Measuring forecast performance in the presence of observation error

C. A. T. Ferro

Department of Mathematics, University of Exeter, UK

*Correspondence to: C. A. T. Ferro, Department of Mathematics, University of Exeter, Laver Building, North Park Road, Exeter EX4 4QE, UK. E-mail: c.a.t.ferro@exeter.ac.uk

A new framework is introduced for measuring the performance of probability forecasts when the true value of the predictand is observed with error. In these circumstances, proper scoring rules favour good forecasts of observations rather than of truth and yield scores that vary with the quality of the observations. Proper scoring rules thus can favour forecasters who issue worse forecasts of the truth and can mask real changes in forecast performance if observation quality varies over time. Existing approaches to accounting for observation error provide unsatisfactory solutions to these two problems. A new class of 'error-corrected' proper scoring rules is defined that solves both problems by producing unbiased estimates of the scores that would be obtained if the forecasts could be verified against the truth. A general method for constructing error-corrected proper scoring rules is given for the case of categorical predictands, and error-corrected versions of the Dawid-Sebastiani scoring rule are proposed for numerical predictands. The benefits of accounting for observation error in ensemble post-processing and in forecast verification are illustrated in three data examples that include forecasts for the occurrence of tornadoes and of aircraft icing. **Copyright © 0000 Royal Meteorological Society**

Key Words: observation errors; probability forecasts; proper; scores; scoring rules; verification

Received . . .

Citation: . . .

1. Introduction: two problems

The performance of probability forecasts is commonly measured by using a scoring rule to assign a score to each forecast. If the probability distribution f is issued as a forecast for a certain predictand and if the value of the predictand is subsequently observed to be y then the scoring rule s assigns the score $s(f, y)$ to the forecast. The performance of a set of forecasts is summarized by their mean score. We assume throughout that scoring rules are chosen to be negatively oriented, which means that lower scores indicate better forecasts.

A scoring rule is called 'proper' if the expected value of the score, taken over any probability distribution, q, for y, is minimized when $f = q$. Proper scoring rules encourage the forecaster to be honest because if the forecaster's belief about the predictand is represented by q then the forecaster's expected score is minimized by issuing q as the forecast. Proper scoring rules also reward forecasts that are calibrated and sharp (e.g. Winkler, 1996; Gneiting *et al.*, 2007), and are widely used as definitive measures of performance for probability forecasts (e.g. Broecker, 2012).

The application of scoring rules typically ignores the fact that the observation, y, may not be the true value of the predictand: the predictand might be observed with error. Such errors may be due to measurements being inexact (instrument error), to the measured quantity differing from the predictand (representativity error), to rounding and mistakes in transcription (recording error) etc. In the presence of observation error, using proper scoring rules can have undesirable consequences. If the forecaster's belief about the true value, x, of the predictand is p, but the forecaster's belief about the observed value, y, is $q \neq p$ then the forecaster's expected score is minimized by issuing q, not p, as the forecast. Proper scoring rules also reward forecasts that are calibrated to the observed values, y, rather than to the true values, x. In other words, proper scoring rules will favour good forecasts of y rather than good forecasts of x. Sometimes this is desirable (for example, if a bet will be decided by the value of the observation) and sometimes there is no observation error (for example, when forecasting stock prices). In environmental forecasting, however, observations tend to be inexact and our usual aim is to forecast the true value because that is what will affect us. The following example, to which we shall refer repeatedly, illustrates this problem with proper scoring rules.

Example 1. *Suppose that the true value, x, is observed with random error, w, to yield the observed value, $y = x + w$. Suppose also that w is independent of x and follows a Normal distribution with mean 0 and variance c^2, denoted $N(0, c^2)$. We shall refer to this form of observation error as white noise. Suppose that the forecast for x is $N(\mu, \sigma^2)$ with probability density function f, and that the true distribution of x (which we may think of as representing the forecaster's honest belief about x) is $N(\mu_0, \sigma_0^2)$. The strictly proper logarithmic scoring rule (Good, 1952) is*

$$\begin{aligned} s(f, y) &= -\log f(y) \\ &= \frac{1}{2}\log(2\pi) + \log\sigma + \frac{(y-\mu)^2}{2\sigma^2}. \end{aligned} \qquad (1)$$

For brevity, we suppress the constant $\log(2\pi)/2$ in all such scoring rules hereafter. The forecaster's expectation for $s(f, y)$ with respect to y is

$$\mathrm{E}_y\{s(f, y)\} = \log\sigma + \frac{(\mu-\mu_0)^2 + \sigma_0^2 + c^2}{2\sigma^2}. \qquad (2)$$

This is minimized when the forecast has $\mu = \mu_0$ and $\sigma^2 = \sigma_0^2 + c^2$. If there is no observation error ($c = 0$) then $\sigma = \sigma_0$ and the optimal forecast equals the true distribution of x, as desired. In the presence of observation error, on the other hand, the optimal forecast has $\sigma > \sigma_0$ and the scoring rule favours over-dispersed forecasts, as illustrated in Figure 1.

This example also highlights a second effect of observation error. The expected score (2) exceeds by an

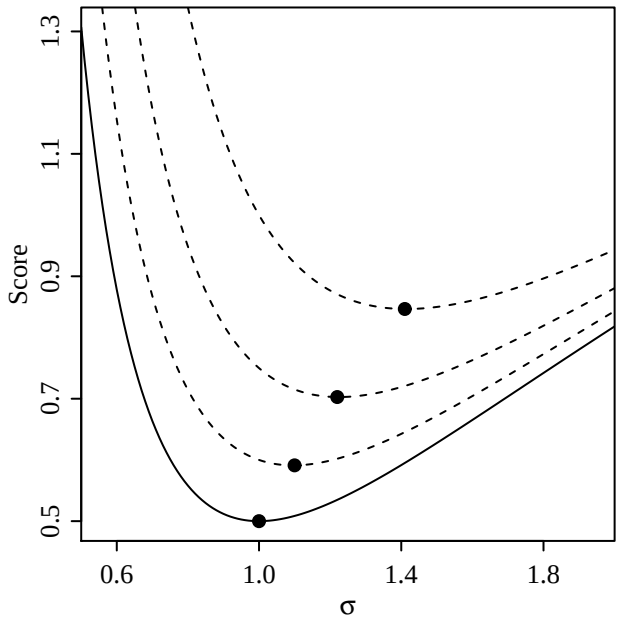


Figure 1. Expected values (2) of the logarithmic scoring rule plotted against the forecast standard deviation, σ, when the forecast mean is correct ($\mu = \mu_0$), when the true standard deviation is $\sigma_0 = 1$, when there is no observation error ($c = 0$, solid line) and when the error standard deviation is $c = 0.2, 0.5$ and 1 (lower, middle and upper dashed lines). The dots (•) mark the minima.

amount $c^2/(2\sigma^2)$ the expected score,

$$\mathrm{E}_x\{s(f,x)\} = \log\sigma + \frac{(\mu-\mu_0)^2+\sigma_0^2}{2\sigma^2}, \qquad (3)$$

that would obtain if there were no observation error. Thus, the logarithmic scoring rule (1) tends to overestimate the score that would be obtained were we able to observe x without error (see Figure 1 again). This means that forecasts appear worse than they actually are, and that the score is sensitive to the quality of the observations: the score will tend to increase as the error variance increases.

We have described how the existence of observation error means that proper scoring rules favour good forecasts of the observations rather than of the truth, and yield scores that vary with the quality of the observations. The first problem has implications for deciding which of two forecasters, or forecasting systems, should be preferred: proper scoring rules can lead us to favour the forecaster who issues worse forecasts of the truth. The second problem has implications for monitoring the performance of a forecaster over time: proper scoring rules can mask real changes in forecast performance if observation quality varies. As we illustrate later, in the appendix, these two problems are likely to be felt most acutely in situations where observation errors are as large as forecast errors, for example at short lead times or when observation quality is poor, and may become more pronounced as forecast quality improves (Bowler, 2008; Mittermaier and Stephenson, 2015).

