ASSESSING THE SKILL OF YES/NO FORECASTS FOR MARKOV OBSERVATIONS

WILLIAM M. BRIGGS AND DAVID RUPPERT

Abstract. Mozer and Briggs (2003) and Briggs and Ruppert (2003) recently introduced a new, easy-to-calculate economic skill score for use in yes/no forecast decisions, of which precipitation forecast decisions are an example. The advantage of this new climate skill score is that the sampling distribution is known, which allows one to perform hypothesis tests on collections of forecasts and to say whether a given skill score is significant or not.

Skill, as ever, is defined as improvement over an optimal naive prediction. We show that the optimal naive prediction depends on both the base rate (the climatology) of the event being forecasted, and the loss one would incur if one were to make an incorrect decision based on the forecast.

Here, we take the climate skill score and extend it to the case where the predicted series is first-order Markov in nature, of which, again, precipitation occurrence series are an example. We show that Markov skill is different and more demanding than is persistence skill. Persistence skill is defined as improvement over forecasts which state that the next value in a series will equal the present value. We also define the optimal naive prediction in the Markov case.

Surprisingly, it turns out that the form of the Markov skill score is identical to the climate skill score, making calculations simple. The distribution of the Markov skill is more complex than is the distribution of the climate skill score, however. The distribution for the Markov skill score is presented, and examples of hypothesis testing for precipitation forecasts are given. We graph these skill scores for a wide range of forecast-user loss functions, a process which makes their interpretation simple.

1. Introduction

In a previous paper, Briggs and Ruppert (2003; from here, BR), developed a test for skill for forecasts of dichotomous events Y. The events Y_i in this test were assumed to be independent of each Y_j for all $i \neq j$. In this paper, we extend the original skill score test to situations where the event is a two-state Markov chain. Precipitation occurrence at a point is often a good example of such series.

Much work has been done in the area of investigating forecast value and forecast verification, most notably in the works of Murphy (Murphy, 1991; Murphy, 1997; Murphy and Winkler, 1987; Murphy and Ehrendorfer, 1987; to name only a few), Schervish (1989), Briggs and Leving (1998), Meeden (1979), and Wilks (2001). Wilks (1995), Mason (2003), and Livezey (2003) provide a detailed list of skill scores for categorical events, such as we consider here. Wilks (1991) began work in showing how the dependent nature of observation process interacts with forecast verification, work which we continue here.

Date: 16 Oct 03.

Thanks to Dan Wilks for his insightful comments and criticisms.

BR defined forecasts $\widetilde{X} \in [0,1]$ made for events Y. Here, we are interested in the two-decision problem, which is when a decision maker acts on the forecast $\widetilde{X}$ and makes one of two decisions: d_1 is he believes $Y = 1$ will occur, or d_0 is he believes $Y = 0$ will occur. The decision maker faces a loss k_{01} is he takes d_1 and $Y = 0$ occurs, and has a loss k_{10} is he takes d_0 and $Y = 1$ occurs. The loss can always be parameterized so that $\theta = k_{01}/(k_{01} + k_{01})$. When $\theta = 1/2$, the loss is said to be symmetric. BR show that this parameterization allows us to transform the forecast $\widetilde{X}$, for the two-decision problem, as $X^E = I(\widetilde{X} \geq \theta)$, where the superscript E designates that X^E is an *expert* forecast, which is any forecast that is not the optimal naive forecast. The optimal naive forecast X^N for Y is the forecast one would make knowing only $p = P(Y = 1)$. It is easy to show that this is $X^N = I(p > \theta)$.

Skill is now defined in two ways. The first is when the expected loss of the expert forecast is less than the expected loss of the optimal naive forecast. The second definition of skill is when $P(X^E = Y) > P(X^N = Y)$. BR shows that these two definitions are identical when $\theta = 1/2$, or when the loss is symmetric. BR developed a skill score and a test statistic for skill, where the key parameter was $p_{1|1} = P(Y = 1|X = 1)$, which was less than or equal to θ under the null hypothesis of no skill.

This work will extend the same concepts developed in BR to events Y_i where $\{Y_i\}$ is a two-state Markov chain. We first define persistence as the forecast $X^P = Y_{i-1}$ for all i. We show in Section 3 that skill, when Y is Markov, is not the same as skill of a persistence forecast. In Section 2, we develop a test for comparing any two forecasts for the same event, which we later apply in Section 4 with a persistence forecast and the optimal naive forecast. Finally, an Appendix is given that may help orient the reader, as it highlights the results originally presented in BR.

2. Comparing competing forecasts

In this Section, we develop a simple framework to compare competing forecasts for the *same* event. In this framework, there are two forecasts X_1 and X_2. Define $Z_1 = I(Y = X_1)$, which is the indicator that the first forecast is correct, and define $Z_2 = I(Y = X_2)$, which is the indicator of the second forecast being correct. We have that $P(Z_1 = Z_2 = 1)$ is the probability that both forecasts are correct and $P(Z_1 = Z_2 = 0)$ is the probability that they are both wrong. The probabilities of interest are $P(Z_1 = 1, Z_2 = 0)$ and $P(Z_1 = 0, Z_2 = 1)$, that is, the factors designating those times when one forecast was correct while the other was wrong. We assume that the loss for one forecast being correct while the other incorrect is symmetric.

The development in this Section is not entirely new. The tests here are the similar to McNemar's test for matched pairs and to its refinement by Mosteller (1952). Depending on the particual null hypothesis chosen, our statistic is only slightly different to the classic statistic and, again depending on the null, one could use the classic statistic in place of this one (this is explained below). The development here shows how the comparison test operates with respect to the skill test.

We assume that there is an i.i.d. sequence $\{(X_{1i}, X_{2i}, Y_i) : i = 1, \ldots, n\}$ and we define $Z_{ji} = I(X_{ji} = Y_i)$, $j = 1, 2$. From these data one obtains a contingency table of counts, such as illustrated in Table 1. One possible null hypothesis is

$$H_0 : P(Z_1 = 1, Z_2 = 0) = P(Z_1 = 0, Z_2 = 1) \tag{2.1}$$

TABLE 1. Forecast comparison contingency table.

	$Z_1 = 1$	$Z_1 = 0$
$Z_2 = 1$	m_{11}	m_{01}
$Z_2 = 0$	m_{10}	m_{00}

with the two-sided alternative

$$P(Z_1 = 1, Z_2 = 0) \neq P(Z_1 = 0, Z_2 = 1). \tag{2.2}$$

This null is similar to the one normally stated for McNemar's test. Another null is

$$P(Z_1 = 1, Z_2 = 0) \leq P(Z_1 = 0, Z_2 = 1) \tag{2.3}$$

with the one-sided alternative

$$P(Z_1 = 1, Z_2 = 0) > P(Z_1 = 0, Z_2 = 1). \tag{2.4}$$

The likelihood is

$$L(\{z_{i,j}\}_{i,j=0,1}|Z_1, Z_2) = \prod_i \prod_j z_{ij}^{m_{ij}},$$

where $z_{11} = P(Z_1 = 1, Z_2 = 1)$ and so on. Under the null (2.1) the estimates are

$$\begin{aligned}
\widehat{P}(Z_1 = 1, Z_2 = 1) &= \frac{m_{11}}{m_{++}}, \\
\widehat{P}(Z_1 = 1, Z_2 = 0) &= \frac{m_{10} + m_{01}}{2m_{++}}, \\
\widehat{P}(Z_1 = 0, Z_2 = 1) &= \frac{m_{10} + m_{01}}{2m_{++}}, \\
\widehat{P}(Z_1 = 0, Z_2 = 0) &= \frac{m_{00}}{m_{++}}.
\end{aligned}$$

The likelihood ratio statistic G_c is computed easily; the terms involving z_{11} and z_{00} drop out, leaving

$$G_c = 2\left\{ m_{10} \log\left(\frac{2m_{10}}{m_{10} + m_{01}}\right) + m_{01} \log\left(\frac{2m_{01}}{m_{10} + m_{01}}\right)\right\}. \tag{2.5}$$

The distribution of G_c, assuming the two-sided null (2.1), has an asymptotic χ^2 distribution with one degree of freedom.

If a one-sided null is chosen the exact form of G_c changes because the MLEs under the null are different than what are given above. But, if one of the forecasts is the optimal naive forecast then G_c is the same as BR's G and has an asymptotic $1/2\chi_0^2 + 1/2\chi_1^2$ distribution (Self and Liang, 1987; BR, Section 2; see also the Appendix).

The classic statistic for independence between (null 2.1) Z_1 and Z_2 is $G_C = (|m_{01} - m_{10}| - 1)^2/(m_{01} + m_{10})$ which has a χ_1^2 distribution.

