# Incorporating misclassification error in skill assessment

**William Briggs**

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

**and**

**Matt Pocernich**

Research Applications Laboratory,
National Center for Atmospheric Research
3450 Mitchell Lane, Boulder, CO 80301
*email:* pocernic@rap.ucar.edu

**and**

**David Ruppert**

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

April 28, 2005

**Submitted to Monthly Weather Review**

ABSTRACT: It is desirable to account for misclassification error of meteorological observations so that the true skill of the forecast can be assessed. Errors in observations can occur in, among other places, pilot reports of icing, and tornado spotting. Not accounting for misclassification error gives a misleading picture of the forecast's true performance. We present an extension to the climate skill score test developed in Briggs and Ruppert (2005) to account for possible misclassification error of the meteorological observation. This extension supposes a statistical misclassification error model where "gold" standard data, or expert opinion, is available to characterize the misclassification error characteristics of the observation. These model parameters are then inserted into the BR skill score for which a statistical test of significance can be performed.

KEY WORDS: Skill testing; Skill score; Forecast value; Misclassification error; Brier score; Expected loss; Finley tornado forecast.

## 1. Introduction

It is desirable to account for misclassification error of meteorological observations so that the true skill of the forecast can be assessed. Not accounting for this error gives a misleading picture of the forecast's true performance. To this end, we develop a misclassification error model to represent the effect of measurement error on skill assessments. To create this misclassification error model requires either the use of external data set (external to the set of forecasts and observations at hand), called a "gold standard", from which we can estimate the amount of error, or a subjective estimate of the amount of error.

This paper extends the work of Briggs and Ruppert (2004) and Briggs and Ruppert (2005; hereafter, BR) on the hypothesis testing of skill for dichotomous forecasts and observations to the case where the observations are possibly measured with error in the sense that the observations can be misclassified. We will use pilot reports (PIREPs) of icing, and tornado observations as given in Finley (1884) as examples of meteorological observations that can be measured with error (Brown et al., 1997; Finley, 1884; Murphy, 1996).

The hypothesis tests for skill for dichotomous events are fully described in BR. Statistical tests of skill and value are needed to ascertain whether observed skill/value is due to chance or is significant. Other important work on the hypothesis testing of skill can be found in Parker and Davis (1999), Hamill (1998), Wilks (1995), Murphy (1996), Murphy and Ehrendorfer (1987), and Murphy and Winkler (1987).

The primary example used throughout this paper is PIREPs and the accompanying forecasts of icing. The PIREPs are error prone, though the importance of accurate icing forecasts for aircraft is obvious. This error is due to several factors - the variability of the atmosphere, the way in which the icing was encountered (with the aircraft

climbing, descending, etc.) and the subjectivity inherent in any description (Sand and Biter, 1997). Traditional verification measures are not suitable because of the large amount of measurement error in PIREPs.

The modelling framework developed also allows us to better estimate the underlying event rates (e.g. icing events) in the presence of measurement error and other parameters of the joint observation-forecast distribution. These parameters can be interesting in their own right, as we show later in an example.

This paper is organized as follows. Section 2 expounds the mathematical structure of the climate skill test developed in BR, deriving both the skill test and climate skill score. The tests developed can also be seen as tests of value as well as skill, and their differentiation is noted. Section 3 shows the development of the misclassification error model and its relation to the skill test and climate skill score. Section 4 discusses a particular example of PIREPs and forecasts and the amount of detail that can be added by using the misclassification model. We also show how the well-known Finley tornado forecast data can be analyzed with this new method. Finally, Section 5 presents some concluding remarks.

## 2. Climate skill

a. **Notation.** The notation of BR is followed in this paper. We are concerned with events $Y$ which are dichotomous, that is $Y \in \{0, 1\}$. Forecasts $\widetilde{X}$ are made for observations $Y$ which can be either dichotomous ($\widetilde{X} \in \{0, 1\}$) or probabilistic ($\widetilde{X} \in [0, 1]$). Here we consider dichotomous decisions $X \in \{0, 1\}$ based on the forecast $\widetilde{X}$. This implies a transformation of a probabilistic forecast into an eventual dichotomous decision $\widetilde{X} \rightarrow X \in \{0, 1\}$, that is, a forecast user acts is if the event $Y = 1$ will occur, or acts as if the event $Y = 0$ will occur.

We also use the following notation developed in Schervish (1989). Let $Y_i \in \{0, 1\}$ designate the $i$th observation of a dichotomous event, that is, $Y_i = 1$ if the event occurs and equals zero if it does not. Also let $X_i \in \{0, 1\}$ designate the $i$th (possibly transformed) forecast.

Loss is written as $k_{YX}$. We assume $k_{11} = k_{00} = 0$, that is, the loss associated with making a correct decision is 0 (incorporating cost or loss for when $X = Y$ is easily done: see Briggs, 2005). 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 $X = 1$ the loss can be written as some $k_{01} = \theta < 1$, which implies that with $Y = 1$ and decision $X = 0$ the loss is $k_{10} = 1 - \theta > 0$. Note that these inequalities are strict.

The user of the forecast, who may also be called the decision maker, minimizes his expected loss and makes decision $X^E$ based on the forecast $\widetilde{X}$ via the relation $X^E = I(\tilde{X} \geq \theta)$, where the superscript indicates an any "expert" forecast, which is one that is not the optimal naive forecast. Let $p = P(Y = 1)$. The optimal naive forecast is $X^N = I(p \leq \theta)$ which equals 0 when $p \leq \theta$ and equals 1 otherwise. What this means for example, is that, since $\theta \approx 1/2$ and $P(Y = 1)$ is low, the optimal naive forecast is to say "no event".

b. **Climate skill test.** BR framed skill and forecast value in terms of expected loss. In order for a collection of forecasts to have value, we desire that its expected loss should be less than the expected loss incurred by using the optimal naive forecast, that is, when $E(k^E) < E(k^N)$. Typical definitions of skill (see Wilks, 1995) refer to skill as relative accuracy of an expert to a naive forecast, a distinction we will keep when developing skill/value scores below. The naive information we have about $Y$ is that we know $p = P(Y = 1)$, the unconditional probability of occurrence, so that

skill (or value), if it exists, is known as *climate skill* to reflect the idea that the expert forecast can beat the simple or climatological forecast.