We shall overcome these two problems by constructing scoring rules that favour good forecasts of the true value of the predictand even in the presence of observation error, and that are insensitive to the quality of the observations. We make precise our definitions of such scoring rules in section 2. In section 3, we critique several ways of handling observation error that have been proposed by other authors. We show how to construct the new scoring rules in section 4, present data examples in section 5 and close with a discussion in section 6.

2. Proper scoring rules and observation error

2.1. *Proper scoring rules*

Before extending the idea of proper scoring rules to account for observation error, let us revise the formal definition of a proper scoring rule (e.g. Gneiting and Raftery, 2007). We write $x \sim p$ to denote that x has distribution p.

Definition 1. *A negatively oriented scoring rule, s, is said to be proper relative to a class, $\mathcal{P}$, of probability forecasts if*

$$\mathrm{E}_x\{s(f,x)\} \geq \mathrm{E}_x\{s(p,x)\} \qquad \textit{for all } f,p \in \mathcal{P} \qquad (4)$$

when $x \sim p$.

This says that if the true distribution, p, of x belongs to the class $\mathcal{P}$ then no other choice of forecast, f, from $\mathcal{P}$ yields a better expected score.

Example 2. *The fact that the expected score* (3) *is minimized when $\mu = \mu_0$ and $\sigma = \sigma_0$ means that the logarithmic scoring rule, $s(f,x) = -\log f(x)$, is proper relative to the class of all Normal distributions. In fact, it is proper relative to the class of all well-behaved distributions (Gneiting and Raftery, 2007).*

Now let x denote the true value of the predictand and let y denote the observed value. We represent observation error by a model (which we call the observation model) for the conditional distribution of y given x. For example, for the white noise observation error in Example 1, the observation model specifies that the conditional distribution of y given x is $N(x, c^2)$. For a generic observation model, r, we denote the conditional density of y given x by $r(y \mid x)$ and write $y \mid x \sim r$. Our development requires that r is known.

Scoring rules must be functions of the forecast and the observed value rather than of the forecast and the true value since the latter will be unknown. We saw in Example 1, however, that evaluating a proper scoring rule for y rather than x will penalize good forecasts of x. Are there other scoring rules that favour good forecasts of x? The following class of scoring rules meets this requirement.

Definition 2. *A negatively oriented scoring rule, s, is said to be proper under observation model r and relative to a class, $\mathcal{P}$, of probability forecasts if*

$$\mathrm{E}_y\{s(f, y)\} \geq \mathrm{E}_y\{s(p, y)\} \qquad \textit{for all } f, p \in \mathcal{P}$$

when $x \sim p$ and $y \mid x \sim r$.

This says that if the observation model is r and if the true distribution, p, of x belongs to the class $\mathcal{P}$ then no other choice of forecast from $\mathcal{P}$ yields a better expected score. This generalizes the notion of a proper scoring rule since Definition 1 arises as a special case when r prescribes no observation error, in which case $y = x$. To be proper under an observation model, r, scoring rules typically need to depend on r as well as f and y, as in the following example.

Example 3. *Recall Example 1 in which we showed that the logarithmic scoring rule* (1) *is not proper under the white noise observation model. Consider the alternative scoring rule*

$$s_*(f, y) = \frac{1}{2}\log(\sigma^2 + c^2) + \frac{(y - \mu)^2}{2(\sigma^2 + c^2)}, \qquad (5)$$

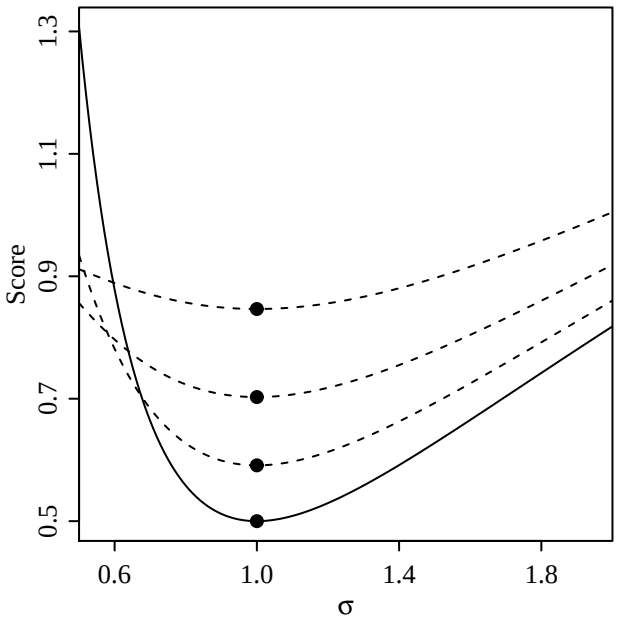


Figure 2. Expected values (6) of the error-convolved logarithmic scoring rule plotted against the forecast standard deviation, σ, when the forecast mean is correct ($\mu = \mu_0$), when the true standard deviation is $\sigma_0 = 1$, when there is no observation error ($c = 0$, solid line) and when the error standard deviation is $c = 0.2, 0.5$ and 1 (lower, middle and upper dashed lines). The dots (•) mark the minima.

which we call the error-convolved logarithmic scoring rule for reasons given in section 3.3. The expected score is

$$\mathrm{E}_y\{s_*(f, y)\} = \frac{1}{2}\log(\sigma^2 + c^2) + \frac{(\mu - \mu_0)^2 + \sigma_0^2 + c^2}{2(\sigma^2 + c^2)}, \qquad (6)$$

which is minimized when $\mu = \mu_0$ and $\sigma = \sigma_0$, as desired. Thus, the error-convolved logarithmic scoring rule is proper under the white noise observation model and relative to the class of Normal distributions, as illustrated in Figure 2.

2.2. *Error-corrected scoring rules*

Although the error-convolved logarithmic scoring rule (5) is proper under the white noise observation model, the expected score (6) typically differs from the expected score (3) that would obtain if there were no observation error (see Figure 2 again). Thus, the error-convolved logarithmic scoring rule, like the logarithmic scoring rule, is sensitive to the quality of the observations. Are there scoring rules that are insensitive to the quality of the observations? The following class of scoring rules meets this requirement.

Definition 3. *A scoring rule, s, is said to be unbiased for a scoring rule, s_0, under observation model r and relative to a class, $\mathcal{P}$, of probability forecasts if*

$$\mathrm{E}_y\{s(f,y)\} = \mathrm{E}_x\{s_0(f,x)\} \qquad \textit{for all } f,p \in \mathcal{P}$$

when $x \sim p$ and $y \mid x \sim r$.

This says that the expected value of the score that would be achieved by evaluating s for f and y equals the expected value of the score that would be achieved by evaluating s_0 for f and x. We refer to s as the error-corrected version of s_0.