This comparison test can be viewed as a test for climate skill in a different guise, if the first forecast is the expert forecast and the optimal naive forecast is the second (the null hypothesis is, of course, (2.3)).

3. Skill tests and Skill Scores

Skill, if it exists when $\{Y_i\}$ is a two-state Markov chain, is known as Markov skill because the optimal naive forecast of each Y_i is based on the previous observation Y_{i-1}. Markov skill is not identical with persistence skill, in which the naive forecast for Y_i is Y_{i-1} for all i, as will be shown below. We assume only that $\{Y_i\}$ is Markov, and not, for example, that $\{Y_i, X_i\}$ is bivariate Markov. No further conditions are put on $\{X_i\}$ except that to require $P(X_i|Y_{i-1})$ be constant for all i.

Daily occurrence of precipitation is a common example of Markov data (Wilks, 1995). We could condition skill scores for forecasts of Markov data on the event Y_{i-1} and use the results from BR. That is, individual tests of climate skill can be carried out for the cases in which $Y_{i-1} = 1$ and $Y_{i-1} = 0$. This is a useful as a performance diagnostic to highlight those experts who possibly forecast badly in one situation but well in the other. This approach in ultimately unsatisfying for formal testing because it requires two scores, one for $Y_{i-1} = 0$ and another for $Y_{i-1} = 1$. It requires two tests for the same reason, where it is desirable to have only one composite score for the entire set of data.

3.1. **Model.** Consider the factorization

$$P(Y_i, X_i, Y_{i-1}) = P(Y_i|X_i, Y_{i-1})P(X_i|Y_{i-1})P(Y_{i-1}). \tag{3.1}$$

Other factorizations are, of course, possible but it turns out that this form is the most mathematically convenient to work with. The full model may be expanded to (with $P(Y_{i-1} = 1) = p$) the following set of equations. The methodology is exactly that used in BR.

$$\begin{aligned}
P(Y_i = 1, X_i = 1, Y_{i-1} = 1) &= p_{1|11}p_{+1|1}p \\
P(Y_i = 1, X_i = 1, Y_{i-1} = 0) &= p_{1|10}p_{+1|0}(1-p) \\
P(Y_i = 1, X_i = 0, Y_{i-1} = 1) &= p_{1|01}(1-p_{+1|1})p \\
P(Y_i = 1, X_i = 0, Y_{i-1} = 0) &= p_{1|00}(1-p_{+1|0})(1-p) \\
P(Y_i = 0, X_i = 1, Y_{i-1} = 1) &= (1-p_{1|11})p_{+1|1}p \\
P(Y_i = 0, X_i = 1, Y_{i-1} = 0) &= (1-p_{1|10})p_{+1|0}(1-p) \\
P(Y_i = 0, X_i = 0, Y_{i-1} = 1) &= (1-p_{1|01})(1-p_{+1|1})p \\
P(Y_i = 0, X_i = 0, Y_{i-1} = 0) &= (1-p_{1|00})(1-p_{+1|0})(1-p),
\end{aligned}$$

where $p_{1|11} = P(Y_i = 1|X_i = 1, Y_{i-1} = 1)$, $p_{+1|1} = P(X_i = 1|Y_{i-1} = 1)$, $p_{1|10} = P(Y_i = 1|X_i = 1, Y_{i-1} = 0)$, $p_{+1|0} = P(X_i = 1|Y_{i-1} = 0)$, $p_{1|01} = P(Y_i = 1|X_i = 0, Y_{i-1} = 1)$, and $p_{1|00} = P(Y_i = 1|X_i = 0, Y_{i-1} = 0)$.

We shall also need to define the parameters that characterize the Markov nature of Y. These are

$$\begin{aligned}
p_{1+|1} &= P(Y_i = 1|Y_{i-1} = 1) \\
p_{0+|1} &= P(Y_i = 0|Y_{i-1} = 1) \\
p_{1+|0} &= P(Y_i = 1|Y_{i-1} = 0) \\
p_{0+|0} &= P(Y_i = 0|Y_{i-1} = 0).
\end{aligned}$$

It happens that $p_{0+|1} = 1 - p_{1+|1}$ and $p_{0+|0} = 1 - p_{1+|0}$ so that only two parameters are needed to fully specify the Markov nature of Y.

It is also helpful to define the following counts. Let $n_{j,k,l}$, where $j,k,l \in \{0,1\}$, be the counts for the cells Y_i, X_i, and Y_{i-1}. For example, $n_{111} = \sum_{i=2}^{n} Y_i X_i Y_{i-1}$, and $n_{000} = \sum_{i=2}^{n}(1-Y_i)(1-X_i)(1-Y_{i-1})$.

3.2. **Markov skill tests.** All of the parameters of this model neatly separate in the likelihood, making estimation easy. For example, the part of the likelihood relating to the parameter $P(Y_i = 1|X_i = 1, Y_{i-1} = 1) = p_{1|11}$ is

$$\prod p_{1|11}^{Y_i X_i Y_{i-1}} (1 - p_{1|11})^{(1-Y_i)X_i Y_{i-1}}.$$

It is simple to differentiate and solve for the MLE for all such parameters. It turns out that the parameters $p, p_{+1|1}$ and $p_{+1|0}$ will not play a role in the likelihood ratio test as their MLEs are the same under both the null and alternative hypotheses for either pair (2.1) and (2.2) or (2.3) and (2.4). We will use the convention that the replacement of an index by "+" means summation over that index so, for example, $n_{++1} = \sum_i \sum_j n_{ij1}$. The unrestricted MLEs are

$$\begin{aligned} \widehat{p} &= \frac{n_{++1}}{n_{+++}} \\ \widehat{p}_{+1|1} &= \frac{n_{+11}}{n_{++1}} \\ \widehat{p}_{+1|0} &= \frac{n_{+10}}{n_{++0}}. \end{aligned}$$

The other parameters do change and the unrestricted MLES are

$$\begin{aligned} \widehat{p}_{1|11} &= \frac{n_{111}}{n_{+11}} \\ \widehat{p}_{1|10} &= \frac{n_{110}}{n_{+10}} \\ \widehat{p}_{1|01} &= \frac{n_{101}}{n_{+01}} \\ \widehat{p}_{1|00} &= \frac{n_{100}}{n_{+00}}. \end{aligned}$$

The optimal naive forecast must now be defined. It turns out that there are four situations, that is, four circumstances that dictate different optimal naive forecasts. In BR there were two situations, but we transformed one if necessary to ensure that $P(Y = 1) \leq 1/2$. This shall also be be done here, leaving us to focus on one situation for the sake of an example. The other three situations will be removed to the Appendix.

We assume that the events $\{Y_i, Y_{i-1}\}$ are such that $p_{1+|1} < \theta$ and $p_{1+|0} < \theta$, that is, the probability that $Y_i = 1$ no matter the value of Y_{i-1} is always less than θ. This gives that the optimal naive forecast is always 0. Note that the optimal naive forecast is different than a true persistence forecast, which would be $X_i = Y_{i-1}$ for all i. One solution for deriving persistence skill (but one which ignores the Markov nature of Y) is to use the comparative forecast test developed earlier with the first set of forecasts assigned to the expert, and the second set of forecasts assigned to persistence. Examples of this will be given later.

It is easier (because of notation) to first define skill in the Markov case in terms of accuracy, rather than on expected loss (details on how to define skill based on expected loss are removed to the Appendix). In BR it was shown that the test for climate skill based on expected loss is equivalent to showing that the probability

of an expert prediction is correct is larger than the probability of $Y = 0$ (where $P(Y = 0) > P(Y = 1)$). In the Markov case, in order for collection of predictions to be skillful the probability of a correct forecast must be greater than the maximum of the probability of $Y_i = 1$ and the probability of $Y_i = 0$ given Y_{i-1}. The null hypothesis of no skill is

$$H_0 : P(Y_i = X_i|Y_{i-1}) \leq \max\left(P(Y_i = 1|Y_{i-1}), P(Y_i = 0|Y_{i-1})\right) \tag{3.2}$$

for all values of $\{Y_i\}$. The right hand side of (3.2) is the probability that the optimal naive forecast is correct. In the case of the optimal naive forecast always equaling zero, this gives:

$$\begin{aligned} H_0 : \quad & P(Y_i = X_i|Y_{i-1} = 1) \leq P(Y_i = 0|Y_{i-1} = 1), \text{ and} \\ & P(Y_i = X_i|Y_{i-1} = 0) \leq P(Y_i = 0|Y_{i-1} = 0). \end{aligned}$$

