Assessing the skill of yes/no forecasts for Markov observations

William Briggs

General Internal Medicine, Weill Cornell Medical College
525 E. 68th, Box 46, New York, NY 10021
email: wib2004@med.cornell.edu

and

David Ruppert

School of Operations Research & Industrial Engineering,
Rhodes Hall, Cornell University, Ithaca, NY 14853
email: dr24@cornell.edu

April 28, 2005

Summary: Briggs and Ruppert (2005) recently introduced a new, easy-to-calculate economic skill/value score for use in yes/no forecast decisions, of which precipitation forecast decisions are an example. The advantage of this new skill/value 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/value score is significant or not.

Here, we take the climate skill/value score and extend it to the case where the predicted series is first-order Markov in nature, of which, again, precipitation occurrence series can be an example. We show that, in general, Markov skill/value 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 show that any naive forecasts based solely on the Markov parameters is always at least as valuable/skillful than are persistence forecasts; in general, persistence forecasts should not be used.

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 previous papers, Briggs and Ruppert (2004) and Briggs and Ruppert (2005; from here, BR), developed a statistical 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 events are 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 Levine (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.

We define forecasts $\widetilde{X} \in [0,1]$ made for events $Y \in \{0,1\}$. 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} if 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 such that $\theta = k_{01}/(k_{01} + k_{01})$. Here, we assume $k_{11} = k_{00} = 0$; Briggs (2005) showed how to modify this so that any $k_{YX} \geq 0$, which includes the class of cost-loss problems (see Wilks, 2001; Richardson, 2000, 2001). Briggs et al. (2005) extended the climate skill test to cases where the observed series is possibly classified with error.

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 climate forecast. The optimal naive climate forecast X^N_{cl} for Y is the forecast one would make knowing only $p = P(Y = 1)$. It is easy to show that this is $X^N_{\text{cl}} = I(p > \theta)$.

Skill is now defined. This is when $P(X^E = Y) > P(X^N_{\text{cl}} = Y)$. *Value* is when the expected loss of the expert forecast is less than the expected loss of the optimal naive climate forecast: $E(k^E) < E(k^N_{\text{cl}})$. BR showed that these two definitions are identical when $\theta = 1/2$, or when the loss is symmetric. BR developed a skill/value score and a test statistic for skill/value, 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; we further show that the expected loss of a persistence forecast is *always* as great or greater than the expected loss of optimal naive forecasts; thus, persistence forecasts should never be used. This result holds for optimal naive climate or Markov forecasts. 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 Markov forecast. Finally, an Appendix is given to detail the mathematical results.

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 (expert) forecasts X_1 and X_2. Define $Z_i = I(Y = X_i)$, which is the indicator that forecast i is correct, $i = 1, 2$. 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 probabilities designating those times when one forecast was correct while the other was wrong. We assume that the loss is symmetric, i.e., the loss for one forecast being correct is the same as for the other being incorrect.

The development in this Section is not new. The test statistic developed here is similar to McNemar's (1947) test for matched pairs and to its refinement by Mosteller (1952). Depending on the particular null hypothesis chosen, our statistic is only slightly different to the classical statistic and, again depending on the null, one could use the classical 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 $Z_{ji} = I(X_{ji} = Y_i)$, $j = 1, 2$. Let $m_{i,j} = \sum Z_{1,i} = Z_{2,j}$, that is, the observed counts.

One possible null hypothesis is

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

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 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 of the model is

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

where $z_{ij} = P(Z_1 = i, Z_2 = j)$, $i, j = 1, 2$. Under the null (2.1) the estimates are $\widehat{P}(Z_1 = i, Z_2 = i) = m_{ii}/m_{++}$, and $\widehat{P}(Z_1 = i, Z_2 = 1 - i) = (m_{10} + m_{01})/2m_{++}$, $i = 0, 1$. The likelihood ratio statistic G_c is computed easily; the terms involving z_{11} and z_{00} drop out, leaving

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

As is well known, the distribution of G_c, assuming the two-sided null (2.1), has an asymptotic χ^2 distribution with one degree of freedom (see for example Agresti, 1990).

