variables at time 0 are also observed so that $\bar{Y}_0(j)$ is available.

SCENARIO 2. For each period, there are nine buckets, three of which have size 2 and two of which have size 5; the other bucket sizes are 24, 30, 35 and 45 (one bucket for each size). The bucket probabilities $p_t(j)$ are i.i.d. random variables generated from Uniform (0,1). The forecast $\hat{p}_{t,k}$ is the same as that in Scenario 1, and there is another forecast $\hat{p}'_{t,k} = \bar{Y}_{t-1}$ that ignores covariate information.

SCENARIO 3. For each period, there are five buckets of size 30 each, and $p_t(j) = -0.1 + j/5$ for $j = 1, \dots, 5$. The two forecasts are the same as in Scenario 2.

SCENARIO 4. This is the same as Scenario 3, except that p_i is uniformly distributed on $[(j-1)/5, j/5]$ for $i \in I_{j,t}$, i.e., the bucket assumption is only approximately correct.

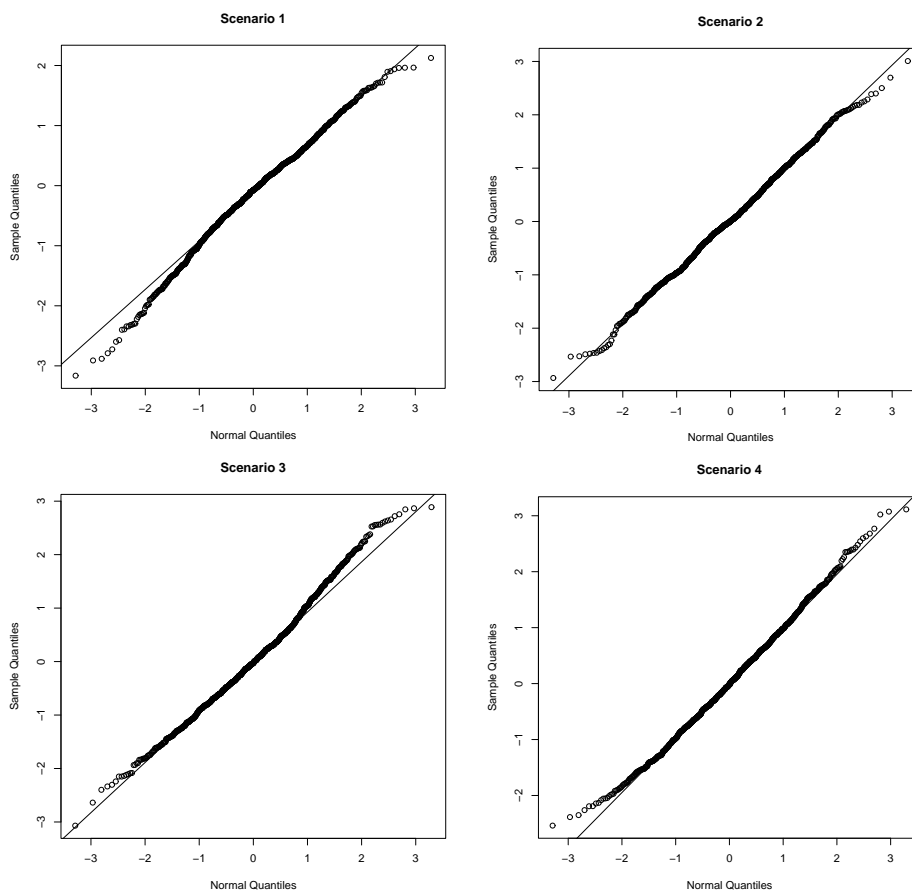


FIG 1. Q-Q plots for Scenarios 1-4.

Figure 1 gives the Q-Q plots of $\sqrt{n}(\hat{L}_n - n^{-1} \sum_{t=1}^2 \sum_{j=1}^{J_t} n_{j,t} \hat{v}_t(j) - L_n) / \hat{\beta}_n$ for Scenario 1 and $\sqrt{n}(\hat{\Delta}_n - \Delta_n) / \hat{s}_n$ for Scenarios 2-4. Despite the deviation from the assumed bucket model in Scenario 4, the Q-Q plot does not deviate much from the 45° line. Table 1 gives the means and 5-number summaries

TABLE 1
Simulation results for $\hat{\beta}_n/\beta_n$ (Scenario 1) and \hat{s}_n/s_n .