The result is analogous to that found in BR. The complete null hypothesis is:

$$H_0 : (p_{1|11} \leq \frac{1}{2}, p_{1|10} \leq \frac{1}{2}).$$

Asymmetric loss can been introduced in the same manner as in BR with the changes to the null hypothesis accomplished in the obvious way (details are left to the Appendix). The final null is then

$$H_0 : (p_{1|11} \leq \theta, p_{1|10} \leq \theta).$$

Once again, all parameters except those indicated in the null hypothesis have the same MLEs in both the null and alternate hypotheses. The LRS (likelihood ration statistic) depends on only two parameters, $p_{1|11}$ and $p_{1|10}$, which are maximized under the null with estimates $\widetilde{p}_{1|11} = \min\{\frac{n_{111}}{n_{+11}}, \theta\}$ and $\widetilde{p}_{1|10} = \min\{\frac{n_{110}}{n_{+10}}, \theta\}$. Substitution leads to the LRS:

$$\begin{aligned} G_M = 2n_{111} \log\left(\frac{\widehat{p}_{1|11}}{\widetilde{p}_{1|11}}\right) + 2n_{011} \log\left(\frac{1 - \widehat{p}_{1|11}}{1 - \widetilde{p}_{1|11}}\right) + \\ 2n_{110} \log\left(\frac{\widehat{p}_{1|10}}{\widetilde{p}_{1|10}}\right) + 2n_{010} \log\left(\frac{1 - \widehat{p}_{1|10}}{1 - \widetilde{p}_{1|10}}\right). \end{aligned}$$

There are four situations under the null: when both $\frac{n_{111}}{n_{+11}}$ and $\frac{n_{110}}{n_{+10}}$ are greater than θ then $\widetilde{p}_{1|11} = \widetilde{p}_{1|10} = \theta$ and $G_M > 0$; when $\frac{n_{111}}{n_{+11}} \leq \theta$ and $\frac{n_{110}}{n_{+10}} > \theta$ then $\widetilde{p}_{1|11} = \widehat{p}_{1|11}$ and $\widetilde{p}_{1|10} = \theta$ and $G_M > 0$; when $\frac{n_{111}}{n_{+11}} > \theta$ and $\frac{n_{110}}{n_{+10}} \leq \theta$ then $\widetilde{p}_{1|11} = \theta$ and $\widetilde{p}_{1|10} = \widehat{p}_{1|10}$ and $G_M > 0$; or when $\frac{n_{111}}{n_{+11}} \leq \theta$ and $\frac{n_{110}}{n_{+10}} \leq \theta$ then $\widetilde{p}_{1|11} = \widehat{p}_{1|11}$ and $\widetilde{p}_{1|10} = \widehat{p}_{1|10}$ and $G_M = 0$. This allows us to rewrite G_M as

$$\begin{aligned} G_M = & \left(2n_{1|1} \log\left[\frac{n_{111}}{n_{+11}\theta}\right] + 2n_{011} \log\left[\frac{n_{011}}{n_{+11}(1-\theta)}\right]\right) I\left(\frac{n_{111}}{n_{+11}} > \theta\right) + \\ & \left(2n_{110} \log\left[\frac{n_{110}}{n_{+10}\theta}\right] + 2n_{010} \log\left[\frac{n_{010}}{n_{+10}(1-\theta)}\right]\right) I\left(\frac{n_{110}}{n_{+10}} > \theta\right) \end{aligned} \tag{3.3}$$

This statistic has an asymptotic mixture distribution under the null of $1/4\chi_0^2 + 1/2\chi_1^2 + 1/4\chi_2^2$ where χ_k^2 is the chi-square distribution with k degrees of freedom and χ_0^2 is point mass at 0 (see Self and Liang, 1987; an extension of their case 5).

3.3. **Markov skill score.** A skill score can now be created, as in BR. A common form for such a score is (see Wilks, 1995 for a more complete discussion of skill scores):

$$K_\theta(y, x^E) = \frac{E(k^N) - E(k^E)}{E(k^N)}, \tag{3.4}$$

where $E(k^N)$ is expected loss for the optimal naive forecast, and $E(k^E)$ is the expected loss for the expert forecast. There are two parts to that equation, $E(k^N) - E(k^E)$ and $E(k^N)$. For $E(k^N) - E(k^E)$, it is easy to show that we have

$$(p_{1|11} - \theta)p_{+1|1}p + (p_{1|10} - \theta)p_{+1|0}(1-p).$$

Also, $E\{k(Y, X^N)\}$ is

$$(1-\theta)(p_{1|11}p_{+1|1}p + p_{1|01}(1-p_{+1|1})p + p_{1|10}p_{+1|0}(1-p) + p_{1|00}(1-p_{+1|0})(1-p)).$$

An estimate for K_θ comes from substituting the estimates for $p_{1|11}$, $p_{1|01}$ and so on into these equations. Details will be left to the Appendix. Upon slugging through the algebra, we find that

$$\widehat{K}_\theta = \frac{(1-\theta)n_{111} - \theta n_{011} + (1-\theta)n_{110} - \theta n_{010}}{(n_{111} + n_{101})(1-\theta) + (n_{110} + n_{100})(1-\theta)}. \tag{3.5}$$

However, it is the case that $n_{111} + n_{110} = n_{11+}$, where n_{11+} is the number of days when $Y_{i-1} = 1$ and $Y_{i-1} = 0$. Similar facts hold for n_{110} and n_{010} and so on. What this means is that (3.5) ultimately collapses to

$$\widehat{K}_\theta = \frac{(1-\theta)n_{11+} - \theta n_{01+}}{(n_{11+} + n_{10+})(1-\theta)}. \tag{3.6}$$

which is identical to the original climate skill score developed in BR, which is not surprising since the optimal naive forecast is always 0 (as it was in the climate skill score). This makes computation simple, but more can be done because (3.5) can be written in a more insightful manner and decomposed into parts for when $Y_{i-1} = 1$ and when $Y_{i-1} = 0$.

Let $D = (n_{111} + n_{101})(1-\theta) + (n_{110} + n_{100})(1-\theta)$, which is the denominator of equation (3.5). We can now rewrite that equation:

$$\begin{aligned}
\widehat{K}_\theta &= \frac{(n_{111} + n_{011})(1-\theta)}{(n_{111} + n_{011})(1-\theta)} \frac{(1-\theta)n_{111} - \theta n_{011}}{D} \\
&\quad + \frac{(n_{110} + n_{010})(1-\theta)}{(n_{110} + n_{010})(1-\theta)} \frac{(1-\theta)n_{110} - \theta n_{010}}{D} \\
&= \frac{(n_{111} + n_{011})(1-\theta)}{D} \widehat{K}_{1,\theta} + \frac{(n_{111} + n_{011})(1-\theta)}{D} \widehat{K}_{0,\theta}.
\end{aligned}$$

where $\widehat{K}_{1,\theta}$ is the same as equation (3.6) but only calculated for those days when $Y_{i-1} = 1$. Similarly, $\widehat{K}_{0,\theta}$ is only calculated for those days when $Y_{i-1} = 0$.

We have that

$$\begin{aligned}
\frac{(n_{111} + n_{011})(1-\theta)}{D} &= \frac{(n_{111} + n_{011})(1-\theta)}{D} \frac{n(n_{111} + n_{011} + n_{101} + n_{001})}{n(n_{111} + n_{011} + n_{101} + n_{001})} \\
&= \frac{\widehat{p}_{1+|1}\widehat{p}}{\widehat{p}_y},
\end{aligned}$$

where $\widehat{p_y} = \widehat{P}(Y_i = 1)$ (note that $\widehat{p} = \widehat{P}(Y_{i-1} = 1)$ does not necessarily equal $\widehat{p}_y = \widehat{P}(Y_1 = 1)$ for any given sample) Similarly,

$$\frac{(n_{110} + n_{010})(1-\theta)}{D} = \frac{\widehat{p}_{1+|0}(1-\widehat{p})}{\widehat{p}_y}.$$

This results in

$$\widehat{K}_\theta = \frac{\widehat{p}_{1+|1}\widehat{p}}{\widehat{p}_y}\widehat{K}_{1,\theta} + \frac{\widehat{p}_{1+|0}(1-\widehat{p})}{\widehat{p}_y}\widehat{K}_{0,\theta}. \tag{3.7}$$

The contribution of each $\widehat{K}_{i,\theta}$ is weighted by the proportion of Y_i's=1 on those days when $Y_{i-1} = 1$ and $Y_{i-1} = 0$. Because $\widehat{p}_{1+|1}\widehat{p}/\widehat{p}_y = \widehat{P}(Y_{i-1} = 1|Y_i = 1)$, and $\widehat{p}_{1+|0}(1-\widehat{p})/\widehat{p}_y = \widehat{P}(Y_{i-1} = 0|Y_i = 1)$, we can also write (3.7) as

$$\widehat{K}_\theta = \widehat{P}(Y_{i-1} = 1|Y_i = 1)\widehat{K}_{1,\theta} + (1 - \widehat{P}(Y_{i-1} = 1|Y_i = 1))\widehat{K}_{0,\theta}. \tag{3.8}$$

This notation is similar to the idea of sensitivity and specificity.