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. When equality holds in (2.3), which is case used to compute p-values, then G_c has an asymptotic $1/2\chi^2_0 + 1/2\chi^2_1$ distribution (Self and Liang, 1987; BR, Section 2; see also the Appendix).

The classical test statistic for Z_1 and Z_2 having the same distributions (null (2.1)) is $G_C = (|m_{01} - m_{10}| - 1)^2/(m_{01} + m_{10})$ which also has a χ^2_1 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 climate forecast is the second (the null hypothesis is, of course, (2.3)), then G_c is the same as BR's G.

3. Markov 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 Markov forecast X^N_{Ma} of each Y_i is based on the previous observation Y_{i-1}. We will show that Markov skill is generally *not* identical with persistence skill. A persistence forecast is one in which the naive forecast for Y_i is Y_{i-1} for all i, and persistence skill means outperforming the persistence forecast. 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 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 ignores the data as a whole and the distribution of the test statistic under the Markov assumption. The following test is needed.

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. This factorization gives:

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

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 will use the convention that the replacement of an index by "+" means summation over that index so, for example, $p_{+1|1} = \sum_i p_{i1|1}$.

We shall also need to define the parameters that characterize the Markov nature of Y. These are $p_{1+|1} = P(Y_i = 1|Y_{i-1} = 1)$, $p_{0+|1} = P(Y_i = 0|Y_{i-1} = 1)$, $p_{1+|0} = P(Y_i = 1|Y_{i-1} = 0)$, $p_{0+|0} = P(Y_i = 0|Y_{i-1} = 0)$. 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 and persistence skill tests.** We now introduce tests of skill relative to optimal naive Markov forecasts and to persistence forecasts.

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).

The unrestricted MLEs are $\widehat{p} = n_{++1}/n_{+++}$, $\widehat{p}_{+1|1} = n_{+11}/n_{++1}$, $\widehat{p}_{+1|0} = n_{+10}/n_{++0}$. The other estimators do change when one switches between the null and alternative, and the unrestricted MLES are: $\widehat{p}_{1|11} = n_{111}/n_{+11}$, $\widehat{p}_{1|10} = n_{110}/n_{+10}$, $\widehat{p}_{1|01} = n_{101}/n_{+01}$ $\widehat{p}_{1|00} = n_{100}/n_{+00}$.

The optimal naive Markov forecast X^N_{Ma} must now be defined; "naive" means that only the transition probabilities $P(Y_i|Y_{i-1})$ are known. It turns out that there are four situations, that is, four circumstances that dictate different optimal naive Markov forecasts. We focus here on just one situation, detailed next, for the sake of an example. The other three situations will be removed to the Appendix. Table 1 lists the four cases of optimal naive Markov forecasts.

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 Markov forecast is always 0, regardless of the value of Y_{i-1}. Note that in this case the optimal naive Markov forecast is different than a persistence forecast, which is $X_i = Y_{i-1}$ for all i. Table 1 shows that in only one case is the optimal naive Markov forecast the same as the persistence forecast.

One solution for deriving a test of climate skill against persistence 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. But we can go further and show that the optimal naive Markov forecast is *always* at least as good as the persistence forecast (in terms of value or skill); this is done in the next Section.