We will use the notation for joint and conditional probabilities $p_{yx} = P(Y = y, X^E = x)$, $p_{y|x} = P(Y = y|X^E = x)$. Thus, $P(Y = 1|X = 1) = p_{1|1}$, $P(Y = 0|X = 0) = p_{0|0}$, $P(X = 1) = p_{+1}$, and $P(Y = 1) = p_{1+} = p$. Table 1 shows the model in graphical form. We assume that the observations $(Y_i, X_i)$ are independent and identically distributed. In particular, all of these probabilities are unvarying for all $i$. Also $Y_i$, $X_j$ are independent for $i \neq j$, that is, the forecast observation process is not dynamic and future observations do not depend on past forecasts nor on past observations. See Briggs and Ruppert (2004) for a skill score and test for when $Y$ is Markov (such as a precipitation series might be).

It is convenient in what follows, but not necessary, to transform both the observations and the loss so that the optimal naive forecast $X^N$ is always 0 (see BR; below we also give the results for untransformed forecasts and observations where appropriate). The null hypothesis for the climate skill/value test can now be formed. It is

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

where $k^E$ corresponds to the loss of the expert forecast and $k^N$ is the loss of the optimal naive forecast, and expectation is taken over both forecasts and observations. It is easy to show that $E(k^E) = \theta p_{01} + (1 - \theta)p_{10}$ and that $E(k^N) = p_{1+}(1 - \theta)$.

Substituting for the expected loss, and noting that $X^N \equiv 0$, we have the null hypothesis

$$H_0 : p_{1|1} \quad \leq \quad \theta \tag{2}$$

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

BR showed that *skill defined in terms of expected loss is the same as skill defined in terms of accuracy when the loss is symmetric, that is, when* $\theta = 0.5$. Thus, when $\theta = 0.5$ the null is equivalent to $H_0 : P(Y = X^E) \leq P(Y = X^N)$, that is, the expert forecast's accuracy is less than or equal to the optimal naive forecast's accuracy. As such, it is more proper to speak of a test of value when $\theta \neq 0.5$ instead of strictly a test of skill. We shall mostly use the term "skill", however, as most general verification schemes center, at least implicitly, on symmetric loss situations.

The probability model (likelihood) and the estimates derived from this model for the parameters $p_{y|x}$ can be found in BR, where a test statistic $G = G(n_{yx}, \theta)$, based on a likelihood ratio test, was also developed: these parameter estimates and test statistics are easily found from their misclassification-error versions given in the next Section. Here $n_{yx}$ are simply the counts of a $2 \times 2$ table for $Y$ and $X$, $x, y = 0, 1$ and $\theta$ again is the loss parameter: the $n_{yx}$ are the cell counts of Table 1. Under the null hypothesis, $G$ has an asymptotic distribution which is related to the $\chi^2$ distribution with one degree of freedom. Since the test is one-sided the actual distribution is $1/2\chi_0^2 + 1/2\chi_1^2$ (Self and Liang, 1987; their case 5). Tests are carried out similarly to a standard $\chi_1^2$ test, except that the user must double his chosen test level and use an ordinary $\chi_1^2$ distribution, or equivalently divide the $p$-value given by the usual $\chi^2$ test by 2.

c. **Skill/Value Score.** BR show that testing the significance of a skill score is the same as the climate skill test if the following skill score is taken

$$
\begin{aligned}
K_\theta(y, x^E) &= \frac{E(k^N) - E(k^E)}{E(k^N)} \\
&= \frac{p_{+1}(p_{1|1} - \theta)}{p(1 - \theta)}
\end{aligned}
\tag{3}
$$

where the expected forecast loss is taken as the error score. A collection of perfect expert forecasts will have a loss of 0, so a perfect skill score will be $K \equiv 1$. A collection with "negative" skill, or negative value, as defined in (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

$$H_0 : \quad K_\theta \leq 0. \tag{4}$$

BR showed that this translates exactly to the hypothesis and test used before, defined in (2).

An estimate for the skill/value score is

$$\widehat{K}_\theta = \frac{n_{11}(1-\theta) - n_{01}\theta}{(n_{11}+n_{10})(1-\theta)}. \tag{5}$$

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

$$\widehat{K}_{1/2} = \frac{n_{11} - n_{01}}{n_{11}+n_{10}}. \tag{6}$$

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).

BR also showed the relationship between $K$ and the popular Brier score $B$ (of the dichotomous forecasts). In particular, they found that

$$\widehat{B} = \widehat{p}(1 - \widehat{K}_{1/2})$$

so a collection of prediction has skill (with symmetric loss) when $\widehat{K}_{1/2} > 0$, or when $\widehat{B} < \widehat{p}$. This relationship is useful to draw a level at which the Brier score is skillful.

## 3. Climate skill and misclassification error

a. **Hypothesis test.** We now extend the basic model of BR to include misclassification error. In some cases it is not possible to observe $Y$ directly or to know with certainty that the event $Y$ has or has not occurred. That is, if $Y$ is observed with some error there is the possibility of misclassification. Some $Y = 1$ may be mistakenly classified as $Y = 0$, and some $Y = 0$ may be mistakenly classified as $Y = 1$. An example might be tornado spotting. It has been known that spotters sometimes confuse other meteorological phenomena with tornados, both reporting the existence of a tornado when one was not present and not reporting them when they were there. Tornado reporting difficulties are explored in Speheger et al. (2002). The Finley data (Murphy, 1996) may even be analyzed in this fashion; we do so later. Another example is pilot reports of aircraft icing: again pilots may miss icing and may also erroneously report it. It is desirable to assess how these kinds of uncertainties affect the climate skill test so that the test and skill score reward the forecast when it truly does well but the observations are poor.

The customary classification error framework that we use assumes that some observations $Y$ are mistakenly classified as the opposite value with certain *known* probabilities. These known probabilities can be estimated or gathered from outside sources (outside of this model framework, that is). This practice is common in medicine where a new diagnostic procedure is tested against an older so-called gold standard that is thought to be error free (Geisler et al., 1988). $X$ is the forecast as before, and $Y$ is the unobserved truth. $W$ is defined to be a *diagnosis* or *spotter's report* of the truth $Y$. $W$ is, for example, the pilot report or the spotter's eyewitness report of a tornado. The observational error can be modelled as

$$P(W = 1|Y = 1) = t \tag{7}$$

and

$$P(W = 1|Y = 0) = u \tag{8}$$

Of course, $P(W = 0|Y = 1) = 1 - P(W = 1|Y = 1) = 1 - t$ and $P(W = 0|Y = 0) = 1 - (W = 1|Y = 0) = 1 - u$. We would hope that $t$ will be close to 1 and $u$ close to 0: a perfect observational model has $t = 1$ and $u = 0$. For this model to be sensible (in a probabilistic sense) we require that $P(W = 1|Y = 1) > P(W = 1|Y = 0)$, or $t > u$; further restrictions are necessary for the parameters, which will be detailed in a moment.