Example 4. *Continuing Example 1 with the white noise observation model, consider a second alternative scoring rule,*

$$s_c(f,y) = \log\sigma + \frac{(y-\mu)^2 - c^2}{2\sigma^2}. \tag{7}$$

The expected score is

$$\mathrm{E}_y\{s_c(f,y)\} = \log\sigma + \frac{(\mu-\mu_0)^2 + \sigma_0^2}{2\sigma^2}, \tag{8}$$

which is minimized when $\mu = \mu_0$ and $\sigma = \sigma_0$, as desired. Thus, this scoring rule (7) *is proper under the white noise observation model and relative to the class of Normal distributions. Moreover, as expectations* (3) *and* (8) *are equal, this scoring rule* (7) *is unbiased for the logarithmic scoring rule* (1) *under the white noise observation model and relative to the class of Normal distributions.*

Our focus is probability forecasts, but it is appropriate to mention that Bowler (2008) provided an example of an unbiased scoring rule in the case of deterministic forecasts. He showed how a variance decomposition used by Ciach and Krajewski (1999) allows the mean squared error of a point forecast, $\hat{x}$, relative to the truth, x, to be estimated. If the observation is $y = x + w$ and the error, w, has zero mean and is uncorrelated with x then subtracting the error variance from the mean squared error of the forecast relative to the observation yields an unbiased estimate of the mean squared error relative to the truth:

$$\mathrm{E}_y\{(\hat{x}-y)^2\} - \mathrm{var}(w) = \mathrm{E}_x\{(\hat{x}-x)^2\}.$$

This motivates the scoring rule $s(\hat{x},y) = (\hat{x}-y)^2 - \mathrm{var}(w)$, which is unbiased for $s_0(\hat{x},x) = (\hat{x}-x)^2$ under the observation model just described.

Returning to probability forecasts, we would like scoring rules to be both proper and unbiased under the observation model. This is achieved by (and only by) scoring rules, such as the error-corrected logarithmic scoring rule (7), that are unbiased for proper scoring rules, as the following proposition shows.

Proposition 1. *A scoring rule is proper under observation model r and relative to a class, $\mathcal{P}$, of probability forecasts if and only if it is unbiased under r and relative to $\mathcal{P}$ for a scoring rule that is proper relative to $\mathcal{P}$.*

Proof. Let s_0 be proper relative to $\mathcal{P}$ and let s be unbiased for s_0 under r and relative to $\mathcal{P}$. Then

$$\begin{aligned}\mathrm{E}_y\{s(f,y)\} &= \mathrm{E}_x\{s_0(f,x)\}\\ &\geq \mathrm{E}_x\{s_0(p,x)\}\\ &= \mathrm{E}_y\{s(p,y)\}\end{aligned}$$

for all $f,p \in \mathcal{P}$ so that s is proper under r and relative to $\mathcal{P}$. Now let s be proper under r and relative to $\mathcal{P}$ and define $s_0(f,x) = \mathrm{E}_{y|x}\{s(f,y) \mid x\}$ to be the conditional expectation of $s(f,y)$ given x. Then $\mathrm{E}_y\{s(f,y)\} = \mathrm{E}_x\{s_0(f,x)\}$ so that s is unbiased for s_0 under r and relative to $\mathcal{P}$ and

$$\begin{aligned}\mathrm{E}_x\{s_0(f,x)\} &= \mathrm{E}_y\{s(f,y)\}\\ &\geq \mathrm{E}_y\{s(p,y)\}\\ &= \mathrm{E}_x\{s_0(p,x)\}\end{aligned}$$

for all $f,p \in \mathcal{P}$ so that s_0 is proper relative to $\mathcal{P}$. □

Scoring rules for which the inequality (4) in Definition 1 is strict when $f \neq p$ are called 'strictly proper'. We shall not dwell on this distinction but we can define scoring rules to be strictly proper under observation models by extending Definition 2 in a similar way, whence Proposition 1 remains true with 'proper' replaced by 'strictly proper' throughout.

We refer to scoring rules that are unbiased for proper scoring rules as 'error-corrected proper scoring rules' and we shall construct such scoring rules in section 4. There, we shall make use of the following, slightly stronger notion of bias. Let $\mathcal{X}$ denote the set of possible values of x (formally, the set of values, x, for which there exists a density $p \in \mathcal{P}$ with $p(x) > 0$).

Definition 4. *A scoring rule, s, is said to be everywhere unbiased for a scoring rule, s_0, under observation model r and relative to a class, $\mathcal{P}$, of probability forecasts if*

$$\mathrm{E}_{y|x}\{s(f,y) \mid x\} = s_0(f,x) \quad \text{for all } x \in \mathcal{X} \text{ and all } f \in \mathcal{P} \qquad (9)$$

when $y \mid x \sim r$.

This says that, for any fixed x, the expected score that would be achieved by evaluating s for f and y equals the actual score that would be achieved by evaluating s_0 for f and x.

Example 5. *Continuing Example 4, we have*

$$\begin{aligned}\mathrm{E}_{y|x}\{s_c(f,y) \mid x\} &= \log\sigma + \frac{(x-\mu)^2}{2\sigma^2} \\ &= -\log f(x)\end{aligned}$$

so that the error-corrected logarithmic scoring rule (7) *is everywhere unbiased for the logarithmic scoring rule under the white noise observation model and relative to the class of Normal distributions.*

The following proposition describes the connection between 'unbiased' and 'everywhere unbiased' scoring rules.

Proposition 2. *If a scoring rule, s, is everywhere unbiased for a scoring rule, s_0, under observation model r and relative to a class, $\mathcal{P}$, of probability forecasts then s is also unbiased for s_0 under r and relative to $\mathcal{P}$. If $\mathcal{P}$ includes deterministic forecasts (that is, point mass distributions) at all $x \in \mathcal{X}$ then the properties 'everywhere unbiased' and 'unbiased' are equivalent.*

Proof. Let $x \sim p \in \mathcal{P}$. If s is everywhere unbiased for s_0 then

$$\mathrm{E}_y\{s(f,y)\} = \mathrm{E}_x[\mathrm{E}_{y|x}\{s(f,y) \mid x\}] = \mathrm{E}_x\{s_0(f,x)\}$$

for all $f, p \in \mathcal{P}$, showing that s is also unbiased for s_0. Now let p be the distribution that places probability 1 at $x = x_0$ for some $x_0 \in \mathcal{X}$ so that $\mathrm{E}_x\{s_0(f,x)\} = s_0(f,x_0)$ and

$$\begin{aligned}\mathrm{E}_y\{s(f,y)\} &= \mathrm{E}_x[\mathrm{E}_{y|x}\{s(f,y) \mid x\}] \\ &= \mathrm{E}_{y|x_0}\{s(f,y) \mid x_0\}.\end{aligned}$$

Then, if s is unbiased for s_0, we have

$$\mathrm{E}_{y|x_0}\{s(f,y) \mid x_0\} = s_0(f,x_0)$$

for any $x_0 \in \mathcal{X}$, showing that s is also everywhere unbiased for s_0. □

2.3. Mean scores

Let s be unbiased for s_0 under an observation model so that $\mathrm{E}_y\{s(f,y)\} = \mathrm{E}_x\{s_0(f,x)\}$. In practice, the performance of a forecaster is summarized by the mean score,

$$\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s(f_i, y_i),$$

calculated for a set of forecasts and observations, $\{(f_i, y_i) : i = 1, \ldots, n\}$. Such a mean score is an unbiased estimate of the expected score, $\mathrm{E}_{f,y}\{s(f,y)\}$, taken over the joint (climatological) distribution of f and y. We would like $\bar{s}$ also to be an unbiased estimate of the expected score,

$\mathrm{E}_{f,x}\{s_0(f,x)\}$, that would be achieved by s_0 if there were no observation error. Usually, $\bar{s}$ is unbiased for $\mathrm{E}_{f,x}\{s_0(f,x)\}$. For example, if s is everywhere unbiased for s_0 and if the observation, y, and the forecast, f, are conditionally independent given the truth, x, then

$$\begin{aligned}\mathrm{E}_{f,y}\{s(f,y)\} &= \mathrm{E}_x[\mathrm{E}_{f,y|x}\{s(f,y) \mid x\}] \\ &= \mathrm{E}_x(\mathrm{E}_{f|x}[\mathrm{E}_{y|f,x}\{s(f,y) \mid f,x\} \mid x]) \\ &= \mathrm{E}_x(\mathrm{E}_{f|x}[\mathrm{E}_{y|x}\{s(f,y) \mid x\} \mid x]) \\ &= \mathrm{E}_x[\mathrm{E}_{f|x}\{s_0(f,x) \mid x\}] \\ &= \mathrm{E}_{f,x}\{s_0(f,x)\}.\end{aligned}$$