Directly from BR, we have that the null hypothesis of no Markov skill is that the expected loss of the expert forecast is less than or equal to the expected loss of the optimal naive Markov forecast (details are in the Appendix). This gives:

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

All parameters except those indicated in the null hypothesis have the same MLEs in both the null and alternate hypotheses. The LRS (likelihood ratio 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

$$\begin{aligned} G_M = &\left(2n_{1|1} \log\left[\frac{n_{111}}{n_{+11}\theta}\right] + 2n_{011} \log\left[\frac{n_{011}}{n_{+11}(1-\theta)}\right]\right) I\left(\frac{n_{111}}{n_{+11}} > \theta\right) + \\ &\left(2n_{110} \log\left[\frac{n_{110}}{n_{+10}\theta}\right] + 2n_{010} \log\left[\frac{n_{010}}{n_{+10}(1-\theta)}\right]\right) I\left(\frac{n_{110}}{n_{+10}} > \theta\right) \end{aligned} \tag{3.2}$$

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. **The optimal Markov forecast is superior to persistence.** We now prove that the optimal naive Markov forecast is always at least as valuable (skilful) as any persistence forecast.

Theorem 3.1. *Let the optimal naive Markov forecast be denoted as $I\{P(Y_i = 1|Y_{i-1}) > \theta\}$ and let the persistence forecast be Y_{i-i}. Then the optimal naive Markov forecast is always at least as valuable (or skillful) than is the persistence forecast in the sense that $E(k_{Ma}^N) \leq E(k_P^N)$ where $E(k_{Ma}^N)$ is the expected loss of the optimal naive Markov forecast and $E(k_P^N)$ is the expected loss of the persistence forecast.*

Proof. There are four cases of optimal naive Markov forecasts; see Table 1. The optimal naive Markov and persistence forecasts are identical for Case 2, so $E(k_P^N) = E(k_{Ma}^N)$.

For Case 1, $I\{P(Y_i = 1|Y_{i-1}) > \theta\} = 0$ regardless of the value of Y_{i-1}. The expected loss for the optimal naive Markov forecast in Case 1 is

$$E(k_{Ma}^N) = (p_{11|0} + p_{10|0})(1-\theta)(1-p) + (p_{11|1} + p_{10|1})(1-\theta)p.$$

The expected loss of the persistence forecast is

$$\begin{aligned} E(k_P^N) &= p_{0+|1}\theta p + p_{1+|0}(1-\theta)(1-p) \\ &= (p_{00|1} + p_{01|1})\theta p + E(k_{Ma}^N) - (p_{11|1} + p_{10|1})(1-\theta)p \\ &= E(k_{Ma}^N) + p\left[(p_{00|1} + p_{01|1})\theta - (p_{11|1} + p_{10|1})(1-\theta)\right] \\ &= E(k_{Ma}^N) + p\left[\theta - (p_{11|1} + p_{10|1})\right] \end{aligned}$$

Now, $p_{11|1}+p_{10|1} = P(Y_i = 1|Y_{i-1} = 1)$ and because we are in Case 1, $P(Y_i = 1|Y_{i-1} = 1) \leq \theta$. So $E(k_P^N) = E(k_{Ma}^N)$ when $P(Y_i = 1|Y_{i-1} = 1) = \theta$, and $E(k_P^N) > E(k_{Ma}^N)$ otherwise.

For Case 3, it is shown by the same means that the expected loss of the persistence forecast equals

$$E(k_P^N) = E(k_{Ma}^N) + p\,[\theta - P(Y_i = 1|Y_{i-1} = 1)] + (1-p)\,[P(Y_i = 1|Y_{i-1} = 1) - \theta]$$

Again, $P(Y_i = 1|Y_{i-1} = 1) \leq \theta$ and $P(Y_i = 1|Y_{i-1} = 0) > \theta$, and $E(k_P^N) \geq E(k_{Ma}^N)$. Case 4 is proved in an identical fashion. □

Independence (where the optimal naive Markov forecast and the optimal naive climate forecast are the same) is a special case of Markov. Here too, the persistence forecast is worse than the optimal naive Markov forecasts (think of trying to predict random coin flips: the best guess is to say "Heads" always; while the persistence "forecast" is to always guess whatever the coin was last flip). The lesson is that, in general and except where the optimal naive Markov forecasts overlaps the persistence, persistence forecasts should not be used.