4. Example

We first start with an example of a simple skill test. We collected probability of precipitation forecasts made for New York City (Central Park) from 16 November 2000 to 17 January 2001 (63 forecasts) for both Accuweather and the National Weather Service (NWS). Both Accuweather and the NWS made 1-day ahead forecasts, though only Accuweather attempted 14-day ahead forecasts. Accuweather presented its forecasts in the form of yes/no predictions, while the NWS issued probability forecasts. Figure 1 shows how the forecasts did.

The NWS did quite well, beating or closely matching Accuweather's performance for the 1-day ahead predictions. The figure shows that the NWS forecast would have value for most users (for many losses). Accuweather performed badly for its 14-day ahead predictions. In fact, any user, regardless of his loss function, would have done better to use the optimal naive prediction during this time.

We next plot the same data (for the 1-day ahead forecasts) but break it into days when $Y_{i-1} = 1$ and for $Y_{i-1} = 0$. The overall probability of precipitation is $\widehat{p}_y = 0.32$. Estimates of the transition parameters are, $\widehat{p}_{1+|1} = 0.42$ and $\widehat{p}_{1+|0} = 0.28$ (tests, due to the small sample size, do not show the Markov nature of this data as "significant", but it is still useful for illustration).

Both Accuweather and the NWS do better on days where $Y_{i-1} = 1$, and do worse on days when $Y_{i-1} = 0$. But graphical analysis is only part of the answer. We next give a fuller analysis of a larger data set.

Brooks et al. (1997) present two sets of 321 precipitation forecasts for Oklahoma City. Forecasts were from one-day to seven-days ahead but only the one-day ahead forecasts are considered here. There are two sources (anonymous forecasts taken from media outlets) which have produced forecasts for the same event. The forecasts were given as probability of precipitation.

We now check to see if the precipitation data for which the Brooks et al. forecasts were produced is Markov. Estimates of the transition parameters are, $\widehat{p}_{1+|1} = 0.27$ and $\widehat{p}_{1+|0} = 0.19$ (this also says that $\widehat{p}_{0+|1} = 0.73$ and $\widehat{p}_{0+|0} = 0.81$). The overall probability of precipitation is $\widehat{p}_y = 0.21$. This data is actually only weakly dependent in time (a test for independence between Y_i and Y_{i-1} gives $G^2 = 1.92$, p-value=0.17), however they will serve as a good illustration. The probability of

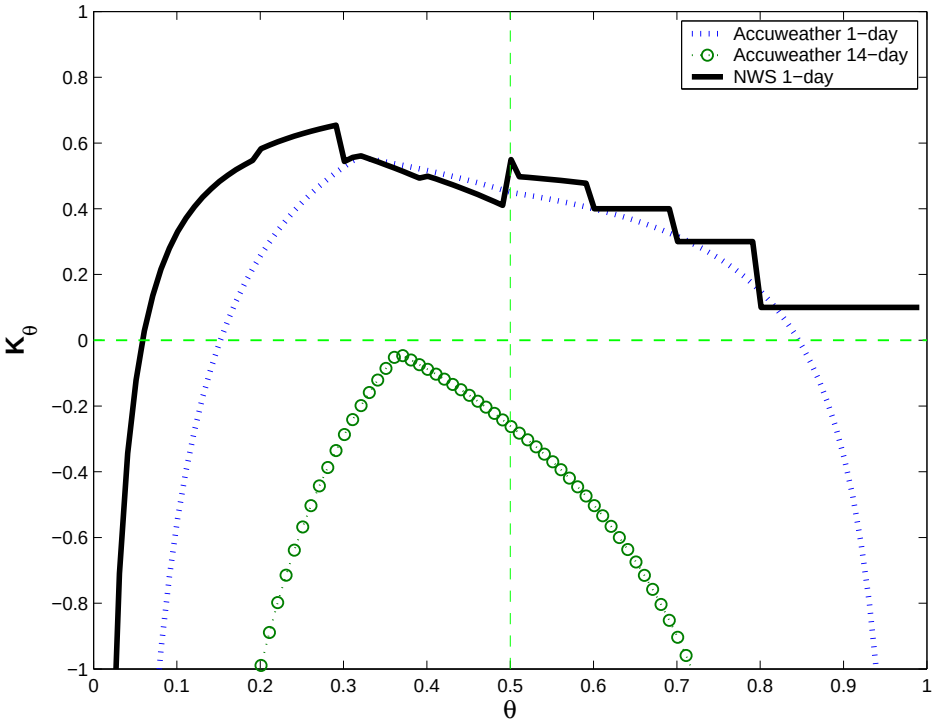

FIGURE 1. Skill score range plot for Accuweather's 1- and 14-day ahead and the NWS's 1-day forecasts. The dashed horizontal line shows 0 and predictions below this line have no skill.

TABLE 2. Skill statistics for Source A (SA) and Source B (SB). See the text for an explanation of the results.

Statistic	SA	SB
$\widehat{K}_{1/2}$	0.254	0.209
$G\ (p)$	14.1 (0.001)	6.34 (0.006)
$\widehat{K}_{1,1/2}$	0.333	0.111
$\widehat{K}_{0,1/2}$	0.225	0.245
$G_M\ (p)$	16.04 (0.0002)	8.04 (0.009)
$G_c\ (p)$	30.8 (< 0.0001)	27.98 (< 0.0001)

a dry day following either wet or dry is greater than the probability of a wet day. This is the situation we developed above with the optimal naive forecast always being 0, regardless of the value of Y_{i-1}. Thus, the optimal naive forecast is not the same as the persistence forecast.

Table 2 lists the relevant statistics. Shown first are $\widehat{K}_{1/2}$, the climate skill statistic developed in BR, the climate skill test statistic G and its p-value. Both sources evidence climate skill, although SA appears somewhat better with a higher skill score; a $\widehat{K}_{1/2} = 0.254$ for SA and a $\widehat{K}_{1/2} = 0.209$ for SB.

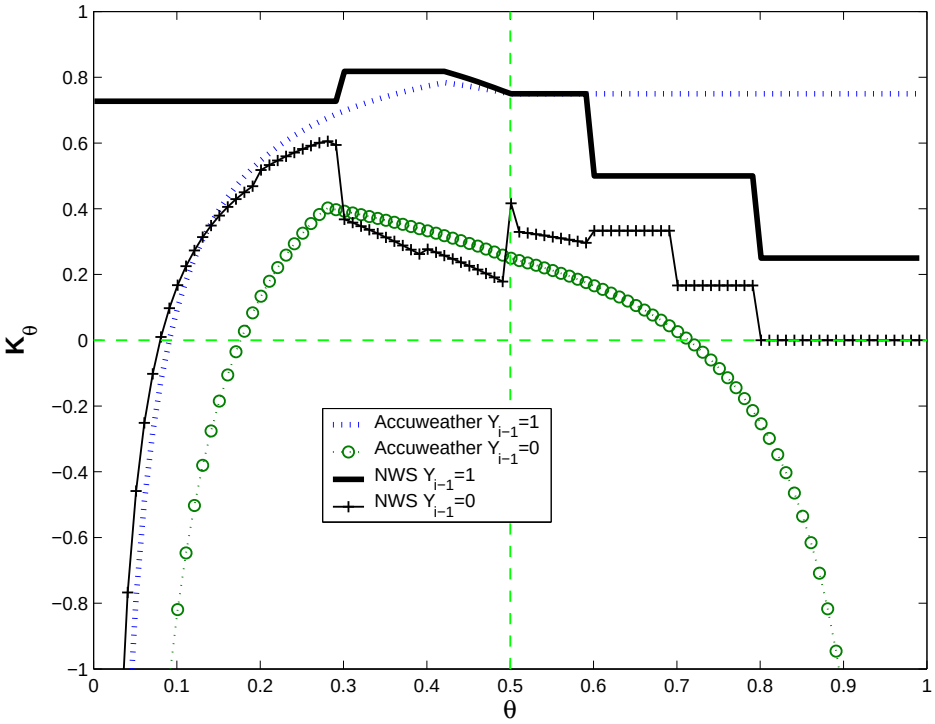

FIGURE 2. Skill score range plot for Accuweather's and NWS's 1-day ahead forecasts, split into days when $Y_{i-1} = 1$ and for $Y_{i-1} = 0$. The dashed horizontal line shows 0 and predictions below this line have no skill.