	Min.	1st Qrt.	Median	3rd Qrt.	Max.	Mean
Scenario 1	.6397	1.0840	1.1810	1.2830	1.6520	1.1780
Scenario 2	.7442	.9647	1.0060	1.0490	1.1970	1.0050
Scenario 3	.7586	.9506	1.0060	1.0570	1.2070	1.0010
Scenario 4	.7420	.9661	1.0180	1.0730	1.2240	1.0160

(minimum, maximum, median, 1st and 3rd quartiles) of \hat{s}_n/s_n for Scenarios 2–4 and $\hat{\beta}_n/\beta_n$ for Scenario 1.

5. Evaluating MOS Forecasts for 6 Cities. In this section we illustrate the proposed approach by considering some MOS probability forecasts of precipitation provided by the US National Weather Service. A commonly used *skill score* to evaluate weather forecasts is the percentage improvement in average score over that provided by climatology, denoted by \hat{p}_i^c and considered as an “unskilled” forecaster, i.e.,

$$S_n = \left\{ n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i^c) - n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i) \right\} / n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i^c). \quad (5.1)$$

Climatology refers to the historic relative frequency, also called the base rate, of precipitation; we can take it to be $\hat{p}_i^c = (M+1)^{-1} \sum_{t=-M}^0 Y_t$.

Noting that (5.1) is not a proper score although it is intuitively appealing, Winkler (1994) proposed to replace the average climatology score in the denominator of (5.1) by individual weights $l(\hat{p}_i, \hat{p}_i^c)$, i.e.,

$$W_n = n^{-1} \sum_{i=1}^n \{L(Y_i, \hat{p}_i) - L(Y_i, \hat{p}_i^c)\} / l(\hat{p}_i, \hat{p}_i^c) \quad (5.2)$$

where $l(p, c) = \{L(1, p) - L(1, c)\}I_{\{p \geq c\}} + \{L(0, p) - L(0, c)\}I_{\{p < c\}}$. Theorem 2(i) can be readily extended to show that Winkler’s score W_n is a consistent estimate of

$$w_n = n^{-1} \sum_{i=1}^n \{L(p_i, \hat{p}_i) - L(p_i, \hat{p}_i^c)\} / l(\hat{p}_i, \hat{p}_i^c) \quad (5.3)$$

and that $\sqrt{n}(W_n - w_n)/\tilde{s}_n$ has a limiting standard normal distribution, where

$$\tilde{s}_n^2 = n^{-1} \sum_{i=1}^n \delta_i^2 p_i (1 - p_i) / l^2(\hat{p}_i, \hat{p}_i^c). \quad (5.4)$$

Winkler (1994) used the score (5.2), in which $L(p, \hat{p}) = (p - \hat{p})^2$, to evaluate precipitation probability forecasts, with a 12- to 24-hour lead time, given by the US National Weather Service for 20 cities in the period between April 1966 and September 1983. Besides the score (5.4), he also computed the Brier score and the skill score (5.1) of these forecasts and found that both the Brier and skill scores have high correlations (0.87 and 0.76) whereas (5.2) has a much lower correlation 0.44 with average climatology, suggesting that (5.2) provides a better reflection of the “skill” of the forecasts over an unskilled forecasting rule (based on historic relative frequency).