3.4. **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; in this paper we also set the expected loss of a perfect forecast equal to 0):

$$K_\theta(y, x^E) = \frac{E(k_{\text{Ma}}^N) - E(k^E)}{E(k_{\text{Ma}}^N)}, \tag{3.3}$$

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

$$E(k_{\text{Ma}}^N) - E(k^E) = (p_{1|11} - \theta)p_{+1|1}p + (p_{1|10} - \theta)p_{+1|0}(1-p).$$

Also, $E\{k(Y, X^N_{\mathrm{Ma}})\}$ 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

$$\widehat{K}_\theta = \frac{(1-\theta)n_{111} - \theta n_{011} + (1-\theta)n_{110} - \theta n_{010}}{(n_{111}+n_{101})(1-\theta) + (n_{110}+n_{100})(1-\theta)}. \tag{3.4}$$

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 (3.4) ultimately collapses to

$$\widehat{K}_\theta = \frac{(1-\theta)n_{11+} - \theta n_{01+}}{(n_{11+}+n_{10+})(1-\theta)}. \tag{3.5}$$

which is identical to the original climate skill score developed in BR, which is not surprising since the optimal naive Markov forecast is always 0 (in Case 1; as it was for the optimal naive climate forecast in the climate skill score).

More can be done because (3.4) can be written in a more insightful manner and decomposed into parts for when $Y_{i-1} = 1$ and when $Y_{i-1} = 0$, with weights (based on the data) for the importance of the skill score for these two regimes. To be clearer, we are seeking a representation of the skill score like the following:

$$\widehat{K}_\theta = w_1\widehat{K}_{\theta,1} + w_0\widehat{K}_{\theta,0}$$

where $\widehat{K}_{\theta,j}$ is the skill score for those times when $Y_{i-1} = j$, and w_j is the weight based on the data. We derive these weights now.

Let $D = (n_{111} + n_{101})(1 - \theta) + (n_{110} + n_{100})(1 - \theta)$, which is the denominator of equation (3.4). 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 (3.5) 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

$$\widehat{K}_\theta = \frac{\widehat{p}_{1+|1}\widehat{p}}{\widehat{p}_y} \widehat{K}_{1,\theta} + \frac{\widehat{p}_{1+|0}(1 - \widehat{p})}{\widehat{p}_y} \widehat{K}_{0,\theta}. \tag{3.6}$$

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 (3.6) as

$$\widehat{K}_\theta = \widehat{P}(Y_{i-1} = 1|Y_i = 1)\widehat{K}_{1,\theta} + (1 - \widehat{P}(Y_{i-1} = 1|Y_i = 1))\widehat{K}_{0,\theta}. \tag{3.7}$$

This also shows that, as we might expect, $w_0 = 1 - w_1$. This last notation is similar to the idea of sensitivity and specificity.

4. Example

We first start with an example of a simple skill test. The first author 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). This data set is not meant to conclusively diagnose the forecasting abilities of these two agencies; it is only chosen to demonstrate the general ideas of the skill scores and skill plots.

Both Accuweather and the NWS made 1-day ahead forecasts. 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. It presents climate skill and value calculated for a range of $\theta \in (0, 1)$; score values less than 0 indicate no value (for $\theta \neq 1/2$) or no skill (for $\theta = 1/2$). These plots are nearly the same as those developed in Richardson (2000), except they are for skill and not value. The NWS did quite well, beating or closely matching Accuweather's performance for the 1-day ahead predictions. The figure shows also that the NWS forecast would have had value for most users (for many loss values θ). Accuweather performed badly for its 14-day ahead predictions. In fact, any user, *regardless of his loss function*, would have done better and would have suffered less loss had they used the optimal naive climate prediction (no precipitation) during this time.

We show, in Figure 2, the climate skill plot for 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. To show this, 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, SA and SB, (anonymous forecasts taken from media outlets) which have produced forecasts for the same event. The forecasts were given as probability of precipitation.