TABLE 3. Skill score weightings for the Brooks et al. data.

$Y_{i-1} = 0$	$Y_{i-1} = 1$
0.73	0.27

Next are the climate skill scores for those days on which $Y_{i-1} = 1$ ($\widehat{K}_1$) and for those days in which $Y_{i-1} = 0$ ($\widehat{K}_0$) (both at $\theta = 1/2$). We can see that SA's advantage has come from scoring better on those days which had $Y_{i-1} = 1$; a $\widehat{K}_{1,1/2} = 0.333$ at SA to a $\widehat{K}_{1,1/2} = 0.111$ at SB. Both Sources did about the same on those days which had $Y_{i-1} = 0$; a $\widehat{K}_{0,1/2} = 0.225$ at SA to a $\widehat{K}_{0,1/2} = 0.245$ at SB. Both Sources evidenced Markov skill; both sources had large G_Ms and small p-values for the test. The weighting (shown in Table 3) for the skill score $\widehat{K}_{1,1/2}$ was 0.27, and for $\widehat{K}_{0,1/2}$ it was 0.73, which shows that the days on which $Y_{i-1} = 0$ receive the majority of the weight and explains why SA and SB are still close in overall performance even though SA scores so well on days when $Y_{i-1} = 1$.

A test against a forecast of persistence was examined for both Sources with the nulls being that the Sources were no more accurate than was persistence. This test assigns the Source forecast as the first forecast and the Persistence forecast as the

second forecast. Recall, a persistence forecast is one in which $X_i = Y_{i-1}$ for all i. Both Sources had no trouble beating persistence; SA had a $G_c = 30.8$ and SB had $G_c = 27.98$, with p-values< 0.0001. For both Sources, the probability that the Source forecast and the Persistence forecast being correct was 0.65. SA and the Persistence forecast were both wrong 11% of the time, while SB and Persistence were both wrong 12% of the time. SA was correct and Persistence was wrong 19% of the time, while Persistence was correct and SA was wrong only 5% of the time. This large discrepancy accounts for why SA evidenced persistence skill. The results are nearly the same for SB, except that SB was correct and Persistence was wrong 18% of the time.

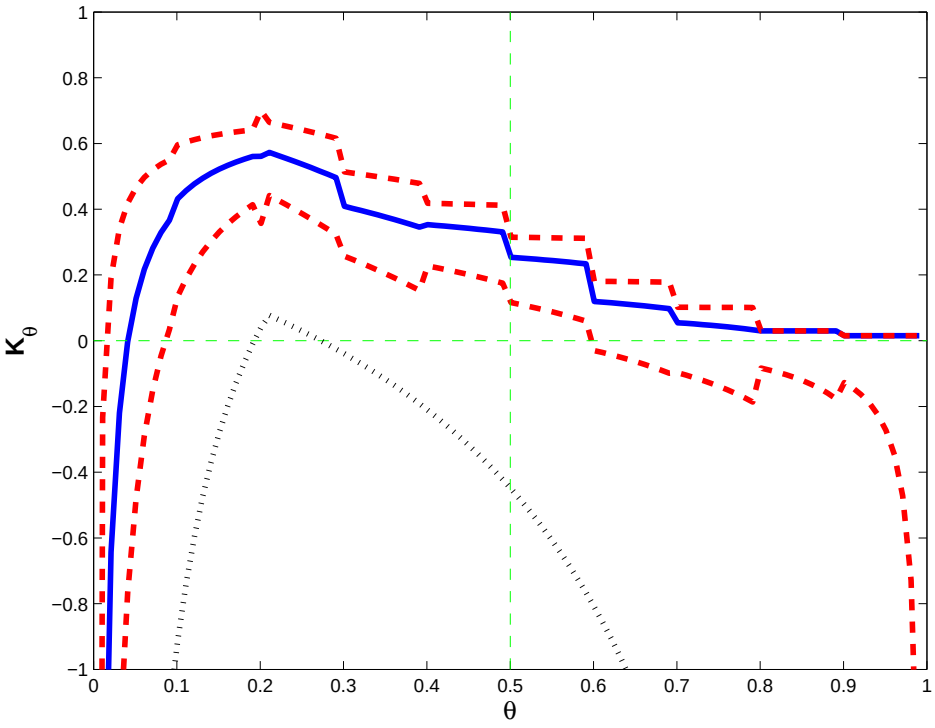

FIGURE 3. Skill score range plot for Accuweather's and NWS's 1-day ahead forecasts, split into days when $Y_{i-1} = 1$ and for $Y_{i-1} = 0$. The dashed horizontal line shows 0 and predictions below this line have no skill.

Finally, we perform an analysis of a persistence forecast. Figure 3 shows the skill plot for Source A and for a persistence forecast. The 95% point-wise confidence bound, created by inverting the test statistic G, for Source A's skill are also shown. Persistence does have some skill, but for a narrow range of θ; in particular, persistence does badly at $\theta = 1/2$. But this plot does highlight the fact that persistence forecasts are not the same as optimal naive forecasts (which would have $K_\theta = 0$ everywhere).

5. Conclusion

We have shown how to extend the basic skill testing framework developed in BR to events that are Markov. We have also proposed a method, which is an extension of well-known contingency table tests, to compare competing forecasts for the same event. We applied this method to test persistence forecasts, which are forecasts, in meteorological terms, that always say tomorrow will be like today. Skill of a persistence forecast is when a persistence forecast beats the optimal naive forecast with respect to the comparison test.

The comparison test, while useful, is not entirely satisfactory because it does not take into account the dependent nature of the observations. The test developed above does use the Markov nature of the observations. We also created a skill score to give a point measure of skill, which we showed reduced to the score given in BR. So we also showed how the score was a weighted sum of two parts, a skill score where the previous observation equalled zero, and a skill score where the previous observation equalled one. The weights were only functions of the observed observations series (not on the forecasts), that is, they were independent of the forecast process.

Scores, like those developed above, will be more useful when they can be applied to field forecasts. An example of such a forecast is a map of PoP forecasts. The skill score can, of course, be calculated for each point on a field and contours can be drawn to gauge performance (Drosdowsky and Zhang, 2003). But naively drawing skill maps won't take into account the dependent nature of observations and forecasts across space. New models are needed.

Appendix A. Climate Skill

This appendix contains material originally presented in BR and is given to orient the reader as all of the above theory takes what is below as given. For a fuller treatment, please see the original BR.

A.1. **Climate Skill Test.** BR defined skill, as above, in two ways. The first says that skill exists when the expected loss of the expert forecast is less than the expected loss of the optimal naive forecast. The second, related to accuracy, says that skill exists when the probability that the expert forecast is correct is larger than the probability that the optimal naive forecast is correct. BR then prove that these definitions are equivalent when the loss is symmetric.

We are concerned with events Y which are dichotomous. Predictions (which may be probabilistic) $\widetilde{X} \in [0, 1]$ are made for Y. Predictions can be either dichotomous or probabilistic but here we only consider decisions based on forecasts that are dichotomous. This implies a transformation of a probabilistic prediction into an eventual dichotomous one, that is, we act is if the event $Y = 1$ will occur, or we act as if the event $Y = 0$ will occur.

We follow the notation developed in Schervish (1989). Let $Y_i \in \{0, 1\}$ designate the ith observation of a dichotomous event, that is, $Y_i = 1$ if the event occurs and equals zero if it does not. Let the loss k_{YX} (k_{11} and k_{00}) associated with making a correct decision equal 0 (this condition is modified later). The finite loss k for making an error can always be quantified such that the total loss is normalized to 1, so that with $Y = 0$ and decision d_1 the loss can be written as some $k_{01} = \theta < 1$, which implies that with $Y = 1$ and decision d_0 the loss is $k_{10} = 1 - \theta$.

The decision maker minimizes his loss and makes decision d_1 whenever (the possibly probabilistic) expert forecast $\tilde{X} \geq \theta$ or makes decision d_0 if $\tilde{X} < \theta$, and where $\tilde{X}$ indicates a probability forecast. So, $X = I(\tilde{X} \geq \theta)$ reflects the (possible) probabilistic prediction transformed to a dichotomous one by the decision maker.

Thus, let $X_i \in \{0, 1\}$ designate the ith (possibly transformed) prediction. Assume that $\{(X_i, Y_i) : i = 1, \ldots, n\}$ is an i.i.d. sequence. Let $P(Y = 1|X = 1) = p_{1|1}$, $P(Y = 1|X = 0) = p_{1|0}$, $P(X = 1) = p_{+1}$, and $P(Y = 1) = p_{1+} = p$. We also assume that each observation Y_i is independent of each other Y_j for $i \neq j$ and that all of these probabilities are unvarying for all i. We also assume that $cov(Y_i, X_j) = 0$ for $i \neq j$, that is, the forecast observation process is not dynamic and that future observations do not depend on past forecasts. In practice, it could be that this covariance will be non-zero for Markov Y since X_i may depend (in unknown ways) on Y_{i-1}. Future work will explore more general models.