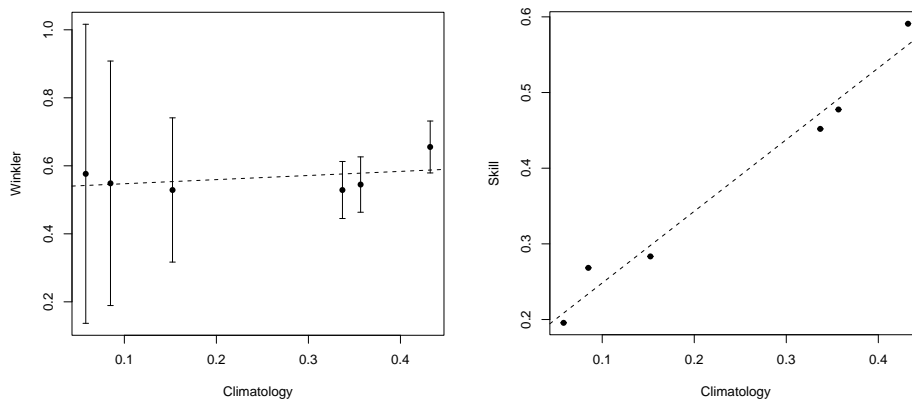


FIG 2. *The Winkler and skill scores versus climatology.*

Instead of using correlation coefficients, we performed a more detailed analysis of Winkler’s and skill scores to evaluate the MOS forecasts of precipitation for six cities: Las Vegas, NV; Phoenix, AZ; Albuquerque, NM; Queens, NY; Boston, MA; and Portland, OR (listed in increasing order of relative frequency of precipitation), during the period January 1, 2005, to December 31, 2009. The period January 1, 2002, to December 31, 2004, is used to obtain the past three years’ climatology, which is used as the reference unskilled score in the calculation of the skill score and Winkler’s score (5.2). The left panel plots Winkler’s score against the relative precipitation frequency taken from the period January 1, 2005, to December 31, 2009, which is simply the percentage of days with rain during that period and represents the climatology in (5.1). The dashed line in the right panel of Figure 2 represents linear regression of the skill scores on climatology and has a markedly positive slope of 0.95. In contrast, the regression line of Winkler’s scores on climatology, shown in the left panel of Figure 2, is relatively flat and has slope 0.12. Unlike skill scores, Winkler’s scores are proper and

TABLE 2
Brier scores B_1 and 95% confidence intervals for $\Delta(k)$.

	B_1	$\Delta(2)$	$\Delta(3)$	$\Delta(4)$	$\Delta(5)$	$\Delta(6)$	$\Delta(7)$
Queens, NY	.125	.021 $\pm .010$.012 $\pm .011$.020 $\pm .012$.010 $\pm .011$.015 $\pm .011$.007 $\pm .010$
Jefferson City, MO	.159	.005 $\pm .010$	-.005 $\pm .011$.007 $\pm .011$.024 $\pm .010$	-.000 $\pm .010$.008 $\pm .008$

provide consistent estimates of the average loss (5.3) involving the actual daily precipitation probabilities p_i for each city during the evaluation period. The vertical bar centered at the dot (representing Winkler's score) for each city is a 95% confidence interval for (5.3), using a conservative estimate of (5.4) that replaces $p_i(1-p_i)$ in (5.4) by $1/4$. The confidence intervals are considerably longer for cities whose relative frequencies \hat{p}_i^c of precipitation fall below 0.1 because $\delta_i^2/l^2(\hat{p}_i, \hat{p}_i^c)$ tends to be substantially larger when \hat{p}_i^c is small.

As another illustration of Theorem 2, we compare the Brier scores B_k for the k -day ahead forecasts $\hat{p}_t^{(k)}$, $1 \leq k \leq 7$, for Queens, NY, and Jefferson City, MO, provided by US National Weather Service from June 8, 2007, to March 31, 2009. Table 2 gives the values of B_1 and $B_k - B_{k-1}$ for $2 \leq k \leq 7$. Using $1/4$ to replace $p_i(1-p_i)$ in (3.5), we can use Theorem 2(i) to construct conservative 95% confidence intervals for

$$\Delta(k) = n^{-1} \left\{ \sum_{t=1}^n (p_t - \hat{p}_t^{(k)})^2 - \sum_{t=1}^n (p_t - \hat{p}_t^{(k-1)})^2 \right\},$$

in which p_t is the actual probability of precipitation on day t . These confidence intervals, which are centered at $B_k - B_{k-1}$, are given in Table 2. The results show significant improvements, by shortening the lead time by one day, in forecasting precipitation for both locations in the case $k = 1$, and for Queens when $k = 2, 3, 4, 6$.

6. Proofs.

PROOF OF THEOREM 1. As noted in the first paragraph of Section 3.1, $d_i := L(Y_i, \hat{p}_i) - L(p_i, \hat{p}_i)$ is a martingale difference sequence with respect to $\{\mathcal{F}_i\}$. In fact, since $L(y, \hat{p})$ is linear in y , we can write $L(y, \hat{p}) = a(\hat{p})y + b(\hat{p})$. Setting $y = 0$ and $y = 1$ in this equation yields $a(\hat{p}) = L(1, \hat{p}) - L(0, \hat{p})$. Moreover, $d_i = a(\hat{p}_i)(Y_i - p_i)$. Since $Y_i | \mathcal{F}_i \sim \text{Bernoulli}(p_i)$ and \hat{p}_i is \mathcal{F}_{i-1} -measurable,