Conditional independence of y and f given x is usually true for numerical predictands and genuinely categorical predictands: if we know x then the observation model tells us the distribution of y; we do not need to know f too. There are situations, however, in which conditional independence does not hold, in which case the third equality above fails and $\bar{s}$ may be a biased estimate of $\mathrm{E}_{f,x}\{s_0(f,x)\}$. Such a situation can arise when categorical predictands are constructed by dichotomizing numerical predictands. For example, let x_t and y_t be the binary quantities that indicate whether x and y lie below a threshold. If y and the forecast are conditionally independent given x then it will typically be the case that y_t and the forecast are not conditionally independent given x_t. Whereas the value of x defines the distribution of the observation, the value of x_t usually does not. The forecast, therefore, can carry additional information about the distribution of the observation and so conditional independence breaks down. In applying the ideas in this paper, therefore, we should be willing to assume that the distribution of the observed value of the predictand would be known if we knew the true value of the predictand. Even if this assumption is reasonable for one predictand, it may be unreasonable for another predictand that is derived from the original predictand, for example by thresholding.

We return to error-corrected proper scoring rules in section 4. Before that, we review some approaches to observation error that have been proposed by other authors and examine whether or not they yield scoring rules that meet our requirements of being proper and unbiased.

3. Other approaches to observation error

3.1. Probabilistic observations

Several authors have proposed accounting for observation error by replacing the observation, y, with a probability distribution and then measuring the difference between this verifying distribution and the forecast distribution. In fact, there are several variations on this approach that differ in terms of what the verifying distribution represents and how the difference between the two distributions is measured.

For some authors, the verifying distribution represents the verifier's uncertainty about the truth, x, and may be constructed using whatever information is available when the forecast is verified. This is the situation envisaged by Weijs and van de Giesen (2011). Such a distribution may be thought of as a posterior predictive distribution for x and might be obtained from a probabilistic analysis or reanalysis, for example. For other authors, the verifying distribution is a distribution of observations. This might be formed from a collection of actual observations, as in Gorgas and Dorninger (2012) and Santos and Ghelli (2012). In contrast, Candille and Talagrand (2008), Pappenberger *et al.* (2009) and Pinson and Hagedorn (2012) form the verifying distribution by randomly perturbing an observation according to a probability model of observation error. This latter approach doubles the error in the observations and the efficacy of doing so is unclear.

Let g denote the verifying distribution, however constructed, and let f denote the forecast distribution for the truth. Some authors measure the difference between f and g by averaging a scoring rule over random draws from g. In other words, they calculate (either analytically or by Monte Carlo simulation) the expected scoring

rule, $\mathrm{E}_y\{s(f,y)\}$, when $y \sim g$. For example, the cross-entropy score proposed by Weijs and van de Giesen (2011) is the expected logarithmic scoring rule, whereas Pinson and Hagedorn (2012) use the expected continuous ranked probability scoring rule. Other authors measure the difference between f and g with a divergence, that is a function, $d(f,g)$, for which $d(g,g) = 0$ and $d(f,g) \geq 0$ for all f and g. For example, Candille and Talagrand (2008) use the quadratic divergence, $(f-g)^2$, in the case of forecasting a binary event (see also Santos and Ghelli, 2012), Pappenberger *et al.* (2009) use the Kullback-Leibler divergence (or relative entropy), $\int g \log(g/f)$, and Friederichs and Thorarinsdottir (2012) propose the integrated quadratic distance, $\int (f-g)^2$. Thorarinsdottir *et al.* (2013) list several other divergences, including the sub-class of 'score divergences' that are formed from proper scoring rules, s, in the following way:

$$d(f,g) = \mathrm{E}_y\{s(f,y)\} - \mathrm{E}_y\{s(g,y)\},$$

where $y \sim g$. Score divergences thus differ from expected (proper) scoring rules by subtracting an amount, $\mathrm{E}_y\{s(g,y)\}$, that is independent of the forecast.

Divergences and expected proper scoring rules are both optimized when $f = g$. If truth really were a distribution, g, rather than a constant, x, and if we could measure g without error, then divergences and expected proper scoring rules might be appropriate ways of measuring forecast performance. If we believe that the truth is not a distribution, however, then the approaches described above are typically inappropriate. For example, a perfect forecast that assigns probability 1 to the truth, x, would receive a worse score than the forecast $f = g$. One might argue that the verifying distribution, g, is as much as we can know about the truth, and that forecast distributions that are more precise than g should be penalized. Over time, however, beliefs about the truth and, therefore, the verifying distribution are likely to change, because more data become available, scientific understanding improves, numerical models develop etc. Such changes in the verifying distribution may change the measured performance of the forecasts, and, as Bowler (2006) and Candille and Talagrand (2008) note, we would rather measure performance in a way that is insensitive to the quality of our observations.

Divergences and expected proper scoring rules also fail to encourage forecasters to issue their honest beliefs as their forecasts. For example, there is no reason that the forecast that optimizes the expected divergence, $\mathrm{E}_g\{d(f,g)\}$, where the expectation is over the forecaster's belief about what g will be when the forecast is verified, should equal the forecaster's belief about x.

Verifying distributions appear to be unfit for purpose for the reasons outlined above. We close this section by mentioning that Bowler *et al.* (2015) describe a situation in which an alternative to the observation is available whose properties ensure that the mean score achieved by the forecast equals the mean score that would be achieved by verifying against truth. They present their result only when the scoring rule is the squared error of a deterministic forecast, however, and the conditions required for their result to hold are restrictive (although not always unrealistic). As we are seeking a class of scoring rules for probability forecasts, we do not consider their approach any further.

3.2. *Deconvolving observation and error distributions*

Another approach to accounting for observation error seeks to estimate the joint distribution of the forecast, f, and the truth, x. Estimates of quantities such as the long-run expected value, $\mathrm{E}_{f,x}\{s(f,x)\}$, of a scoring rule can then be derived from this joint distribution. The approach is as follows. First, construct estimates of the conditional distributions of the observations given the forecasts, that is of $\pi(y \mid f)$ for all f. Then apply a deconvolution algorithm to estimate the conditional distributions, $\pi(x \mid f)$, of the truth given the forecasts, where

$$\pi(y \mid f) = \int \pi(x \mid f) r(y \mid f, x) dx$$

and $r(y \mid f, x)$ denotes the conditional density of the observation given the forecast and the truth. This latter distribution is assumed to be known. Finally, combine the conditional distributions, $\pi(x \mid f)$, with an estimate of the marginal distribution of the forecasts to obtain an estimate of the joint distribution, $\pi(f, x)$, of the forecasts and the truth.

Briggs *et al.* (2005) and Bowler (2006) propose this approach for deterministic forecasts of binary events, where the forecasts are either 0 or 1. Briggs *et al.* (2005) assume that the observation is conditionally independent of the forecast given the truth, in which case $r(y \mid f, x) = r(y \mid x)$ is our familiar observation model, whereas Bowler (2006) assumes that the observation *error* is conditionally independent of the truth given the forecast. In both cases, the conditional distributions, $\pi(y \mid f)$, are estimated by fitting probability distributions to the observations for which the corresponding forecasts are 0 or 1, and the marginal distribution of the forecasts is estimated using the proportions of forecasts equal to 0 and 1. The joint distribution of the forecasts and the truth is then used to estimate the expected entries in a contingency table, from which various scores are calculated. Bowler (2008) uses the same idea to adjust Relative Operating Characteristic curves for observation error.