The expected loss for the optimal naive forecast depends both on p and on the value of θ. If $p \leq \theta$ the optimal naive forecast is $X^N = 0$, where the superscript N denotes the optimal naive forecast. If $p > \theta$ the optimal naive forecast is to always answer $X^N = 1$.

The null hypothesis for the skill test can now be formed. It is

$$H_0 : \quad E(k^E) \geq E(k^N) \tag{A.1}$$

where k^E corresponds to the loss of the expert prediction and k^N is the loss of the optimal naive prediction, and expectation is taken over both forecasts and observations.

Note that $p = P(Y = 1, X = 1) + P(Y = 1, X = 0)$. Substituting for the expected loss, and noting that $X^N \equiv 0$, (A.1) gives

$$\begin{aligned}
\theta P(Y = 0, X^E = 1) + (1 - \theta)P(Y = 1, X^E = 0) &\geq p(1 - \theta) \\
\theta P(Y = 0, X^E = 1) &\geq P(Y = 1, X^E = 1)(1 - \theta) \\
\theta &\geq \frac{P(Y = 1, X^E = 1)}{p_{+1}} \\
p_{1|1} &\leq \theta \qquad \text{(A.2)}
\end{aligned}$$

The alternative is that $p_{1|1} > \theta$.

BR prove that $H_0 : P(Y = X^E) \leq P(Y = X^N)$ is *equivalent* to (A.2) when the loss is symmetric, that is, when $\theta = 1/2$. This neatly ties together the views on skill defined in terms of loss and accuracy.

Let n_{YX} be the observed count of when $Y = y$ and $X = x$. The unrestricted maximum likelihood estimates (MLEs) of the model are easily found as each parameter separates in the likelihood:

$$\begin{aligned}
\widehat{p}_{+1} &= \frac{n_{11} + n_{01}}{n_{++}} \\
\widehat{p}_{1|1} &= \frac{n_{11}}{n_{11} + n_{01}} \\
\widehat{p}_{1|0} &= \frac{n_{10}}{n_{10} + n_{00}},
\end{aligned}$$

where $n_{++} = n_{11} + n_{10} + n_{01} + n_{00}$. Under the null the MLE for p_{+1} and the estimate for $p_{1|0}$ remain unchanged as might be expected. The null is that $p_{1|1} \leq \theta$, with an estimate maximizing the likelihood of $\widetilde{p}_{1|1} = \min\{\frac{n_{11}}{n_{11}+n_{01}}, \theta\}$. These facts

makes calculation of the likelihood ratio statistic (LRS) particularly simple as the terms involving p_{+1} and $p_{1|0}$ drop out, leaving only the terms involving $p_{1|1}$.

The LRS, G, is

$$G = 2n_{11} \log \left[\frac{\widehat{p}_{1|1}}{\widetilde{p}_{1|1}} \right] + 2n_{01} \log \left[\frac{1-\widehat{p}_{1|1}}{1-\widetilde{p}_{1|1}} \right].$$

There are two situations under the null: when $\frac{n_{11}}{n_{+1}} > \theta$ then $\widetilde{p}_{1|1} = \theta$ and $G > 0$, and when $\frac{n_{11}}{n_{+1}} \leq \theta$ then $\widetilde{p}_{1|1} = \widehat{p}_{1|1}$ and $G = 0$. This allows us to rewrite G as

$$G = \left(2n_{11} \log \left[\frac{n_{11}}{n_{+1}\theta} \right] + 2n_{01} \log \left[\frac{n_{01}}{n_{+1}(1-\theta)} \right] \right) I\left(\frac{n_{11}}{n_{+1}} > \theta \right) \tag{A.3}$$

where $n_{+1} = n_{11} + n_{01}$. As a practical matter, when making calculations with real data the often-used definition $0 \log(0) = \lim_{x \downarrow 0} x \log(x) = 0$ is invoked.

G has an asymptotic distribution which is related to the χ^2 distribution with one degree of freedom. Since the test is one-sided the actual distribution is $1/2\chi^2_0 + 1/2\chi^2_1$ (Self and Liang, 1987; their case 5). Tests are carried out similar to a standard χ^2_1 test, except that where a normal χ^2_1 statistic W has that $P(W > w) = \alpha$; here, because there is a probability mass of $1/2$ at 0, the χ^2_1 statistic G has that $P(G > w) = \alpha/2$. In practice, the user only has to double his chosen test level and use an ordinary χ^2_1 distribution. Equivalently, one can half the p-value.

It is easy to show, because of the symmetry of the problem, that the null hypothesis when $p > \theta$ is $H_0 : p_{0|0} \geq 1 - \theta$. The complete null combines this with an indicator $I(p > \theta)$ with the null in A.2 with an indicator $I(p \leq \theta)$. The test statistic G in (A.3) for $p > \theta$ is

$$G = \left(2n_{00} \log \left[\frac{n_{00}}{n_{+0}\theta} \right] + 2n_{10} \log \left[\frac{n_{10}}{n_{+0}(1-\theta)} \right] \right) I\left(\frac{n_{00}}{n_{+0}} > \theta \right) \tag{A.4}$$

A.2. **Climate skill score.** Testing the significance of a skill score is the same as the climate skill test if the following skill score is taken

$$K = K(y, x^E) = \frac{E(k^N) - E(k^E)}{E(k^N)}, \tag{A.5}$$

where the expected forecast loss is taken as the error score. A collection of perfect expert forecasts will have a loss of 0, so, for us, a perfect skill score will be $K \equiv 1$. A collection with "negative" skill, as defined in (A.1), will have either an expected loss the same as the naive forecasts or even greater so that the skill score will be 0 or less. The null hypothesis is

$$H_0 : \quad K \leq 0. \tag{A.6}$$

It can be easily seen that this translates exactly to the hypothesis and test used before, defined in (A.2).

An estimate for the skill score is

$$\begin{aligned} \widehat{K}_\theta &= \frac{\widehat{p}(1-\theta) - \theta(1-\widehat{p}_{1|1})\widehat{p}_{+1} - (1-\theta)\widehat{p}_{1|0}(1-\widehat{p}_{+1})}{\widehat{p}(1-\theta)} \\ &= \frac{(\widehat{p}_{1|1} - \theta)\widehat{p}_{+1}}{\widehat{p}(1-\theta)} \\ &= \frac{n_{11}(1-\theta) - n_{01}\theta}{(n_{11}+n_{10})(1-\theta)}. \end{aligned} \tag{A.7}$$

For general verification purposes a plausible loss is symmetric loss, that is $\theta = 1/2$. Symmetric loss is further justified below. Symmetric loss gives

$$\widehat{K}_{1/2} = \frac{n_{11} - n_{01}}{n_{11} + n_{10}}. \tag{A.8}$$

This has a particularly simple form which shows easily whether forecasts have skill: this is when $n_{11} > n_{01}$, which makes $\widehat{K}_{1/2} > 0$. Our score for symmetric loss is also similar in form to other skill scores which are summarized in, among other places, Wilks (1995).

Let $I_0 = I(p \leq \theta)$ and $I_1 = 1 - I_0$. Finally, and in full form, the estimate of the skill score is

$$\widehat{K}_\theta = \frac{n_{11}(1-\theta) - n_{01}\theta}{(n_{11} + n_{10})(1-\theta)} I_0 + \frac{n_{00}\theta - n_{10}(1-\theta)}{(n_{00} + n_{01})\theta} I_1. \tag{A.9}$$

A.2.1. *Brier score.* The most popular score is the Brier score, which is given as

$$B = (Y - X)^2.$$

Our climate skill score and the Brier score have an interesting relationship.

BR prove that (1) Testing for skill using the Brier score, where skill is defined as when a collection of expert predictions have a lower Brier score than the Brier score for the optimal naive predictions, is equivalent to the climate skill test with symmetric loss, and (2) A collection of predictions has skill (with symmetric loss) when $\widehat{B} < \widehat{p}$, where $\widehat{B} = \sum (Y_i - X_i)^2$. Further, $\widehat{B} = \widehat{p}(1 - \widehat{K}_{1/2})$. A collection of forecasts has skill (with symmetric loss) when $\widehat{K}_{1/2} > 0$, so that skillful forecasts have $\widehat{B} < \widehat{p}$.