$$E(d_i^2 | \mathcal{F}_{i-1}) = a^2(\hat{p}_i) p_i(1-p_i). \quad (6.1)$$

By (6.1), $\sum_1^n E(d_i^2 | \mathcal{F}_{i-1}) = \sum_1^n \{L(1, \hat{p}_i) - L(0, \hat{p}_i)\}^2 p_i(1 - p_i) = O(n)$ a.s., and therefore $n^{-1} \sum_{i=1}^n d_i \rightarrow 0$ a.s. by the martingale strong law (Williams, 1991, Sect. 12.14), proving $\hat{L}_n - L_n \rightarrow 0$ a.s. Moreover, if $n^{-1} \sum_1^n E(d_i^2 | \mathcal{F}_{i-1})$ converges in probability to a non-random positive constant, then we can apply the martingale central limit theorem (Durrett, 1996, Sect. 7.7) to conclude that $\sqrt{n}(\hat{L}_n - L_n)/\sigma_n$ has a limiting standard normal distribution, noting that the conditional Lindeberg condition is satisfied since Y_i is bounded. \square

PROOF OF LEMMA 1. Note that $\hat{v}_t(j) = \sum_{i \in I_{j,t}} (Y_i - \bar{Y}_t(j))^2 / (n_{j,t} - 1)$ and that

$$E(\hat{v}_t(j) | \mathcal{G}_{t-1}) = p_t(j)(1 - p_t(j)), \quad (6.2)$$

which is the variance of Y_i associated with $I_{j,t}$. Therefore $\sum_{j=1}^{J_t} \{\hat{v}_t(j) - p_t(j)(1 - p_t(j))\} \{\sum_{i \in I_{j,t}} [L(1, \hat{p}_i) - L(0, \hat{p}_i)]^2\}$ is a martingale difference sequence with respect to $\{\mathcal{G}_t\}$. Hence we can apply the martingale strong law as in the proof of Theorem 1 to show that $\hat{\sigma}_n^2 - \sigma_n^2$ converges a.s., and the same argument also applies to $\hat{s}_n^2 - s_n^2$. \square

PROOF OF LEMMA 2. Use $L(p, \hat{p}) = (p - \hat{p})^2$ to express $n\{\hat{L}_n - n^{-1} \sum_{t=1}^T \sum_{j=1}^{J_t} n_{j,t} \hat{v}_t(j) - L_n\}$ as

$$\sum_{i=1}^n (1 - 2\hat{p}_i) (Y_i - p_i) - \sum_{t=1}^T \sum_{j=1}^{J_t} n_{j,t} [\hat{v}_t(j) - p_t(j)(1 - p_t(j))], \quad (6.3)$$

which is the difference of two martingales and is therefore a martingale. To compute the conditional variance (or predictable variation) of (6.3), we can use the ‘‘angle bracket’’ notation and formulas for predictable variation and covariation (Williams, 1991, Sect. 12.12) to obtain

$$\begin{aligned} \left\langle \sum_{i=1}^n (1 - 2\hat{p}_i) (Y_i - p_i) \right\rangle &= \sum_{i=1}^n (1 - 2\hat{p}_i)^2 E\left((Y_i - p_i)^2 | \mathcal{F}_{i-1}\right) \\ &= \sum_{i=1}^n (1 - 2\hat{p}_i)^2 p_i(1 - p_i), \end{aligned} \quad (6.4)$$

$$\begin{aligned} &\left\langle \sum_{i=1}^n (1 - 2\hat{p}_i) (Y_i - p_i), \sum_{t=1}^T \sum_{j=1}^{J_t} n_{j,t} [\hat{v}_t(j) - p_t(j)(1 - p_t(j))] \right\rangle \\ &= \sum_{t=1}^T \sum_{j=1}^{J_t} \left[\sum_{i \in I_{j,t}} (1 - 2\hat{p}_i) \right] p_t(j)(1 - p_t(j))(1 - 2p_t(j)), \end{aligned} \quad (6.5)$$

$$\begin{aligned}
& \left\langle \sum_{t=1}^T \sum_{j=1}^{J_t} n_{j,t} [\hat{v}_t(j) - p_t(j)(1 - p_t(j))] \right\rangle \\
&= \sum_{t=1}^T \sum_{j=1}^{J_t} \{n_{j,t} p_t(j)(1 - p_t(j)) [1 - 4p_t(j)(1 - p_t(j))] \\
&\quad + 2n_{j,t} p_t^2(j)(1 - p_t(j))^2 / (n_{j,t} - 1)\}. \quad (6.6)
\end{aligned}$$

Combining (6.4), (6.5), and (6.6) yields the formula (3.8) for the conditional variance of (6.3) divided by n .