The terms $P(W|X)$ may be written incorporating $Y$. For example

$$\begin{aligned} P(W = 1|X = 1) &= P(W = 1, Y = 1|X = 1) + P(W = 1, Y = 0|X = 1) \\ &= P(W = 1|Y = 1)P(Y = 1|X = 1) \\ &\quad +P(W = 1|Y = 0)P(Y = 0|X = 1). \end{aligned}$$

Other terms are modelled in a similar fashion. The key assumption is that, conditional on the true value, the spotter's report is independent of the forecast, that is,

$$P(W|Y, X) = P(W|Y). \tag{9}$$

This is a reasonable assumption if the person responsible for the ultimate diagnosis is unaware of the forecast, or does not let it influence him. This won't always be the case if, for example, a spotter never tries to "spot" if the forecast is $X = 0$ (unfortunately, this may be likely for some meteorological phenomena). As stated above, the probabilities $P(W|Y)$ are assumed to be known. The notation of previous sections will continue to be used, that is, $P(Y = 1|X = 1) = p_{1|1}$ and so on. The full

model may thus be written

$$\begin{aligned}
P(W=1,X=1) &= (u + p_{1|1}(t-u))p_{+1} \\
P(W=0,X=1) &= (1-u-p_{1|1}(t-u))p_{+1} \\
P(W=1,X=0) &= (t - p_{0|0}(t-u))(1-p_{+1}) \\
P(W=0,X=0) &= (1-t+p_{0|0}(t-u))(1-p_{+1}).
\end{aligned}$$

The constants $n_{ij}$ defined before will continue to be used with the modification that the *observed* $W$s and $X$s are counted instead of *unobserved* $Y$s.

The full likelihood, $L$, of the misclassification model can now be built. This model will be used in a test of climate skill in the presence of misclassification error. The likelihood is:

$$\begin{aligned}
(10) \qquad & L(p_{1|1}, p_{0|0}, p_{+1} | W, X, t, u) = \\
& \textstyle\prod_{i=1}^{n} p_{+1}^{X_i}(1-p_{+1})^{1-X_i}(p_{1|1}(t-u)+u)^{W_i X_i}(t-p_{0|0}(t-u))^{W_i(1-X_i)} \\
& (1-u-p_{1|1}(t-u))^{(1-W_i)X_i}(1-t-p_{0|0}(t-u))^{(1-W_i)(1-X_i)}.
\end{aligned}$$

The maximum likelihood parameter estimates for $p_{+1}$, $p_{1|1}$, and $p_{0|0}$ are simple to find:

$$\widehat{p}_{+1} = \frac{n_{11}+n_{01}}{n_{++}}, \qquad \widehat{p}_{1|1} = \frac{n_{11}(1-u) - n_{01}u}{(n_{11}+n_{01})(t-u)}, \qquad \widehat{p}_{0|0} = \frac{n_{00}t - n_{10}(1-t)}{(n_{10}+n_{00})(t-u)},$$

Further, $\widehat{p} = \widehat{P}(Y=1)$ is

$$\widehat{p} = \frac{n_{11}+n_{10}-nu}{n(t-u)}$$

The estimates of these parameters may be interesting in their own right, besides for their use in the skill test. In particular, it is interesting to have a better estimate of the event of interest $\widehat{p} = \widehat{P}(Y=1)$ in the presence of misclassification. For example, the estimate $\widehat{p}$ may be larger or smaller than the error-free naive estimate depending on the values of $t$ and $u$. Examples will be given later. Additionally, the error-free estimates, as given in BR, are easily derived from these by setting $t=1$ and $u=0$.

The estimates for $\widehat{p}_{1|1}$ and $\widehat{p}_{0|0}$ may be rewritten. First let $\widehat{q}_{1|1} = n_{11}/(n_{11}+n_{01})$ and $\widehat{q}_{0|0} = n_{00}/(n_{10}+n_{00})$ (these are the misclassification error-free estimates of $p_{1|1}$

and $p_{0|0}$). Then it is easy to show that

$$\widehat{p}_{1|1} = \frac{\widehat{q}_{1|1} - u}{t - u}, \qquad \widehat{p}_{0|0} = \frac{\widehat{q}_{0|0} - (1-t)}{t-u}. \tag{11}$$

For this model to be probabilistically sensible we need to have the parameter estimates bounded by 0 and 1; that is, we need $0 \leq \widehat{p}_{i|j} \leq 1$ for $i, j = 0, 1$ and $0 \leq \widehat{p}, 1 - \widehat{p} \leq 1$. This leads to restrictions on the possible values of $t$ and $u$. Finding these restrictions is simple. For example, we require $\widehat{p}_{1|1} = \frac{\widehat{q}_{1|1}-u}{t-u} \geq 0$, which means $u \leq \widehat{q}_{1|1}$. But we also require $1 - \widehat{p}_{0|0} = 1 - \frac{\widehat{q}_{0|0}-(1-u)}{t-u} \geq 0$ which means $u \leq 1 - \widehat{q}_{0|0} = \widehat{q}_{1|0}$. Finally, we require $\widehat{p} = \frac{n_{11}+n_{10}-nu}{n(t-u)} \geq 0$ which means $u \leq (n_{11} + n_{10})/n = \widehat{P}(W = 1)$. Similar calculations are made for the parameter $t$ so that we can gather the requirements together in the following form:

$$t \geq \max[n_{10}/(n_{00} + n_{10}), n_{11}/(n_{11} + n_{01}), (n_{11} + n_{10})/n]$$

$$u \leq \min[n_{10}/(n_{00} + n_{10}), n_{11}/(n_{11} + n_{01}), (n_{11} + n_{10})/n].$$

Not unexpectedly, these restrictions are symmetric.

The null hypothesis in this case is identical to the regular climate skill test, that is equation (2). This is because we are interested in parameters modeling the $X$ and $Y$ relationship; we are not specifically interested in the relationship between $X$ and $W$. Thus

$$H_0 : \; p_{1|1} \leq \theta. \tag{12}$$

The alternate is that $p_{1|1} > \theta$, with a maximum likelihood estimate of $\widetilde{p}_{1|1} = \min\{\widehat{p}_{1|1}, \theta\}$.