We first check to see if the precipitation data for which the Brooks et al. forecasts were produced exhibit dependence (independence is of course a special case of 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 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 Markov forecast always being 0, regardless of the value of Y_{i-1}. Obviously, the optimal naive Markov 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.

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 $Y_{i-1} = 1$.

Figure 3 shows the skill plot for Source A along with the 95% point-wise confidence bound, created by inverting the likelihood ratio test statistic G. This confidence interval plot "builds in" the test at the various values of θ.

5. Conclusion

We have shown how to extend the basic skill testing framework developed in BR to events that are Markov. We have also shown how (modifications to) McNemar's test can be used to test for persistence skill, or to compare competing forecasts for the same event.

The climate skill test, while useful, is not entirely satisfactory because it does not take into account the dependent nature of the observations when it exists. 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.

We have also shown that persistence forecasts should never be used, and that the optimal naive Markov (in the usual dependance or independence case) forecast is always better.

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. 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 Markov forecasts.

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

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

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

Case (1)

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

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 climate 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. Agresti, 1990. *Categorical Data Analysis*, Wiley, New York, 558pp.
2. Briggs, W.M., 2005. A general method of incorporating forecast cost and loss in value scores *Monthly Weather Review.* In review.
3. 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.
4. Briggs, W.M., M. Pocernich, and D. Ruppert, 2005. Incorporating misclassification error in skill assessment. *Monthly Weather Review.* In review.
5. Briggs, W.M., and D. Ruppert, 2005. Assessing the skill of yes/no predictions. *Biometrics.* In press.
6. 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.
7. 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.
8. Livezey, R.E., 2003. Categorical events. In *Forecast Verification*, Jolliffe, I.T., and D.B. Stephenson, eds. Wiley, New York, 77–96.
9. Mason, I.B., 2003. Binary events. In *Forecast Verification*, Jolliffe, I.T., and D.B. Stephenson, eds. Wiley, New York, 37–76.
10. McNemar, I., 1947. Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika*, **12**, 153–157.
11. Meeden, G., 1979. Comparing two probability appraisers. *JASA*, **74**, 299–302.
12. Mosteller, F, 1952. Some statistical problems in measuring the subjective response to drugs. *Biometrics*, 220–226.

13. 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).

14. Murphy, A.H., 1991. Forecast verification: its complexity and dimensionality. *Monthly Weather Review*, **119**, 1590–1601.

15. 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.

16. 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.

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

18. Richardson, D.S., 2000. Skill and relative economic value of the ECMWF ensemble prediction system. *Q.J.R. Meteorol. Soc.*, **126**, 649-667.

19. Richardson, D.S., 2001. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q.J.R. Meteorol. Soc.*, **127**, 2473-2489.

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

21. 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.

22. Wilks, D.S., 1991. Representing serial correlation of meteorological events and forecasts in dynamic decision-analytic models. *Monthly Weather Review*, **119**, 1640–1662.

23. Wilks, D.S., 1995. *Statistical Methods in the Atmospheric Sciences*, Academic Press, New York. 467 pp.

24. Wilks, D.S., 2001. A skill score based on economic value for probability forecasts. *Meteorological Applications*, **8**, 209–219.

List of Tables

List of Figures

Table 1. The four cases of optimal naive Markov forecasts. Notice that in Case 2 the optimal naive Markov and persistence forecast are identical. Case 1, when the optimal naive Markov forecast is always 0, is the one used for examples in the main text.

Case	$I\{P(Y_i = 1 \vert Y_{i-1}) > \theta\}$	
	$Y_{i-1} = 0$	$Y_{i-1} = 0$
1	0	0
2	0	1
3	1	0
4	1	1

.

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)

.

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

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

.

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

.