This deconvolved observation approach is suitable if one wishes only to estimate average quantities such as $\mathrm{E}_{f,x}\{s(f, x)\}$. An advantage is that the joint distribution of the forecasts and the truth for any dichotomized events may be derived from $\pi(f, x)$ and so the limitation discussed in section 2.3 does not apply. On the other hand, this approach may not yield an unbiased estimate of $\mathrm{E}_{f,x}\{s(f, x)\}$ and may be difficult to implement if a continuum of forecasts can be issued. Moreover, the approach does not provide a way of awarding a score to each individual forecast and so does not provide a means of encouraging forecasters to issue their honest belief as the forecast on any given occasion.

3.3. Convolving forecast and error distributions

A third approach to accounting for observation error is to add error to the forecast by convolving the forecast distribution with the observation model. If f is the density forecast for x and r is the observation model then the implied density forecast for y is the convolution

$$(f * r)(y) = \int r(y \mid x) f(x) dx \tag{10}$$

and this can be scored with a proper scoring rule.

Proposition 3. *Let r be an observation model, let $\mathcal{P}$ be a class of probability forecasts, and let $\mathcal{P}_r$ be the class of distributions formed by convolving members of $\mathcal{P}$ with r. If a scoring rule, s_0, is proper relative to $\mathcal{P}_r$ then the scoring rule*

$$s(f, y) = s_0(f * r, y)$$

is proper under r and relative to $\mathcal{P}$.

Proof. Let $x \sim p \in \mathcal{P}$ so that $y \sim p * r$. Then

$$\begin{aligned} \mathrm{E}_y\{s(f, y)\} &= \mathrm{E}_y\{s_0(f * r, y)\} \\ &\geq \mathrm{E}_y\{s_0(p * r, y)\} \\ &= \mathrm{E}_y\{s(p, y)\} \end{aligned}$$

for all $f, p \in \mathcal{P}$. □

Example 6. *Recall Example 1 in which the true distribution of x is $N(\mu_0, \sigma_0^2)$, the forecast distribution for x is $N(\mu, \sigma^2)$ and the conditional distribution for y given x specified by the observation model is $N(x, c^2)$. The convolved forecast distribution for y is then $N(\mu, \sigma^2 + c^2)$ and, taking $s_0(f, y) = -\log f(y)$ to be the logarithmic scoring rule, we find that $s_0(f * r, y)$ is the error-convolved logarithmic scoring rule* (5)*. We noted in Example 3 that this scoring rule is proper under the white noise observation model and relative to the class of Normal distributions, but that it is not unbiased for s_0.*

This error-convolved approach was first proposed by Anderson (1996), who added observation error to ensemble members before forming rank histograms. See also Hamill (2001). Saetra *et al.* (2004) also proposed convolving the forecast and error distributions, and investigated the impact on both rank histograms and reliability diagrams. In a similar vein, Candille *et al.* (2007) assessed bias and dispersion using the residual $(y-\mu)/\sqrt{\sigma^2+c^2}$ rather than $(y-\mu)/\sigma$, where μ is the forecast mean, σ^2 is the forecast variance, and c^2 is the observation error variance. This inflates the forecast variance to account for an additive observation error. Bröcker and Smith (2007) explicitly proposed the scoring rule given in Proposition 3 and the same idea is part of the Bayesian approach proposed by Röpnack *et al.* (2013).

Candille and Talagrand (2008) compared the error-convolved approach and the probabilistic observation approach of section 3.1 and showed that neither yields unbiased estimates of the performance that would be achieved were there no observation error. So this approach yields scoring rules that are proper, but not usually unbiased, in the presence of observation error.

4. Error-corrected proper scoring rules

4.1. Categorical predictands

We would like to construct scoring rules that are both proper and unbiased under the observation model. Such scoring rules will encourage forecasters to issue their honest belief as the forecast and will yield an unbiased estimate of the score that the forecasts would receive were we able to verify them against the truth. In this sense, the scores will be insensitive to the quality of the observations. In light of Proposition 1, therefore, we investigate how to construct error-corrected proper scoring rules, that is those that are unbiased for proper scoring rules under an observation model. We consider, first, probability forecasts of categorical predictands, where the number of categories is finite. Examples include whether or not a tornado occurred, or the type of synoptic weather regime.

Let x and y take values in the set $\mathcal{X}=\{1,2,\ldots,k\}$ and let $\mathcal{P}$ be the set of all probability distributions on $\mathcal{X}$, that is the set of all vectors $(p_1,\ldots,p_k)$ for which $p_1+\ldots+p_k=1$ and $p_i\geq 0$ for $i=1,\ldots,k$. An observation model, r, defines the misclassification probabilities $r_{b|a}=\Pr(y=b\mid x=a)$ for $a,b\in\mathcal{X}$. In this setting, the properties 'unbiased' and 'everywhere unbiased' are equivalent by Proposition 2 and the conditions (9) required for the scoring rule s to be unbiased for the scoring rule s_0 under r and relative to $\mathcal{P}$ may be written as

$$\mathbf{RS}=\mathbf{S}_0 \qquad \text{for all } f\in\mathcal{P}, \tag{11}$$

where $\mathbf{R}$ is the $k\times k$ matrix whose (a,b)th element is $r_{b|a}$, $\mathbf{S}=(s(f,1),\ldots,s(f,k))^T$ and $\mathbf{S}_0=(s_0(f,1),\ldots,s_0(f,k))^T$. If $\mathbf{R}$ is invertible then the unique scoring rule, s, that is unbiased for s_0 under r and relative to $\mathcal{P}$ is defined by

$$\mathbf{S}=\mathbf{R}^{-1}\mathbf{S}_0. \tag{12}$$

As long as $\mathbf{R}$ is invertible, therefore, we can use this formula to construct scoring rules that are unbiased for proper scoring rules in the presence of observation error just by choosing s_0 to be proper. We describe situations in which $\mathbf{R}$ is singular later in this section.

Let us consider this general construction (12) in more detail for the special case of binary predictands. Changing notation slightly, let $\mathcal{X}=\{0,1\}$ and denote the misclassification probabilities by $r_0=\Pr(y=1\mid x=0)$ and $r_1=\Pr(y=0\mid x=1)$. Here, $\mathbf{R}$ is invertible if and only if $r_0+r_1\neq 1$, in which case the scoring rule, s, that is unbiased for s_0 under r and relative to $\mathcal{P}$ is

$$s(f,y)=s_0(f,y)+\frac{r_y\{s_0(f,y)-s_0(f,1-y)\}}{1-r_0-r_1}. \tag{13}$$

If $\mathbf{R}$ is singular then no scoring rule is unbiased for s_0 under r and relative to $\mathcal{P}$ unless s_0 is trivial, in the sense that $s_0(f,x)$ is independent of x.

We saw in Example 1 that proper scoring rules favour good forecasts of y rather than of x. We can obtain a general expression for this effect in the binary case. If $\Pr(x=1)=p$ defines the true distribution of x then the true distribution of y is given by $\Pr(y=1)=q$, where

$$\begin{aligned} q &= \Pr(y=1 \mid x=1)\Pr(x=1) \\ &\quad + \Pr(y=1 \mid x=0)\Pr(x=0) \\ &= (1-r_1)p + r_0(1-p). \end{aligned}$$

Hence, $q=p$ for all p if and only if there is no observation error ($r_0=r_1=0$). If $s_0(f,y)$ is proper, its expectation is optimized when $f=q$ rather than when $f=p$. The bias in the optimal value of f that is caused by observation error thus varies linearly from r_0 when $p=0$ to $-r_1$ when $p=1$ and is zero only when $p=r_0/(r_0+r_1)$. In other words, observation error biases the optimal forecast towards $r_0/(r_0+r_1)$.