The likelihood ratio statistic is defined as -2 times the log of the likelihood under the null divided by the general likelihood (as given above). Calculation of the likelihood ratio statistic $G$ is simple as the terms involving $p_{+1}$ and $p_{0|0}$ drop out and

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

Substituting the estimates leads to the likelihood ratio statistic of

$$G = \left( 2n_{11} \log \left[ \frac{n_{11}}{n_{+1}(\theta(t-u)+u)} \right] + 2n_{01} \log \left[ \frac{n_{01}}{n_{+1}(1-u-\theta(t-u))} \right] \right) \times I\left(\widetilde{p}_{1|1} > \theta\right). \tag{13}$$

$G$ has the same distribution as the climate skill test (because, again, the error parameters are fixed and not random in this model).

The role of the misclassification error parameters is now clear. When $u > 0$, $\widehat{p}_{1|1}$ (from 11) decreases, making it harder for skill to be confirmed by the observations. And when $t < 1$ it means that some $Y = 1$ have been not been classified as $W = 1$ when they should have been, and so $G$ (on average) rewards those times when $X = 1$ and $Y = 1$ was mistakenly classified as $W = 0$. That is, the forecast was correct, but because of misclassification error was not believed to be so.

The opposite discussion of the misclassification error parameters is true when the optimal naive forecast is 1. There, it is easy to show that $H_0 : p_{0|0} \geq 1 - \theta$, and so $t < 1$ makes $\widehat{p}_{0|0}$ (from 11) larger which makes it more difficult for there to be skill. And $u > 0$ means that some $Y = 0$ have been not been classified as $W = 0$ when they should have been, and so $G$ rewards those times when $X = 0$ and $Y = 0$ was mistakenly classified as $W = 1$. Again, the forecast was correct, but because of measurement error was not believed to be so.

b. **Measurement Error Skill/Value Score.** The same framework as before is used to develop a skill score. This is a slightly expanded version of equation (3).

$$K_\theta = \frac{(p_{1|1} - \theta)p_{+1}}{(1-\theta)(p_{1|1}p_{+1} + p_{1|0}(1-p_{+1}))}. \tag{14}$$

$\widehat{K}_\theta$ is derived by substituting the estimates for each parameter

$$\begin{aligned}
\widehat{K}_\theta &= \frac{(\widehat{p}_{1|1}-\theta)\widehat{p}_{+1}}{(1-\theta)(\widehat{p}_{1|1}p_{+1}+\widehat{p}_{1|0}(1-\widehat{p}_{+1}))} \\
&= \frac{\left(\frac{n_{11}(1-u)-n_{01}u}{(n_{11}+n_{01})(t-u)}-\theta\right)\frac{n_{11}+n_{01}}{n_{++}}}{(1-\theta)\left(\frac{n_{11}(1-u)-n_{01}u}{(n_{11}+n_{01})(t-u)}\frac{n_{11}+n_{01}}{n_{++}}+\frac{n_{10}(1-u)-n_{00}u}{(n_{10}+n_{00})(t-u)}(1-\frac{n_{11}+n_{01}}{n_{++}})\right)} \\
&= \frac{n_{11}(1-u-\theta(t-u))-n_{01}(u+\theta(t-u))}{(n_{11}+n_{10})(1-\theta)-n_{++}u(1-\theta)} \qquad (15)
\end{aligned}$$

When $t = 1$ and $u = 0$, that is, error free observations, this skill score is identical to equation (5).

Note: some do not prefer to recode the forecast and observations so that the optimal naive forecast is always 0. We can always write the skill score fully: $K_\theta = K_{\theta,0}I(p \leq \theta) + K_{\theta,1}I(p > \theta)$ where $I(p \leq \theta) = 1$ when $p \leq \theta$ (when $X^N = 0$) and 0 otherwise, and $I(p > \theta) = 1$ when $p > \theta$ (when $X^N = 1$) and 0 otherwise. Here, we present the estimate of the skill score for when the optimal naive forecast is 1.

$$\widehat{K}_{\theta,1} = \frac{n_{00}(t-(1-\theta)(t-u))-n_{10}(1-t+(1-\theta)(t-u))}{(n_{00}+n_{01})\theta-n_{++}(1-t)\theta} \qquad (16)$$

## 4. Examples

a. **Prediction and Observation of Aviation Icing in Aviation.** Ice accumulation threatens aircraft and the people on board. To predict this event, the National Center for Atmospheric Research (NCAR) Research Applications Laboratory (RAL) has developed a model known as the Current Icing Potential (CIP) model. The CIP is a deterministic, expert-based model that yields an icing potential value. Icing potentials are expressed on the [0,1] interval (and are similar to the pseudo-probability forecasts $\widetilde{X}$). Since this potential is not calibrated, it is not considered a probability. Calibration has proved to be difficult due the non-random nature of observations and the bias pilots have in reporting icing conditions more frequently during bad weather.

Verification of the presence or absence of icing requires an aircraft be present at that location. Most typically, the presence or absence of icing is documented in a pilot report (PIREP: here, our $W$s). PIREPs note the location, description and intensity of icing. This description is subjective and subject to further errors in its reporting and recording. Further, when the CIP forecasts high icing potentials, general aviation aircraft tend to avoid that area, reducing the number of PIREPS. To overcome these problems, specially equipped research aircraft are deliberately flown into areas with high icing potential. This study considers data collected by these aircraft as a "gold" standard to quantify the effect of measurement error (these aircraft measure our $Y$s). Further descriptions of this data may be found in Wolff and Bernstein (2004).

Computing verification statistics for the CIP forecasts is complicated by the difficulty in obtaining a random sample of icing conditions. The spatial distribution of icing as indicated by the PIREP locations does not reflect the actual spatial distribution of icing conditions. For example, without knowing the reported values, if one knew there was a concentrated area of PIREPs, one could correctly surmise that there was either icing conditions or a high potential for icing in that area. However, the potential location of PIREPs is also dictated by the locations of commercial air traffic routes, clustering around major cities. These factors mean the values of the probability model are poorly estimated, a consequence which is explored below.