As shown in the proof of Lemma 1, $\hat{v}_t(j)$ is a “conditionally unbiased” estimate of $p_t(j)(1 - p_t(j))$ in the sense of (6.2). If $Y \sim \text{Bernoulli}(p_i)$ then $E(Y - p)^3 = p(1 - p)(1 - 2p)$. Hence a conditionally unbiased estimate of $p_t(j)(1 - p_t(j))(1 - 2p_t(j))$ is

$$[n_{j,t}/(n_{j,t} - 1)]^3 \sum_{i \in I_{j,t}} (Y_i - \bar{Y}_t(j))^3, \quad (6.7)$$

analogous to (6.2). Let X_1, \dots, X_m be i.i.d. random variables. As is well known, the sample variance $\hat{v} = \sum_{i=1}^m (X_i - \bar{X})^2 / (m - 1)$ is a U -statistic of order 2, with kernel $h(X_i, X_k) = (X_i - X_k)^2 / 2$. An unbiased estimate of the variance of the U -statistic is provided by the jackknife estimate

$$\frac{4(m-1)}{m(m-2)^2} \sum_{i=1}^m \left\{ \frac{1}{m-1} \sum_{\substack{k=1 \\ k \neq i}}^m h(X_i, X_k) - \hat{v} \right\}^2; \quad (6.8)$$

see Arvesen (1969). This corresponds to the estimate used in the last summand of (3.9) and can be shown to yield a conditionally unbiased estimate. The rest of the argument is similar to that of Lemma 1. \square

7. Discussion. The average score $n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i)$ measures the divergence of the predicted probabilities \hat{p}_i , which lie between 0 and 1, from the indicator variables Y_i that can only have values 0 or 1. As noted by Lichten Dahl and Winkler (2007), this tends to encourage more aggressive bets on the 0/1 outcomes, rather than the forecaster’s estimates of the event probabilities. For example, an estimate of 95% probability may lead to a probability forecast of 100% for a higher reward associated with the 0/1 indicator variable Y_i ; see also Mason (2008), who gives an example in which a forecaster is encouraged to give such “dishonest” forecasts. This difficulty would disappear if one uses L to compare \hat{p}_i with the actual p_i , rather than with the Bernoulli(p_i) random variable Y_i . Because the p_i are unknown, this

is not feasible and therefore a proper score $L(Y_i, \hat{p}_i)$ has been used to evaluate a probability forecast. In Section 3.2 and Section 5, we have shown that it is possible to use $L(p_i, \hat{p}_i) - L(p_i, \hat{p}'_i)$ for comparing two forecasters and to construct confidence intervals of the average loss difference. A key idea underlying this development is the linear equivalent of a loss function introduced in Section 2.

Another important idea is the application of martingale theory (strong laws and central limit theorems) in Section 3 and the Appendix to develop consistent estimators of $n^{-1} \sum_{i=1}^n L(p_i, \hat{p}_i)$ and their standard errors. Dawid (1982, Appendix) has also applied martingale theory to prove that Bayesian forecasts are well calibrated in the following sense. He assumes a “subjective probability distribution” for the events so that Bayesian forecasts are given by

$$\hat{p}_i = E(Y_i | \mathcal{B}_{i-1}), \quad (7.1)$$

where \mathcal{B}_t is the “totality of events known to the forecaster” up to time t . Letting $\xi_t = 1$ or 0 according to whether time t is included in the “test set” to evaluate forecasts, he calls the test set “admissible” if ξ_t depends only on \mathcal{B}_{t-1} , and uses martingale theory to show that under the subjective probability measure,

$$\frac{\sum_{i=1}^n \xi_i Y_i - \sum_{i=1}^n \xi_i \hat{p}_i}{\sum_{i=1}^n \xi_i} \rightarrow 0 \text{ with probability 1 on } \left\{ \sum_{i=1}^n \xi_i = \infty \right\}. \quad (7.2)$$

From (7.2), it follows that for any $0 < x < 1$, the long-run average of Y_i (under the subjective probability measure) associated with $\hat{p}_i = x$ (i.e., $\xi_i = I_{\{\hat{p}_i=x\}}$) is equal to x provided that $\sum_{i=1}^n I_{\{\hat{p}_i=x\}} \rightarrow \infty$. Note that Dawid’s well-calibration theorem (7.2) involves the subjective probability measure, whereas the asymptotic theory of our loss function approach uses the actual probability measure that generates the unobserved p_i . This has the obvious advantage that the evaluator’s prior beliefs need not be the same as the forecaster’s.