Now let us consider the impact of observation error on the value of the score by comparing the expected values of $s_0(f,y)$, the error-convolved $s_0(f*r,y)$ and the error-corrected $s(f,y)$ for a popular, proper scoring rule, s_0, as we did for the logarithmic scoring rule and white noise observation error in sections 1 and 2.

Example 7. *The quadratic scoring rule (Brier, 1950) is* $s_0(f,y)=(f-y)^2$. *If we ignore observation error then our expected score is*

$$\mathrm{E}_y\{s_0(f,y)\} = (f-q)^2 + q(1-q).$$

If we account for observation error by using the error-convolved quadratic scoring rule, $s_0(f*r,y)=(g-y)^2$, *where* $g=(1-r_1)f+r_0(1-f)$, *then our expected score is*

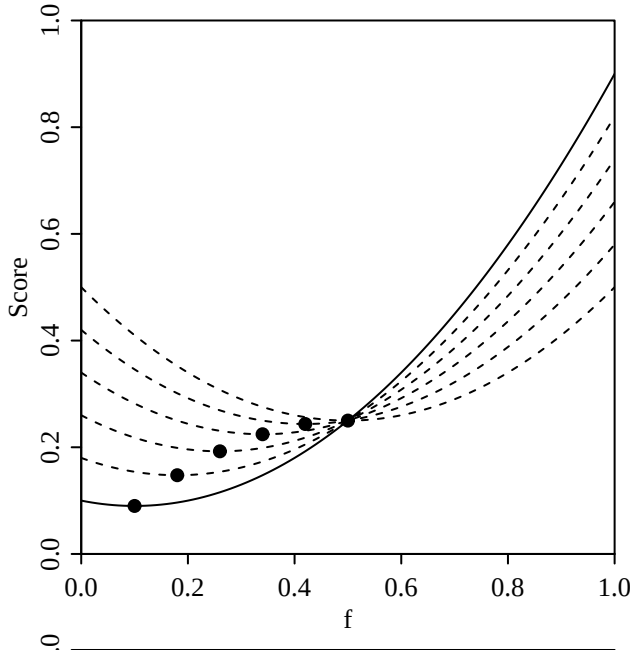
$$\mathrm{E}_y\{s_0(f*r,y)\} = (1-r_0-r_1)^2(f-p)^2 + q(1-q).$$

If we account for observation error by using the error-corrected scoring rule (13) *then our expected score is*

$$\mathrm{E}_y\{s(f,y)\} = (f-p)^2 + p(1-p).$$

All three expectations are equal if there is no observation error, but only the last expectation is unaffected by observation error. We plot the expectations as functions of f *for various values of* r_0 *and* r_1 *when* $p=0.1$ *in Figure 3. The solid lines are the expected scores when there is no observation error and these are minimized at* $f=p$. *The dashed lines in the upper panel are graphs of* $\mathrm{E}_y\{s_0(f,y)\}$ *for different degrees of observation error. These show that* $s_0(f,y)$ *is biased for the score that would be obtained without observation error and that observation error biases the optimal forecasts (indicated by the dots) away from* p *and towards* $r_0/(r_0+r_1)=0.5$. *The dashed lines in the lower panel are graphs of* $\mathrm{E}_y\{s_0(f*r,y)\}$. *These show that* $s_0(f*r,y)$ *is also biased for the score that would be obtained without observation error but that the optimal forecasts are now correct. The graphs of* $\mathrm{E}_y\{s(f,y)\}$ *coincide with the solid lines whatever the degree of observation error, which means that* $s(f,y)$ *is unbiased for the score that would be obtained without observation error and yields the correct optimal forecasts.*

We compare the actual scores, $s_0(f,y)$, $s_0(f*r,y)$ *and* $s(f,y)$, *rather than their expectations, by plotting them as functions of* f *for* $y=0$ *and* $y=1$ *in Figure 4. We see that the original and error-convolved scores,* $s_0(f,y)$ *and* $s_0(f*r,y)$, *always lie between* 0 *and* 1, *but that, in accounting for the observation error, it is necessary for the error-corrected score sometimes to lie below* 0 *or above* 1. *In finite samples, therefore, it is possible for the mean error-corrected score to be negative. Such values might be truncated at* 0 *when reporting a mean score, but truncating* $s(f,y)$ *itself would mean that it were no longer unbiased and proper.*



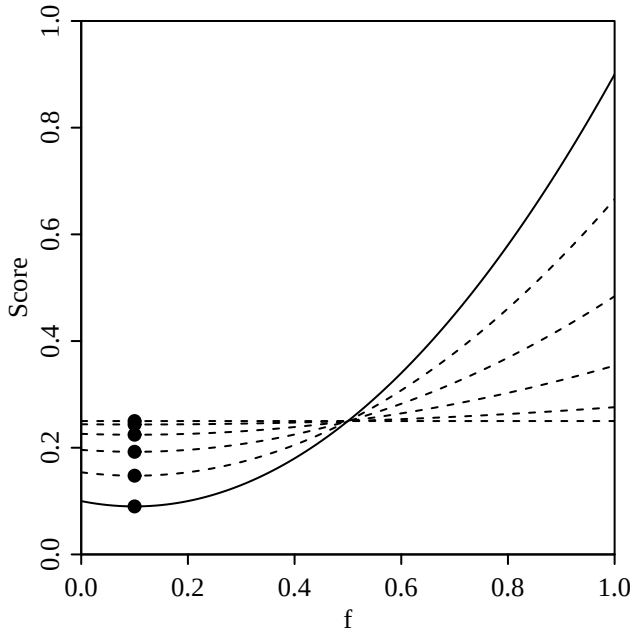
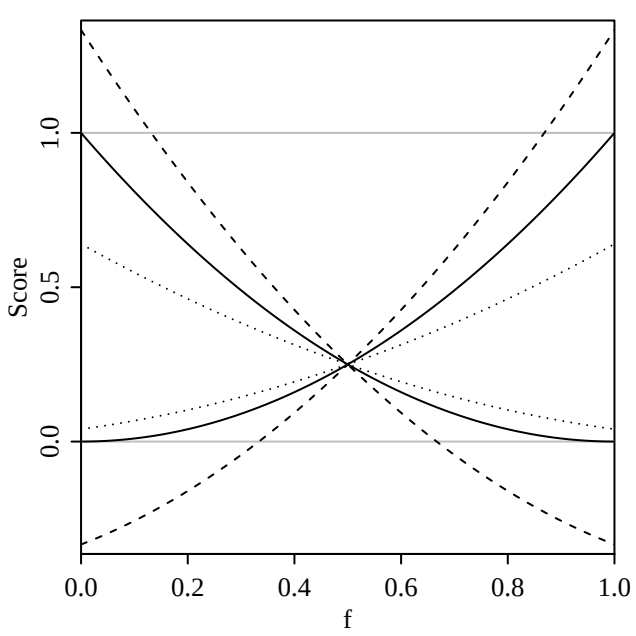


Figure 3. Expected values of the quadratic scoring rule (top) and error-convolved quadratic scoring rule (bottom) plotted against the forecast probability, f, when the true probability is $p = 0.1$, when there is no observation error ($r_0 = r_1 = 0$, solid lines) and when the misclassification probabilities are $r_0 = r_1 = 0.1, 0.2, \ldots, 0.5$ (dashed lines with increasing intercepts). The expected values of the error-corrected quadratic scoring rule coincide with the solid lines for all r_0 and r_1. The dots (•) mark the minima.