Currently, real-time instrumentation is being installed on aircraft to continually measure turbulence and icing conditions. This provides a more systematic and less subjective way in which to collect data over a much larger area. As this technology is introduced, pilots are still required to file PIREPs, permitting development of a better "gold standard." (Of course, one could also imagine over time that pilots of planes fitted with such equipment would become more complacent about reporting

threatening conditions since it's being measured automatically.) The measurement error concepts illustrated here would be applicable to this new source of data.

b. **Eliciting Loss Values.** We consider three loss values: $\theta = \{0.1, 0.2, 0.5\}$. These three values have been selected to explore a range of likely loss values. The actual cost for a false positive or a false negative varies based on the type of aircraft. The general aviation community is the most vulnerable to a false negative forecast, implying $\theta \leq 0.5$. Often general aviation sector planes are not equipped to fly in icing conditions and the pilots don't have experience flying in such conditions. On the other hand, twin engine commuters and larger commercial aircraft have equipment that allows them to shed ice build-up. This makes icing less of a lethal threat and more of an inconvenience and argues for a slightly larger $\theta \approx 0.5$. Briggs (2005) shows how to incorporate non-zero loss (or cost); for example, in the cost-loss problem when $k_{11} = k_{01} > 0$.

c. **Transformations into a dichotomous forecasts and observations.** As mentioned earlier, the CIP ($\widetilde{X}$) is not a calibrated forecast. For this reason, the threshold $h$ used to transform this forecast into a dichotomous forecast is modeled independently from the loss value $\theta$. For specified $h$ values, CIP potentials greater than $h$ are considered a "yes" forecast ($X = I(\widetilde{X} > h)$); potentials less than or equal to $h$ are considered "no" forecasts.

In PIREPs, icing intensity is reported as one of nine ordinal descriptions ranging from none to extreme amounts. For this study, icing descriptions greater than "trace" are considered to indicate the presence of icing. This is a procedure commonly used within RAL. A histogram of CIP potentials ($P(\widetilde{X}|Y = j), j = 0, 1$) for icing and non-icing events is presented in Figure 1. This figure is also referred to as a discrimination plot. Ideally, one would see to distinct clusters of values. In this plot, we see that the

forecast is nowhere near perfect. Table 2 shows the PIREPS observations ($X$) and forecasts ($W$).

The presence and absence of icing on the research aircraft was determined by the union of several conditions. The temperature had to be suitably cold, the moisture content had to be non-zero and the a conductivity probe had to indicate the build-up of ice. In cases where multiple aircraft measurements corresponded to a single CIP potential, icing was considered to be present if any reading indicated positive. Further details on these data can be found in Pocernich et al. (2004).

Note that a concern and potential criticism of linking research plane data with PIREPs is that often there remains a reasonable distance between the two aircraft. Conditions favorable to icing may have a spatial resolution that is much less than this distance, so conceptually, while in disagreement, both planes can be completely accurate in the description of the icing conditions.

d. **Skill score as a function of forecast threshold.** Since the CIP is not calibrated, skill scores $K_\theta$ (15) were first calculated for $\theta = \{0.1, 0.2, 0.5\}$ across a range of transformation thresholds $h$ assuming no misclassification error in the PIREPS ($t = 1, u = 0$). Note that the optimal naive forecast based on Table 2 is to always say 1 (or to always say icing will be present), therefore the value of $K_\theta$ is given by (16). Figure 2 illustrates the effect of $\theta$ on skill scores. For $\theta$ values of 0.1 and 0.2, the skill score remains entirely negative. For $\theta = 0.5$, a positive skill score exists for transformation thresholds less than 0.55. This is disturbing since given the nature of the problem, $\theta$, the cost for a false positive, is likely to be small, while the cost for a false negative, $1 - \theta$, will be large.

e. **Incorporation of misclassification error.** The gold-standard data was collected by a Twin Otter aircraft operated by the NASA-Glenn Research Station. This

aircraft was flown to measure meteorological conditions related to aviation icing. This data was matched to PIREPs occurring in the vicinity. A total of 74 pairs of research aircraft ($Y$) and PIREP ($W$) reports were gathered. Table 3 summarizes the data.

From Table 3 we may estimate the misclassification error model parameters. A first estimate for the parameters might be: $\widehat{P}(W = 1|Y = 1) = t = 0.8113$ and for $\widehat{P}(W = 1|Y = 0) = u = 0.8095$. Recall that these parameters are going to be taken as *fixed* in the misclassification error skill model to come.

The restrictions on the parameters of $t$ and $u$ require that $t \geq \max[0.49, 0.83, 0.60]$ and that $u$ be less than or equal to the minimum of this set. The CIP data suggest that $t < 0.83$ and $u > 0.49$ so that neither model constraint is met. We now have to consider how valid these misclassification error parameters are.

It can be argued that the data from Table 3 is incomplete because the research aircraft would usually not be sent up unless there was suspicion of icing; therefore the count when $Y = 0$ and $W = 0$ is far too low. This would imply that the estimate for $u$ is far too large. Unfortunately, there is no other way to estimate $u$ except subjectively because of this, a common difficulty in misclassification error models. It is our opinion that $u$ is probably closer to 0, somewhere around 0.1-0.3. We have more confidence that the value for $t$ is better estimated, although its value would also clearly change if the research aircraft were sent when icing potentials were low.

A skill score (with symmetric loss) applied to the data from the winter of 2003 (from Table 2) which was converted from a forecasted potential to a binary forecast using the threshold of $h = 0.5$, and ignoring measurement error, is

$$K_{1/2} = 0.017$$

with a $G = 1.08$ and a p-value of 0.15, which indicates slight skill but that it is not statistically significant by the usual criterion.

We now account for measurement error by allowing $t$ and $u$ to range over their possible values, calculating $\widehat{K}_{1/2}$ for each pair, plotting the result in Figure 3. For the $t$ value of $\approx 0.82$ estimated with NASA Twin Otter data we see that $u$ must be at least about 0.18 for there to be useful skill. Fix $t$ at 0.82: as $u$ increases, the skill score increases. This is due to the times when the forecast was for no icing ($X = 0$), but icing was incorrectly observed ($W = 1$ but $Y = 0$). $u = P(W = 1|Y = 0)$, so that $u$ approaching 1/2 says that the pilots ability to correctly identify actual non-icing is quite poor. Also note the wide range of the skill score as $u$ changes. Skill can be anywhere from very large, to nonexistent depending on the value of $u$. This shows skill has a heavy dependence on $u$, so much so that a decision maker may regard the possibility of actual skill for PIREP forecasts as undecidable unless better gold standard data could be had. A similar fact emerges in the picture of $\widehat{p}_{0|0}$ which must be greater than 0.5 for skill to exist. Interestingly, the estimate for icing, $\widehat{p} = \widehat{P}(Y = 1)$, with $t \approx 0.82$ and $u = 0.18$, is about 0.65, which is higher than the naive estimate of $(4028 + 5161)/15254 = 0.60$. This means that misclassification error causes us to underestimate the true frequency of icing. Due to observation error and the difficulties associated with it, icing may occurring more frequently then the raw data indicate. The estimate of $\widehat{p}_{1|1}$ is included for completeness, and may be of interest for its own sake.

f. **Finley.** This measurement error technique can also be used to reexamine the Finley tornado forecasting data (Finley, 1884; summarized in Murphy 1996 and Wilks, 1995). Table 4 lists the data.