A.2.2. *Loss for perfect forecasts.* It is possible to add loss for making correct predictions (losses which we assume are less than the losses for making an incorrect prediction). Let k_{11} and k_{00} be the losses for making correct predictions with the minimal requirement that $k_{00} < k_{01}$ and $k_{11} < k_{10}$. It's easy to show that, using the definition of skill that a collection of expert predictions has less expected loss than a collection of optimal naive predictions, that the null originally given in (A.2) is modified to

$$p_{1|1} \leq \frac{k_{01} - k_{00}}{k_{01} - k_{00} + k_{10} - k_{11}} = \theta'. \tag{A.10}$$

Calculation of the test statistic and so on goes on as before. An example of this is the Value Score (VS) proposed in Wilks (2001) which was developed in the cost-loss scenario. Wilks has that a perfect prediction has either a loss $k_{11} = k_{01}$ or a loss $k_{00} = 0$. Loss for imperfect prediction is the same as before. In these situations it only makes sense to talk of losses where $k_{01} < k_{10}$; see Wilks (2001) for a complete explanation. We have $VS = (E(k^N) - E(k^E))/(E(k^P) - E(k^N))$ where $E(k^P)$ is the expected loss for a perfect prediction. This gives an estimate of

$$\widehat{VS}_\theta = \frac{n_{11}(1-2\theta) - n_{01}\theta}{(n_{11} + n_{10})(1-2\theta)}.$$

which is nearly the same as K_θ. This is highlighted in the VS hypothesis test for skill, which in this case is

$$H_0 : \quad p_{1|1} \leq \frac{\theta}{1-\theta}.$$

TABLE 4. The four separate cases where different optimal naive forecasts are implied. The conditions are set in the Prob. columns, with the optimal naive forecasts listed.

Case	Prob.	Optimal Naive	Prob.	Optimal Naive
1	$p_{0+\|1} \leq 1-\theta$	1	$p_{0+\|0} \leq 1-\theta$	1
2	$p_{0+\|1} \leq 1-\theta$	1	$p_{1+\|0} \leq \theta$	0
3	$p_{1+\|1} \leq \theta$	0	$p_{0+\|0} \leq 1-\theta$	1
4	$p_{1+\|1} \leq \theta$	0	$p_{1+\|1} \leq \theta$	0

For small θ, which is likely in the cost-loss problem, $\theta \approx \theta/(1-\theta)$, and for larger $\theta < 1/2$ the skill test based on the Value Score is more conservative because $\theta/(1-\theta) > \theta$.

A.2.3. *Skill and dependence.* BR also show that skill is stronger than the condition of dependence (the usual tests for 2×2 tables), and that skill implies dependence.

APPENDIX B. MARKOV DETAILS

There are four cases to capture all the possibilities when $\{Y_i\}$ is Markov. These correspond to the probabilities p_{ij} which, depending on their values, represent different optimal naive forecasts.

We developed Case 4 earlier. These four cases imply four separate null hypotheses. These are

Case (1)

$$\begin{aligned} H_{0,1}: \quad & P(Y_i = X_i | Y_{i-1} = 1) \leq P(Y_i = 1 | Y_{i-1} = 1), \\ & P(Y_i = X_i | Y_{i-1} = 0) \leq P(Y_i = 1 | Y_{i-1} = 0). \end{aligned}$$

Case (2)

$$\begin{aligned} H_{0,2}: \quad & P(Y_i = X_i | Y_{i-1} = 1) \leq P(Y_i = 1 | Y_{i-1} = 1), \\ & P(Y_i = X_i | Y_{i-1} = 0) \leq P(Y_i = 0 | Y_{i-1} = 0). \end{aligned}$$

Case (3)

$$\begin{aligned} H_{0,3}: \quad & P(Y_i = X_i | Y_{i-1} = 1) \leq P(Y_i = 0 | Y_{i-1} = 1), \\ & P(Y_i = X_i | Y_{i-1} = 0) \leq P(Y_i = 1 | Y_{i-1} = 0). \end{aligned}$$

Case (4)

$$\begin{aligned} H_{0,4}: \quad & P(Y_i = X_i | Y_{i-1} = 1) \leq P(Y_i = 0 | Y_{i-1} = 1), \\ & P(Y_i = X_i | Y_{i-1} = 0) \leq P(Y_i = 0 | Y_{i-1} = 0). \end{aligned}$$

Or, incorporating the possibility of asymmetric loss,

$$\begin{aligned} H_{0,1}: \quad & (p_{0|01} \leq 1-\theta, p_{0|00} \leq 1-\theta) \\ H_{0,2}: \quad & (p_{0|01} \leq 1-\theta, p_{1|10} \leq \theta) \\ H_{0,3}: \quad & (p_{1|11} \leq \theta, p_{0|00} \leq 1-\theta) \\ H_{0,4}: \quad & (p_{1|11} \leq \theta, p_{1|10} \leq \theta). \end{aligned}$$

Likelihood ratio statistics are found in the same manner as before. The results are:

Case (1)

$$G_{1M} = 2n_{101}\log\left(\frac{\widehat{p}_{1|01}}{\widetilde{p}_{1|01}}\right) + 2n_{001}\log\left(\frac{1-\widehat{p}_{1|01}}{1-\widetilde{p}_{1|01}}\right) + 2n_{100}\log\left(\frac{\widehat{p}_{1|00}}{\widetilde{p}_{1|00}}\right) + 2n_{000}\log\left(\frac{1-\widehat{p}_{1|00}}{1-\widetilde{p}_{1|00}}\right).$$

Case (2)

$$G_{2M} = 2n_{101}\log\left(\frac{\widehat{p}_{1|01}}{\widetilde{p}_{1|01}}\right) + 2n_{001}\log\left(\frac{1-\widehat{p}_{1|01}}{1-\widetilde{p}_{1|01}}\right) + 2n_{110}\log\left(\frac{\widehat{p}_{1|10}}{\widetilde{p}_{1|10}}\right) + 2n_{010}\log\left(\frac{1-\widehat{p}_{1|10}}{1-\widetilde{p}_{1|10}}\right).$$

Case (3)

$$G_{3M} = 2n_{111}\log\left(\frac{\widehat{p}_{1|11}}{\widetilde{p}_{1|11}}\right) + 2n_{011}\log\left(\frac{1-\widehat{p}_{1|11}}{1-\widetilde{p}_{1|11}}\right) + 2n_{100}\log\left(\frac{\widehat{p}_{1|00}}{\widetilde{p}_{1|00}}\right) + 2n_{000}\log\left(\frac{1-\widehat{p}_{1|00}}{1-\widetilde{p}_{1|00}}\right).$$

Case (4)

$$G_{4M} = 2n_{111}\log\left(\frac{\widehat{p}_{1|11}}{\widetilde{p}_{1|11}}\right) + 2n_{011}\log\left(\frac{1-\widehat{p}_{1|11}}{1-\widetilde{p}_{1|11}}\right) + 2n_{110}\log\left(\frac{\widehat{p}_{1|10}}{\widetilde{p}_{1|10}}\right) + 2n_{010}\log\left(\frac{1-\widehat{p}_{1|10}}{1-\widetilde{p}_{1|10}}\right).$$

A slightly different notation will be needed to keep track of the different skill scores for the different cases. Let $K_{ij,\theta}$ be the climate skill score for optimal naive forecast i when the day before $Y_{-1} = j$. For example, in Case 4, the climate skill score estimate is now

$$\widehat{K}_{4,\theta} = \frac{\widehat{p}_{1+|1}\widehat{p}}{\widehat{p}_y}\widehat{K}_{01,\theta} + \frac{\widehat{p}_{1+|0}(1-\widehat{p})}{\widehat{p}_y}\widehat{K}_{00,\theta},$$

where $\widehat{K}_{01,\theta}$ is the climate skill score for those days in which $Y_{-1} = 1$ and the optimal naive forecast is 0, and $\widehat{K}_{00,\theta}$ is the climate skill score for those days in which $Y_{-1} = 0$ and the optimal naive forecast is 0. To be complete,

$$\widehat{K}_{0j,\theta} = \frac{n_{11j}(1-\theta) - n_{01j}\theta}{(n_{11j} + n_{10j})(1-\theta)},$$

and

$$\widehat{K}_{1j,\theta} = \frac{n_{00j}\theta - n_{10j}(1-\theta)}{(n_{00j} + n_{01j})\theta}.$$

Skill scores are slightly more complicated, except in Case 1 and Case 4 (which was derived earlier). Case 1 is similar to Case 4 because no matter the value of Y_{i-1} the optimal naive forecast is always 1 in Case 4 the optimal naive forecast is always 0). Because of this, the skill score for Case 1 is easy:

$$\widehat{K}_{1,\theta} = \frac{\widehat{p}_{0+|1}\widehat{p}}{1-\widehat{p}_y}\widehat{K}_{11,\theta} + \frac{\widehat{p}_{0+|0}(1-\widehat{p})}{1-\widehat{p}_y}\widehat{K}_{10,\theta},$$