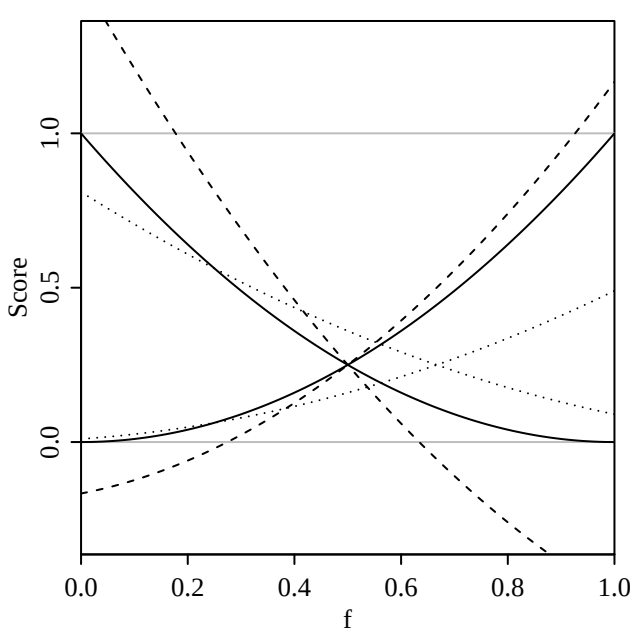


Figure 4. Graphs of the quadratic scoring rule, $s_0(f, y)$ (solid lines), the error-convolved quadratic scoring rule, $s_0(f * r, y)$ (dotted lines), and the error-corrected quadratic scoring rule, $s(f, y)$ (dashed lines), for $y = 0$ and $y = 1$ when $r_0 = r_1 = 0.2$ (top) and when $r_0 = 0.1$ and $r_1 = 0.3$ (bottom).

Qualitatively similar graphs and comments obtain for other scoring rules such as the logarithmic and pseudo-spherical scoring rules (not shown).

4.2. Numerical predictands

Now we consider constructing error-corrected proper scoring rules for probability forecasts of scalar numerical predictands, such as temperatures and numbers of hurricanes. We seek scoring rules, s, that are everywhere unbiased for proper scoring rules, s_0, and so satisfy the conditions (9) in Definition 4.

We consider additive and multiplicative errors, w, in turn. For additive errors, $y = x + w$ and we assume that the mean error is linear in x, so that

$$\mathrm{E}(y \mid x) = a + bx \tag{14}$$

for known constants a and b, and the error variance is constant, so that

$$\mathrm{var}(y \mid x) = c^2 \tag{15}$$

for a known constant c. Setting $a = 0$ and $b = 1$ yields additive errors with zero mean. We make no other assumptions about the shape of the error distribution and so this observation model will be adequate in many situations. Even if we knew that the error mean and variance depended on x in more complicated ways, this model may still be an acceptable approximation. The following example gives an error-corrected proper scoring rule for all observation models of this form.

Example 8. *Let $\mathcal{P}$ be the class of probability distributions that have finite variance, and let μ and σ denote the mean and standard deviation of a forecast, $f \in \mathcal{P}$. With no observation error, the scoring rule*

$$s_0(f, x) = \log \sigma + \frac{(x - \mu)^2}{2\sigma^2} \tag{16}$$

is proper relative to $\mathcal{P}$ (Dawid and Sebastiani, 1999; Gneiting and Raftery, 2007). For our additive observation model (14, 15)*, the scoring rule*

$$s(f, y) = \log \sigma + \frac{(y - a - b\mu)^2 - c^2}{2b^2\sigma^2} \tag{17}$$

satisfies the conditions (9) *for it to be everywhere unbiased for s_0 under this observation model and relative to $\mathcal{P}$.*

For multiplicative errors, $y = xw$ and we assume that $\mathrm{E}(y \mid x) = bx$ and $\mathrm{var}(y \mid x) = c^2x^2$ for known constants b and c.

Example 9. *Continuing the previous example, the scoring rule*

$$s(f, y) = \log \sigma + \frac{(y - b\mu)^2 - y^2c^2/(b^2 + c^2)}{2b^2\sigma^2} \tag{18}$$

is everywhere unbiased for the Dawid-Sebastiani scoring rule (16) *under our multiplicative observation model and relative to $\mathcal{P}$.*

In these examples, the class, $\mathcal{P}$, relative to which s is unbiased for s_0 is the same class (denoted here as $\mathcal{P}_0$) relative to which s_0 is proper. Such parity is not always possible because the expectation in the conditions (9) of Definition 4 may not exist for some distributions, f, in $\mathcal{P}_0$. Error-corrected scoring rules, therefore, are often unbiased relative to only a subset of $\mathcal{P}_0$, and this subset may depend on the scoring rule, s_0, and the observation model.

Such restrictions on $\mathcal{P}$ are undesirable because they limit the scope of the error-corrected proper scoring rule. If the true distribution of x (representing the forecaster's honest belief about x) lies outside $\mathcal{P}$ then the scoring rule fails to encourage the forecaster to issue the true distribution as the forecast. Secondly, if $s(f, y)$ is evaluated for forecasts that lie outside $\mathcal{P}$ and for which the expectation (9) does not exist then scores may be volatile and will no longer be unbiased estimates of the scores that would be obtained from $s_0(f, x)$.

Knowledge of $\mathcal{P}$ is thus important. Identifying $\mathcal{P}$ for error-corrected scoring rules, however, can require detailed analysis. It is possible to obtain general formulae for error-corrected versions of other popular scoring rules, such as the logarithmic, quadratic, pseudo-spherical and continuous ranked probability scoring rules (e.g. Gneiting and Raftery, 2007), for various observation models, but establishing the classes of distributions relative to which they are proper and unbiased requires significant effort. For this reason, we leave the development of other error-corrected proper scoring rules to future work and propose using, in the meantime, the widely applicable error-corrected Dawid-Sebastiani scoring rules (17, 18).

4.3. Rounded observations

We have mentioned that there are situations in which error-corrected scoring rules may not exist. In the case of categorical predictands, for example, the system of equations (11) that defines the error-corrected scoring rule typically has no solution if $\mathbf{R}$ is singular. As Briggs *et al.* (2005) note, we might hope that realistic observation

models will yield invertible $\mathbf{R}$ in the binary case because $r_0 + r_1 < 1$ is guaranteed if the probability of observing $y = 1$ is greater when $x = 1$ than when $x = 0$. More generally, however, there are feasible observation models for which it is usually impossible to find error-corrected scoring rules. For example, if there are three categories and

$$\mathbf{R} = \begin{pmatrix} .6 & .4 & 0 \\ .3 & .4 & .3 \\ 0 & .4 & .6 \end{pmatrix}$$

then the system of equations is inconsistent and there is no error-corrected scoring rule unless $2s_0(f, 2) = s_0(f, 1) + s_0(f, 3)$, in which special case there would be infinitely many error-corrected scoring rules.

An important situation in which error-corrected scoring rules usually do not exist is when the observation is rounded, for example if the observation y is recorded whenever x lies in some interval containing y. Rounding yields singular $\mathbf{R}$. More generally, the conditions (9) that must hold for s to be everywhere unbiased for s_0 require $s(f, y) = s_0(f, x)$ for all values of x that would be rounded to y. In other words, s_0 must be constant over such sets of values of x so that s_0 must, effectively, already be defined with respect to rounded observations. All digitally recorded observations are rounded to some degree, so what are the implications for how we score forecasts?

If rounding is the only observation error present then perhaps the best we can do, in the absence of error-corrected scoring rules, is to find a scoring rule that is proper under rounding. This can be done by following section 3.3 and applying the same rounding to the forecast distribution. For example, if any true value, x, in the interval $[y - \delta, y + \delta)$ is rounded to the observed value y then apply the convolution (10) with $r(y \mid x) = I(y - \delta \leq x < y + \delta)$, where $I(A) = 1$ if A is true and $I(A) = 0$ otherwise. In the absence of any information about x at a higher resolution than is available after rounding, we should essentially verify the rounded forecast with a scoring rule, s, that is proper relative to a class of probability forecasts on the space of rounded values of x. We cannot hope to obtain an unbiased estimate of the score that would be achieved were we able to verify the forecasts against the unrounded truth.