The optimal naive forecast, assuming symmetric loss, is to say no tornado. The climate skill score is $K = -0.86$, which indicates, as is well known, a lack of skill.

Certainly, observations of tornados at this time (and the present) were not perfect; one can imagine $t$ and $u$ being different than ideal. So we can ask for what values of $t$ and $u$ would Finley's data have skill.

Fixing $u = 0$, and $t \leq 0.55$ just gives $K > 0$. We don't know if this is a plausible value for $t$ to rescue Finley. Having a $u > 0$ at this $t$ decreases $K$ and makes it less likely for skill to exist. A contour plot of $\widehat{K}$ for various values of $t$ and $u$ (the restrictions were $u \leq 0.009$ and $t \geq 0.280$) is presented in Fig. 4: for almost no values of $t$ and $u$ is $\widehat{K} > 0$, making it unlikely that these forecasts possess any skill.

## 5. Conclusions

Formal tests of simple skill for forecasts for dichotomous events that are possibly measured with error have been developed. Skill/Value is when the expected loss incurred while using an expert forecast was less than the expected loss that would have been incurred had the optimal naive forecasts been used instead.

It was shown that measurement error can both positively and negatively affect the assessment of skill, and that if measurement error isn't accounted for a misleading picture of true forecast performance can be created.

This new statistic is a simple extension, but requires the user supply measurement error parameters. These parameters may not always be easy to get, and the user may have to settle for a subjective estimate of them.

Many types of weather phenomena can now be remotely estimated. For example, rainfall can be estimated away from weather stations through the use of radar. This is not true icing and turbulence. These weather conditions still require an aircraft be present to be observed. Presently, this information is most commonly gained from PIREPs although slowly this is changing. Some commercial aircraft are being

equipped with instruments to measure and immediately relay icing and turbulence data. In the future, this will provide a better gold standard than the more limited research aircraft data used in this paper. Results shown here indicate that due to observation error, the frequency of icing in the vicinity of PIREPs may be higher than naively assumed.

## Acknowledgements

Thanks to Dan Wilks for his comments and clarifications and to Persi Diaconis for his encouragement and many helpful comments during the early stages of this research. We are also especially grateful to three reviewers who helped us to clarify the exposition of this paper.

## References


1. Briggs, W.M., 2005: Incorporating Forecast Cost in the new Skill Score. Submitted.
2. Briggs, W.M., and D. Ruppert, 2004: Assessing the skill of yes/no forecasts for Markov observations. *17th Conf. on Probability and Statistics in the Atmospheric Sciences*, Seattle, WA, Amer. Meteor. Soc.
3. Briggs, W.M., and D. Ruppert, 2005: Assessing the skill of yes/no forecasts. *Biometrics*, accepted.
4. Brown, B.G. and G. Thompson and R.T. Bruintjes and R. Bullock and T. Kane: 1997: Intercomparison of in-flight icing algorithms. Part II: Statistical verification results, *Weather Forecasting*, **12**, 890-914.
5. Finley, J.P., 1884: Tornado forecasts. *Amer. Meteor. J.*, **1**, 85-88.
6. Geisler, F.H., Jelinek, J.J., Joslyn, J.N., and F. Gelland, 1988: Acute cervical spine trauma: evaluation with 1.5-T MR imaging. *Radiology*, **166**, 807-816.
7. Hamill, T. M., 1998. Hypothesis tests for evaluating numerical precipitation forecasts: *Weather and Forecasting*, **14**, 155-167.
8. 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).
9. Murphy, A.H., 1996: The Finley affair: a signal event in the history of forecast verification. *Weather and Forecasting*, **11**, 3-20.

10. 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.

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

12. Parker, R.A., and R.B. Davis, 1999: Evaluating whether a binary decision rule operates better than chance. *Biometrical Journal*, **41**, 25-31.

13. Pocernich, M.J., C. Wolff and T. Fowler, 2004: Statistical models of aircraft icing. 17th Conf. Prob and Stat in Atmos. Sci., Seattle WA. Amer. Meteor. Soc.

14. W.R. Sand and C. Biter, 1997, Pilot response to icing: It depends, *Preprints, 7th Conf. on Aviation, Range, and Aerospace Meteorology, 2-7 February, Long Beach, CA, American Meteorological Society*, 116-119.

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

16. 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.

17. Speheger, D. A., C. A. Doswell III, and G. J. Stumpf, 2002: The Tornadoes of 3 May 1999: Event Verification in Central Oklahoma and Related Issues. *Wea. Forecasting*, **17**, 362-381.

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

19. Wolff, C.A. and B.C. Bernstein, 2004: Scales of aircraft icing: A comparison of icing PIREPs to liquid water measurements from research aircraft, *11th Conference on Aviation, Range and Aerospace Meteorology, Hyannis MA, 11-14 October, Amer. Meteor. Soc., Boston. Available on CD from the AMS.*

# List of Tables

# List of Figures

**Figure 1** Histogram of CIP forecast during icing and non-icing events. Based on 15,254 events during Winter 2003.

**Figure 2** Error-free skill score $K_\theta$ as a function of transformation threshold $h$, assuming no misclassification of the PIREPS.

**Figure 3** Contour plot of the skill score $K_{1/2}$, $\widehat{p}_{0|0}$, $\widehat{p}_{1|1}$ and $\widehat{p}$ as functions of $t$ and $u$. For the observed $t \approx 0.82$ $u$ must be larger than about 0.18 for there to be skill. The estimate $\widehat{p} = \widehat{P}(Y = 1)$ is also larger than the naive, misclassification-free estimate.

**Figure 4** Contour plot of the skill score $K_{1/2}$, $\widehat{p}_{0|0}$, $\widehat{p}_{1|1}$ and $\widehat{p}$ as functions of $t$ and $u$. Note that $t$ must be less than about 0.5 for most values of $u$ to rescue Finely's forecasts.