DeGroot and Fienberg (1983) have noted that well-calibrated forecasts need not reflect the forecaster’s “honest subjective probabilities,” i.e., need not satisfy (7.1). On the other hand, the Bayesian forecasts (7.1) “are not necessarily accurate in all respects and they are not necessarily of much use to anyone.” They therefore use a criterion called “refinement” to compare well-calibrated forecasts. Using a loss function to compare the forecast \hat{p}_i and the subsequent indicator variable Y_i is more direct, and Schervish

(1989, Sect. 3) has used a framework of two-decision problems involving these loss functions to develop a method for comparing forecasters. Our approach that considers $L(p_i, \hat{p}'_i) - L(p_i, \hat{p}''_i)$ can be regarded as a further step in this direction.

References.

- ARVESEN, J. N. (1969). Jackknifing U -statistics. *Ann. Math. Statist.*, **40** 2076–2100.
- BASEL COMMITTEE ON BANKING SUPERVISION (2006). Basel II: International convergence of capital measurement and capital standards: A revised framework. Available online, URL <http://www.bis.org/publ/bcbs128.htm>.
- BRIER, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78** 1–3.
- COX, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, **45** 562–565. URL <http://biomet.oxfordjournals.org/cgi/reprint/45/3-4/562.pdf>.
- DAWID, A. P. (1982). The well-calibrated Bayesian. *J. Amer. Statist. Assoc.*, **77** 605–613. With comments by Joseph B. Kadane and D. V. Lindley and a rejoinder by the author, URL [http://links.jstor.org/sici?sici=0162-1459\(198209\)77:379<605:TWB>2.0.CO;2-Q&origin=MSN](http://links.jstor.org/sici?sici=0162-1459(198209)77:379<605:TWB>2.0.CO;2-Q&origin=MSN).
- DEGROOT, M. H. and FIENBERG, S. E. (1983). The comparison and evaluation of forecasters. *Statistician*, **32** 12–22.
- DURRETT, R. (1996). *Probability: Theory and Examples*. 2nd ed. Duxbury Press, Belmont, CA.
- FOX, C. R. and BIRKE, R. (2002). Forecasting trial outcomes: Lawyers assign higher probability to possibilities that are described in greater detail. *Law and Human Behavior*, **26** 159–173.
- GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **69** 243–268. URL <http://dx.doi.org/10.1111/j.1467-9868.2007.00587.x>.
- GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.*, **102** 359–378. URL <http://dx.doi.org/10.1198/016214506000001437>.
- GOOD, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. Ser. B.*, **14** 107–114.
- HARI, P. N., ZHANG, M.-J., ROY, V., PEREZ, W. S., BASHEY, A., TO, L. B., ELFENBEIN, G., FREYTES, C. O., GALE, R. P., GIBSON, J., KYLE, R. A., LAZARUS, H. M., MCCARTHY, P. L., MILONE, G. A., PAVLOVSKY, S., REECE, D. E., SCHILLER, G., VELA-OJEDA, J., WEISDORF, D. and VESOLE, D. (2009). Is the international staging system superior to the Durie–Salmon staging system? A comparison in multiple myeloma patients undergoing autologous transplant. *Leukemia*, **23** 1528–1534.
- LICHTENDAHL, K. C., JR. and WINKLER, R. L. (2007). Probability elicitation, scoring rules, and competition among forecasters. *Management Sci.*, **53** 1745–1755.
- MASON, S. J. (2008). Understanding forecast verification statistics. *Meteorol. Appl.*, **15** 31–40.
- MURPHY, A. H. and WINKLER, R. L. (1984). Probability forecasting in meteorology. *J. Amer. Statist. Assoc.*, **79** 489–500.
- REDELMEIER, D. A., BLOCH, D. A. and HICKAM, D. H. (1991). Assessing predictive accuracy: How to compare Brier scores. *J. Clin. Epidemiol.*, **44** 1141–1146.
- SCHERVISH, M. J. (1989). A general method for comparing probability assessors. *Ann. Statist.*, **17** 1856–1879. URL <http://dx.doi.org/10.1214/aos/1176347398>.

- SPIEGELHALTER, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Stat. Med.*, **5** 421–433.
- WILLIAMS, D. (1991). *Probability with Martingales*. Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge.
- WINKLER, R. L. (1994). Evaluating probabilities: Asymmetric scoring rules. *Management Sci.*, **40** 1395–1405.

SEQUOIA HALL, 390 SERRA MALL
STANFORD, CA 94305-4065
E-MAIL: lait@stanford.edu