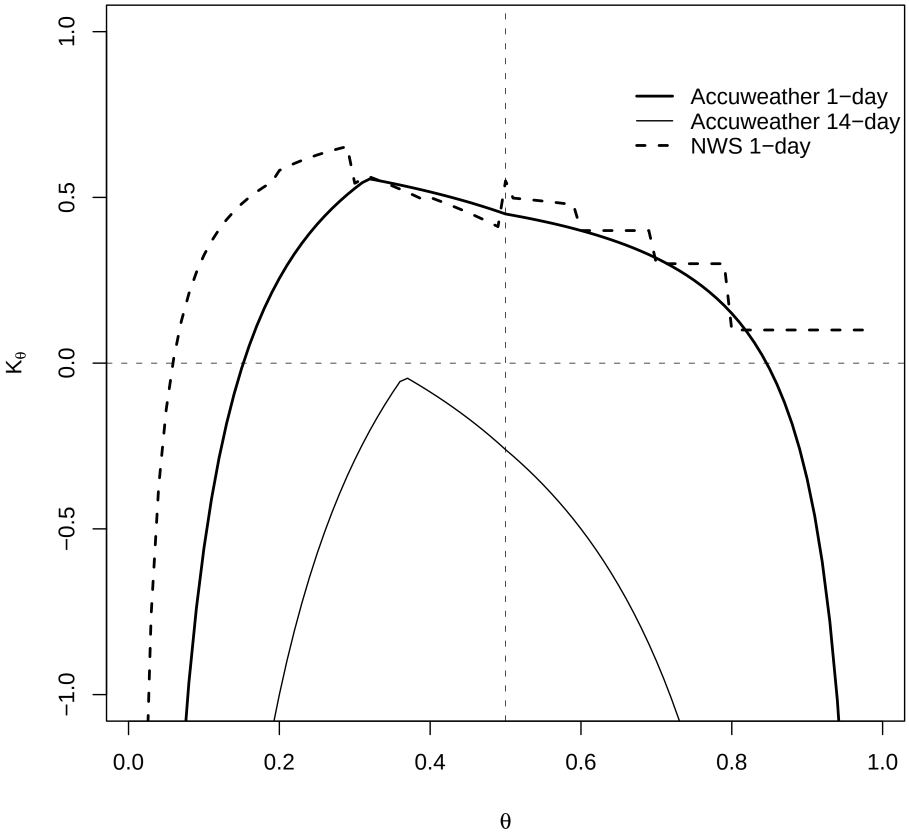

Figure 1. Climate skill score range plot for Accuweather's 1- and 14-day ahead and the NWS's 1-day forecasts. The dashed horizontal line shows 0 and predictions below this line have no skill.