TABLE 1. Parameters used in the probability model. Each cell contains the estimate for $P(Y = i|X = j)P(X = j)$ for $i, j = 0, 1$. The marginals contain the estimate for the $P(Y = i)$ and $P(X = i)$, $i = 0, 1$. The loss $k_{ij}$ for an observation/forecast pair $i, j = 0, 1$ is also shown: note that when $Y = X$ the loss is 0. See Briggs (2005) for cases where this need not hold.

| | | Y: 1 | Y: 0 | |
|---|---|---|---|---|
| X | 1 | $p_{1\|1}p_{+1}$; $k_{11} = 0$ | $(1 - p_{1\|1})p_{+1}$; $k_{01} = \theta$ | $p_{+1}$ |
| | 0 | $(1 - p_{0\|0})(1 - p_{+1})$; $k_{10} = 1 - \theta$ | $p_{0\|0}(1 - p_{+1})$; $k_{00} = 0$ | $1 - p_{+1}$ |
| | | $p$ | $1 - p$ | |

.

TABLE 2. PIREP forecast data from the winter of 2003. $W = 1$ indicates a pilot reported icing, and $W = 0$ they did not. $X = 1$ indicates icing was forecasted, and $X = 0$ indicates it was not.

| | $W = 1$ | $W = 0$ |
|---|---|---|
| $X = 1$ | 4028 | 798 |
| $X = 0$ | 5161 | 5267 |

.

TABLE 3. PIREP gold standard data, from the CIP Twin Otter study. $Y = 1$ and $Y = 0$ indicate the presence/ absence of icing as measured by the research aircraft. $W = 1$ or $W = 0$ indicates the presence or absence of icing as reported on a PIREP.

| | $Y = 1$ | $Y = 0$ |
|---|---|---|
| $W = 1$ | 43 | 17 |
| $W = 0$ | 10 | 4 |

.

TABLE 4. Finley's tornado data.

| | Y=1 | Y=0 |
|---|---|---|
| X=1 | 28 | 72 |
| X=0 | 23 | 2680 |

.

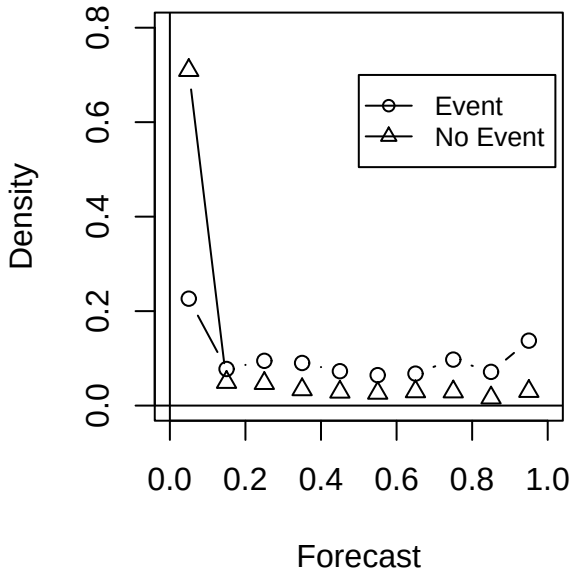


FIGURE 1. Histogram of CIP forecast during icing and non-icing events. Based on 15,254 events during Winter 2003.

.

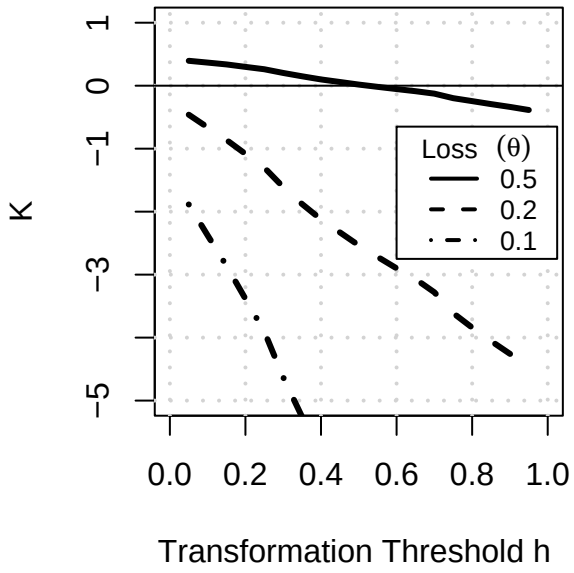


FIGURE 2. Error-free skill score $K_\theta$ as a function of transformation threshold $h$, assuming no misclassification of the PIREPS.

.