Cases 2 and 3 are more difficult, but related. Focus on Case 3, where the optimal naive forecast on day i is 0 on those days when $Y_{i-1} = 1$ and is 1 on those days when $Y_{i-1} = 0$. The expected loss for the optimal naive forecasts is

$$p(1-\theta)(p_{1|11}p_{+1|1}+p_{1|01}(1-p_{+1|1}))+(1-p)\theta((1-p_{1|10})p_{+1|0}+(1-p_{1|00})(1-p_{+1|0})).$$

Substituting the estimates of these parameters gives

$$D = (1/n)((1-\theta)(n_{111}+n_{101})+\theta(n_{010}+n_{000})).$$

The expected loss of the optimal naive forecast minus the expected loss of the expert forecasts is

$$pp_{+1|1}(p_{1|11}-\theta)+(1-p)(1-p_{+1|0})(\theta-p_{1|00}).$$

After substituting the expected values we get

$$(1/n)(n_{111}(1-\theta)-n_{011}\theta+n_{000}\theta-n_{100}(1-\theta)).$$

We now arrive the estimate for $K_{3,\theta}$

$$\begin{aligned}\widehat{K}_{3,\theta} &= \frac{(n_{111}+n_{101})(1-\theta)}{(n_{111}+n_{101})(1-\theta)}\frac{n_{111}(1-\theta)-n_{011}\theta}{D}+\\ &\quad \frac{(n_{111}+n_{101})(1-\theta)}{(n_{111}+n_{101})(1-\theta)}\frac{n_{000}\theta-n_{100}(1-\theta)}{D}\\ &= \frac{(n_{111}+n_{101})(1-\theta)}{D}\widehat{K}_{11,\theta}+\frac{(n_{111}+n_{101})(1-\theta)}{D}\widehat{K}_{10,\theta}.\end{aligned}$$

Now,

$$\begin{aligned}\frac{(n_{111}+n_{101})(1-\theta)}{D} &= \frac{n_{111}+n_{011}+n_{101}+n_{001}}{n_{111}+n_{011}+n_{101}+n_{001}}\frac{(n_{111}+n_{101})(1-\theta)}{D}\\ &= (1-\theta)p_{1+|1}\frac{n_{111}+n_{011}+n_{101}+n_{001}}{D}.\end{aligned}$$

Further,

$$\begin{aligned}\frac{D}{n_{111}+n_{011}+n_{101}+n_{001}} &= \frac{(1-\theta)(n_{111}+n_{101})}{n_{111}+n_{011}+n_{101}+n_{001}}+\\ &\quad \frac{(1-\theta)(n_{010}+n_{000})}{n_{111}+n_{011}+n_{101}+n_{001}}\\ &= (1-\theta)\widehat{p}_{1+|1}+\theta\widehat{p}_{0+|0}\frac{1-\widehat{p}}{\widehat{p}}.\end{aligned}$$

So,

$$\frac{(n_{111}+n_{101})(1-\theta)}{D} = \frac{(1-\theta)p_{1+|1}}{(1-\theta)\widehat{p}_{1+|1}+\theta\widehat{p}_{0+|0}\frac{1-\widehat{p}}{\widehat{p}}}.$$

This can also be written

$$\frac{(n_{111}+n_{101})(1-\theta)}{D} = \frac{(1-\theta)\widehat{P}(Y_i=Y_{i-1}=1)}{(1-\theta)\widehat{P}(Y_i=Y_{i-1}=1)+\theta\widehat{P}(Y_i=Y_{i-1}=0)}.$$

Similarly,

$$\frac{(n_{111}+n_{101})(1-\theta)}{D} = \frac{\theta p_{0+|0}}{\theta\widehat{p}_{0+|0}+(1-\theta)\widehat{p}_{1+|1}\frac{\widehat{p}}{1-\widehat{p}}}.$$

Which is also

$$\frac{(n_{111}+n_{101})(1-\theta)}{D} = \frac{\theta\widehat{P}(Y_i = Y_{i-1} = 0)}{\theta\widehat{P}(Y_i = Y_{i-1} = 0) + (1-\theta)\widehat{P}(Y_i = Y_{i-1} = 1)}.$$

This finally gives

$$\widehat{K}_{3,\theta} = \frac{(1-\theta)p_{1+|1}}{(1-\theta)\widehat{p}_{1+|1} + \theta\widehat{p}_{0+|0}\frac{1-\widehat{p}}{\widehat{p}}}\widehat{K}_{01,\theta} + \frac{\theta p_{0+|0}}{\theta\widehat{p}_{0+|0} + (1-\theta)\widehat{p}_{1+|1}\frac{\widehat{p}}{1-\widehat{p}}}\widehat{K}_{10,\theta}.$$

A similar argument leads to the estimate of $K_{2,\theta}$

$$\widehat{K}_{2,\theta} = \frac{\theta p_{0+|1}}{\theta\widehat{p}_{0+|1} + (1-\theta)\widehat{p}_{1+|0}\frac{\widehat{p}}{1-\widehat{p}}}\widehat{K}_{11,\theta} + \frac{(1-\theta)p_{1+|0}}{(1-\theta)\widehat{p}_{1+|0} + \theta\widehat{p}_{0+|1}\frac{\widehat{p}}{1-\widehat{p}}}\widehat{K}_{00,\theta}.$$

References

1. Briggs, W.M., and R.A. Levine, 1998. Comparison of forecasts using the bootstrap. *14th Conf. on Probability and Statistics in the Atmospheric Sciences*, Phoenix, AZ, Amer. Meteor. Soc., 1-4.
2. Briggs, W.M., and D. Ruppert, 2003. Assessing the skill of yes/no predictions. Submitted to *Biometrics*
3. Brooks, H. E., A. Witt, and M. D. Eilts, 1997. Verification of public weather forecasts available via the media. *Bull. Amer. Meteor. Soc.*, **77**, 2167-2177.
4. Drosdowsky, W., and H. Zhang, 2003. Verification of spatial fields. In *Forecast Verification*, Jolliffe, I.T., and D.B. Stephenson, eds. Wiley, New York, 121-136.
5. Livezey, R.E., 2003. Categorical events. In *Forecast Verification*, Jolliffe, I.T., and D.B. Stephenson, eds. Wiley, New York, 77-96.
6. Mason, I.B., 2003. Binary events. In *Forecast Verification*, Jolliffe, I.T., and D.B. Stephenson, eds. Wiley, New York, 37-76.
7. Meeden, G., 1979. Comparing two probability appraisers. *JASA*, **74**, 299-302.
8. Mosteller, F, 1952. Some statistical problems in measuring the subjective response to drugs. *Biometrics*, 220-226.
9. Mozer, J.B., and Briggs, W.M., 2003. Skill in real-time solar wind shock forecasts. *J. Geophysical Research: Space Physics*, **108 (A6)**, SSH 9 p. 1-9, (DOI 10.1029/2003JA009827).
10. Murphy, A.H., 1991. Forecast verification: its complexity and dimensionality. *Monthly Weather Review*, **119**, 1590-1601.
11. Murphy, A.H., 1997. Forecast verification. In *Economic Value of Weather and Climate Forecasts*. Katz, R.W., and A.H. Murphy (eds.). Cambridge, London, 19-74.
12. Murphy, A.H., and A. Ehrendorfer, 1987. One the relationship between the accuracy and value of forecasts in the cost-loss ratio situation. *Weather and Forecasting*, **2**, 243-251.
13. Murphy, A.H., and R. L. Winkler, 1987. A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330-1338.
14. Schervish, M.J., 1989. A general method for comparing probability assessors. *Annals of Statistics*, **17**, 1856-1879.
15. Self, S.G., and K.Y. Liang, 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. American Statistical Association*, **82**, 605-610.
16. Wilks, D.S., 1991. Representing serial correlation of meteorological events and forecasts in dynamic decision-analytic models. *Monthly Weather Review*, **119**, 1640-1662.
17. Wilks, D.S., 1995. *Statistical Methods in the Atmospheric Sciences*, Academic Press, New York. 467 pp.
18. Wilks, D.S., 2001. A skill score based on economic value for probability forecasts. *Meteorological Applications*, **8**, 209-219.

(W. Briggs) General Internal Medicine, Weill Cornell Medical College, 525 E. 68th St. Box # 45, New York, NY 10021
E-mail address: wib2004@med.cornell.edu

(D. Ruppert) School of OR & IE, Cornell University, 225 Rhodes Hall, Ithaca, NY 14859
E-mail address: dr24@cornell.edu