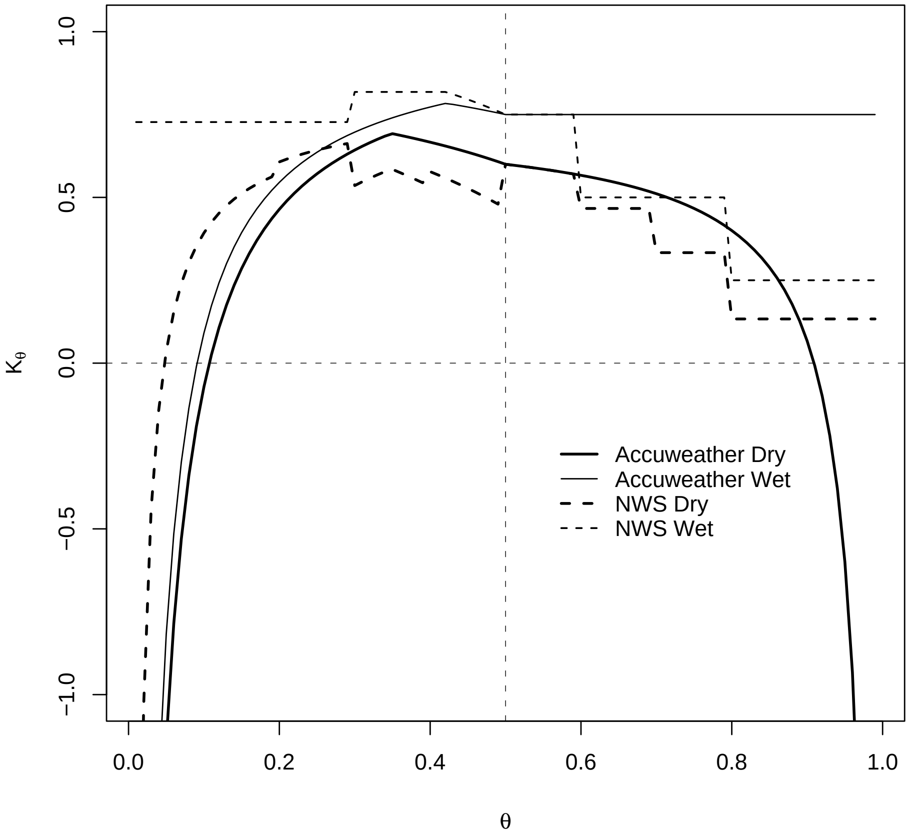

Figure 2. Climate skill score range plot for Accuweather's and NWS's 1-day ahead forecasts, split into days when $Y_{i-1} = 1$ and for $Y_{i-1} = 0$. The dashed horizontal line shows 0 and predictions below this line have no skill.

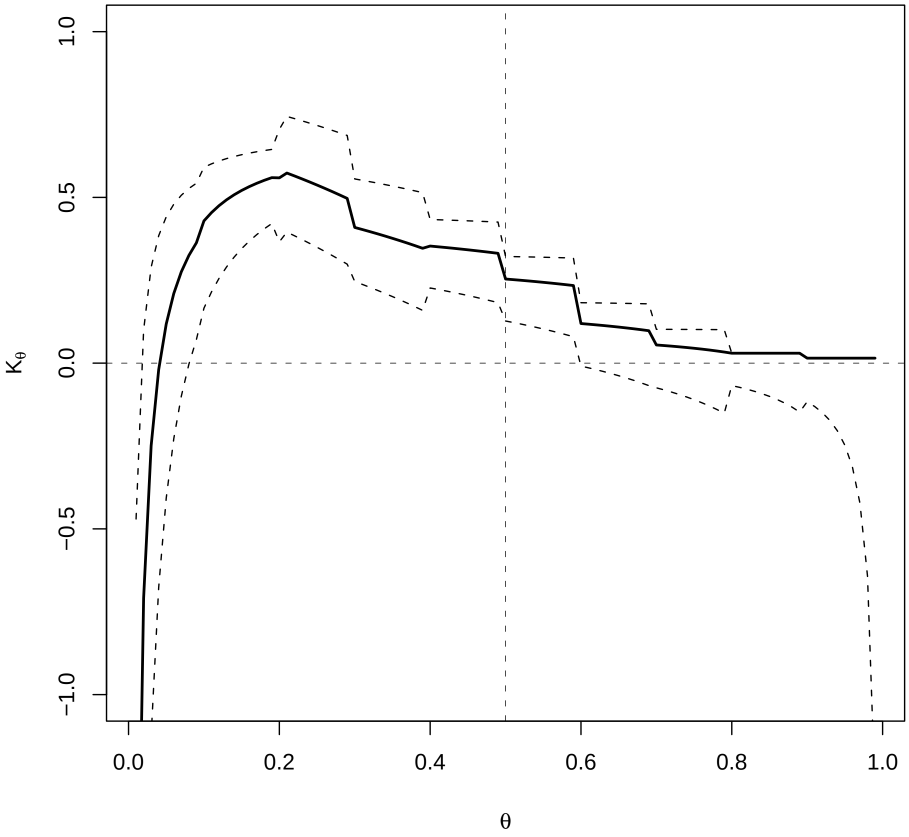

FIGURE 3. Markov skill score (solid line) and point-wise 95% confidence interval (dashed lines) for Source A.