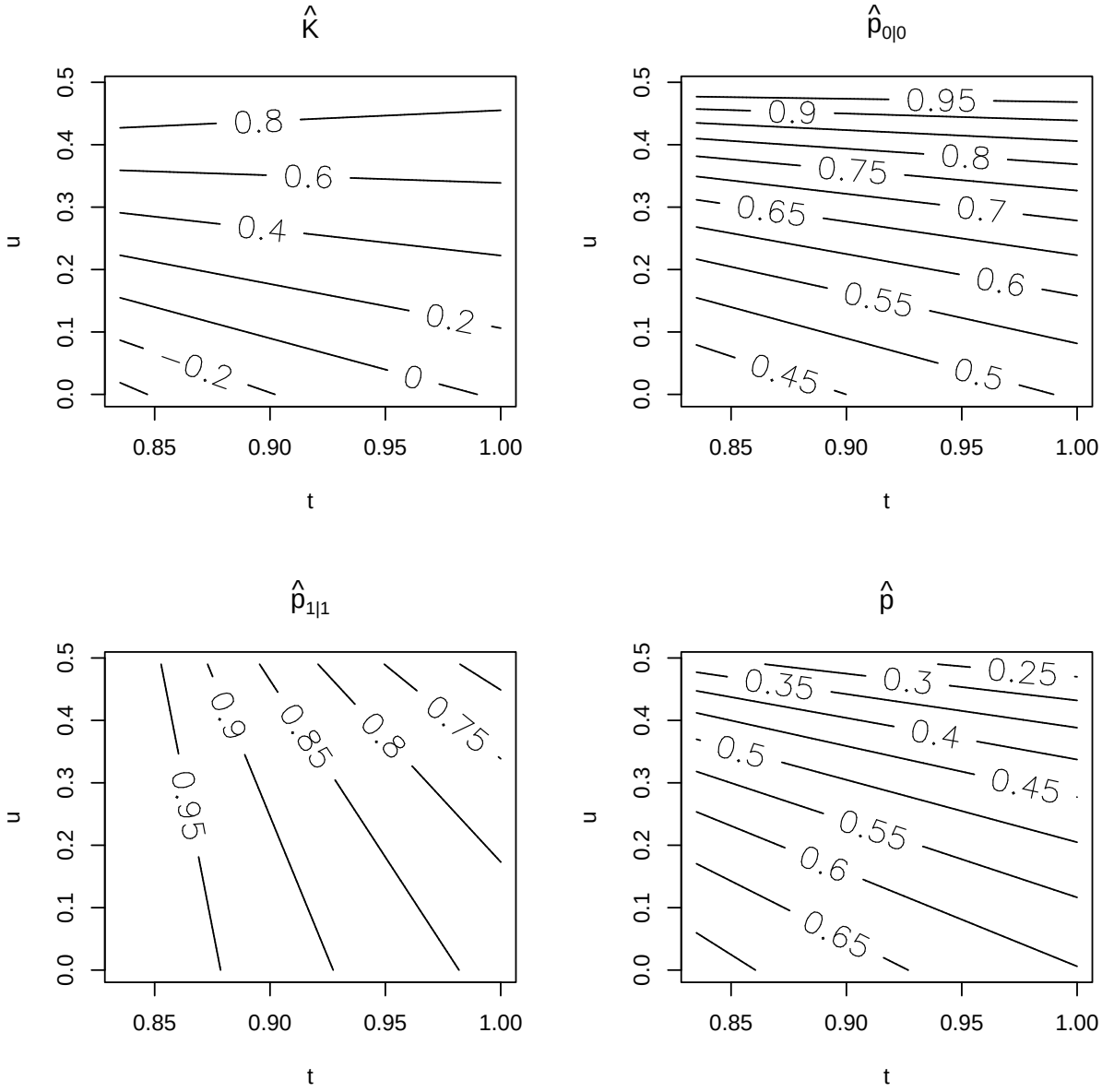


FIGURE 3. Contour plot of the skill score $K_{1/2}$, $\widehat{p}_{0|0}$, $\widehat{p}_{1|1}$ and $\widehat{p}$ as functions of $t$ and $u$. For the observed $t \approx 0.82$ $u$ must be larger than about 0.18 for there to be skill. The estimate $\widehat{p} = \widehat{P}(Y = 1)$ is also larger than the naive, misclassification-free estimate.

.

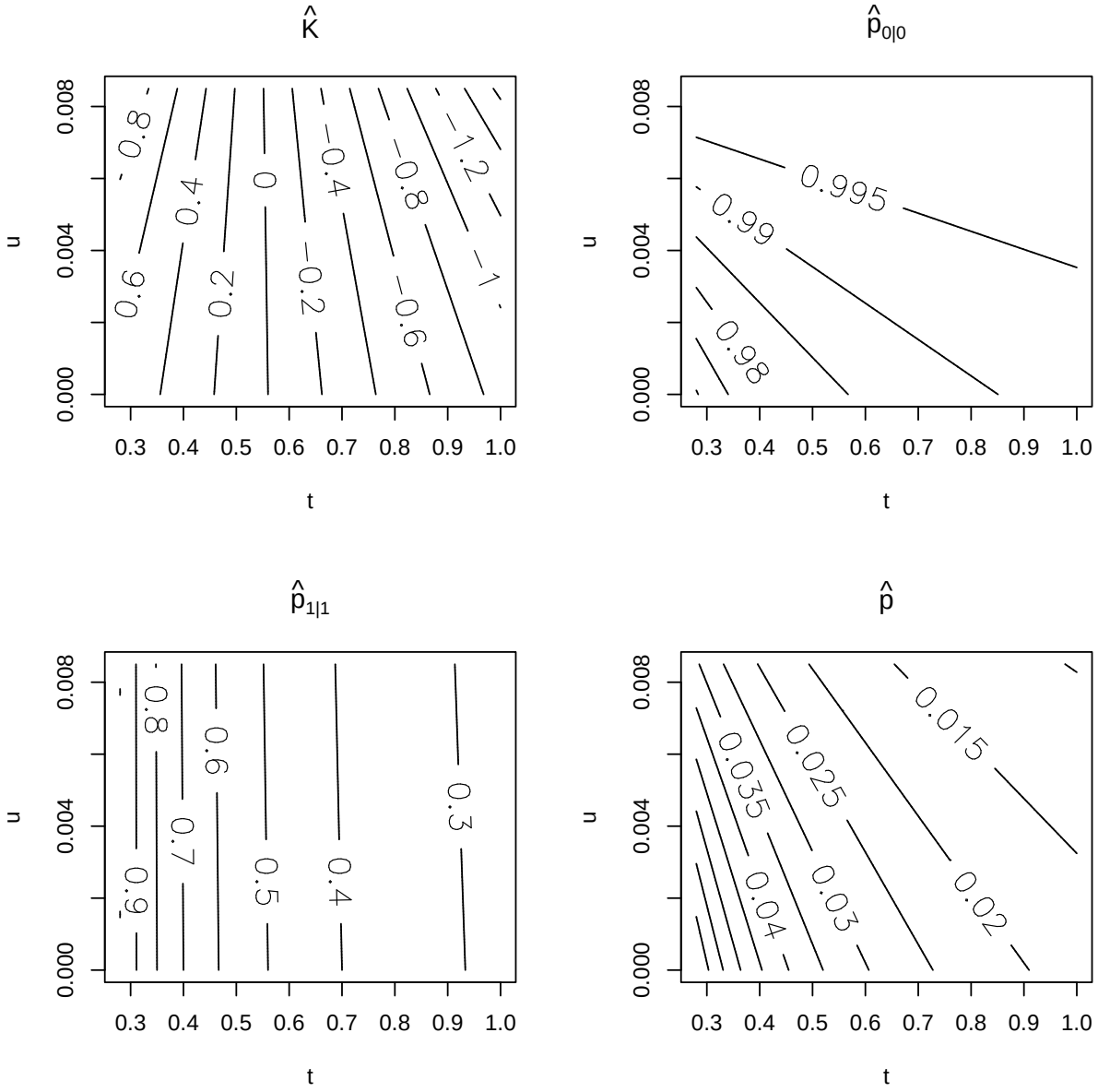


FIGURE 4. Contour plot of the skill score $K_{1/2}$, $\widehat{p}_{0|0}$, $\widehat{p}_{1|1}$ and $\widehat{p}$ as functions of $t$ and $u$. Note that $t$ must be less than about 0.5 for most values of $u$ to rescue Finely's forecasts.

.