If rounding is not the only observation error present then we should find a scoring rule that is unbiased for the proper scoring rule, s, described in the previous paragraph. This should be achievable by applying the ideas outlined earlier in this section with the truth, x, assumed to take values in the space of rounded values.

We leave an exploration of the practical impact of rounding to future work. For the rest of this paper, we follow common practice by assuming that there is no rounding and focus instead on the effects of other sources of observation error.

5. Data examples

5.1. *Continuous predictand*

We apply our scoring rules to forecasts constructed for an artificial data set in order to illustrate the ideas discussed above. We generate truth, x, from a $N(0, \sigma_0^2)$ distribution and generate perfect ensemble members, $z_1, \ldots, z_m$, from the same distribution. The correlations, ρ, between pairs of ensemble members and between each ensemble member and the truth are all equal. The observation error is white noise so that $y \mid x \sim N(x, c^2)$. We set $\sigma_0 = c = 2$, $\rho = 0.8$ and $m = 10$ to reflect typical values for short-range operational ensembles for surface temperatures (Bowler, 2006) and we generate a sample of $n = 300$ days.

We form probability forecasts by post-processing the ensembles using non-homogeneous Gaussian regression (NGR: see Gneiting *et al.*, 2005). For n observations, $\{y_i : i = 1, \ldots, n\}$, and corresponding ensembles, the standard approach is to model the conditional distribution of y_i given the ensemble as

$$N(\alpha + \beta \bar{z}_i, \gamma^2 + \delta^2 s_i^2), \tag{19}$$

where $\bar{z}_i$ is the sample mean of the ith ensemble, s_i^2 is the sample variance of the ith ensemble and α, β, γ and δ are parameters to be estimated. One way to estimate the parameters is to minimize the logarithmic score,

$$-\frac{1}{n}\sum_{i=1}^{n}\log f_i(y_i) = \frac{1}{2n}\sum_{i=1}^{n}\log(\gamma^2+\delta^2 s_i^2) + \frac{1}{2n}\sum_{i=1}^{n}\frac{(y_i-\alpha-\beta\bar{z}_i)^2}{\gamma^2+\delta^2 s_i^2}, \tag{20}$$

where f_i denotes the density of the NGR model (19). This standard approach is designed to produce forecasts of observed values, y, rather than of true values, x. If we wish to produce forecasts of true values then we should account for observation error when post-processing ensembles. To do this, we adopt the NGR model (19) as the conditional distribution of the *truth* given the ensemble and then derive the conditional distribution of y_i as

$$N(\alpha+\beta\bar{z}_i, \gamma^2+\delta^2 s_i^2+c^2). \tag{21}$$

This is the convolution of the forecast distribution (19) and the error distribution, as in Example 6. The parameters may now be estimated by minimizing the error-convolved logarithmic score,

$$\frac{1}{2n}\sum_{i=1}^{n}\log(\gamma^2+\delta^2 s_i^2+c^2) + \frac{1}{2n}\sum_{i=1}^{n}\frac{(y_i-\alpha-\beta\bar{z}_i)^2}{\gamma^2+\delta^2 s_i^2+c^2}. \tag{22}$$

Unless c is a relatively small contributor to the variance in the model distribution (21), this approach sometimes produces estimates of γ and δ, and hence of the forecast variance, that are too small. This is the cause of the relatively large sampling variation in the corresponding logarithmic scores reported later in Table I. Other approaches may yield better estimates but we do not try them here. Note that we should not use the error-corrected logarithmic score,

$$\frac{1}{2n}\sum_{i=1}^{n}\log(\gamma^2+\delta^2 s_i^2) + \frac{1}{2n}\sum_{i=1}^{n}\frac{(y_i-\alpha-\beta\bar{z}_i)^2-c^2}{\gamma^2+\delta^2 s_i^2},$$

to fit the NGR model because we need to reflect the information content of the data (the observations) rather than the information that we would have if we had access to the true values of the predictand. Note also that the three logarithmic scores in this example equal the original, error-convolved and error-corrected Dawid-Sebastiani scores of section 4.2 because the forecasts are Normal distributions and we omit the term $\log(2\pi)/2$.

We fit our NGR model to the full data set and do not attempt to form out-of-sample forecasts as our purpose is merely illustrative. We compare the performance of the forecasts obtained by fitting the NGR model with (22) and without (20) accounting for observation error. For each set of forecasts, we evaluate the logarithmic score and the error-corrected logarithmic score using the observed values. As we know the true values of the predictand in this example, we also calculate the logarithmic score using the true values. Results are in Table I. For both forecasts, we find that the logarithmic score evaluated for the observations overestimates the logarithmic score evaluated for the truth, but that the error-corrected logarithmic score provides accurate estimates of the latter. We also find that the forecasts obtained by ignoring observation error are better as forecasts of the observations, but that the forecasts obtained by accounting for observation error are better as forecasts of the truth. If we want good forecasts of the truth, rather than of the observations, then we need to compare the error-corrected logarithmic scores in order to avoid being misled into preferring the standard forecasts.

Table I. The mean logarithmic score evaluated for the observations, $s_0(f,y)$, and the truth, $s_0(f,x)$, and the mean error-corrected logarithmic score, $s(f,y)$, evaluated for the observations. Scores are for forecasts obtained with (Adjusted) and without (Standard) accounting for observation error. Estimated standard errors are about 0.04 for the Standard forecasts, 0.16 for the Adjusted forecasts and 0.12 for the differences.

	$s_0(f,y)$	$s(f,y)$	$s_0(f,x)$
Standard forecasts	1.33	0.95	0.91
Adjusted forecasts	2.22	0.61	0.44
Difference	−0.89	0.34	0.47

Similar conclusions apply for other ensemble sizes, m, and sample sizes, n (not shown). Reducing the observation

Table II. The frequency with which each of six probabilities was issued as a forecast of tornado occurrence, and the frequency with which at least one tornado was subsequently observed.

Probability (%)	1	5	25	50	75	95
Forecast frequency	2	22	49	68	22	3
Tornado frequency	0	2	9	32	14	3

error variance, c^2, shrinks the differences between the scores of the two forecasts, and shrinks the differences between the original and error-corrected scores for each forecast, but all differences remain statistically significant even for small values of c (not shown). This indicates that accounting for observation error can be important even when the errors are small.

5.2. *Binary predictands*

Now we illustrate the impact of observation error on the scores of two sets of probability forecasts for binary predictands: occurrences of tornadoes and of aircraft icing. Our first example uses data from an experiment reported by Vescio and Thompson (2001). Probability forecasts were made for the event of at least one tornado occurring in the area of each of 166 severe weather watches issued across the United States during 1997 and 1998. The forecasts and corresponding observations (reconstructed from Figures 2 and 5 of Vescio and Thompson, 2001) are shown in Table II. If we assume that there is no observation error then the mean quadratic score for these forecasts is 0.19 with standard error 0.01.

The dominant observation error associated with tornadoes is under-reporting owing to limited observers or radar coverage. This suggests setting $r_0 = 0$ in our observation model and estimating r_1. Vescio and Thompson (2001) give no information about the possible magnitude of under-reporting for their data but other authors have estimated the probability of failing to observe US tornadoes. Ray *et al.* (2003) estimate the probability to be about 0.4 in the 1980s. Anderson *et al.* (2007) find that the probability varies geographically between about 0 and 0.6 for the period 1950–2000, with evidence that the probability is lower later

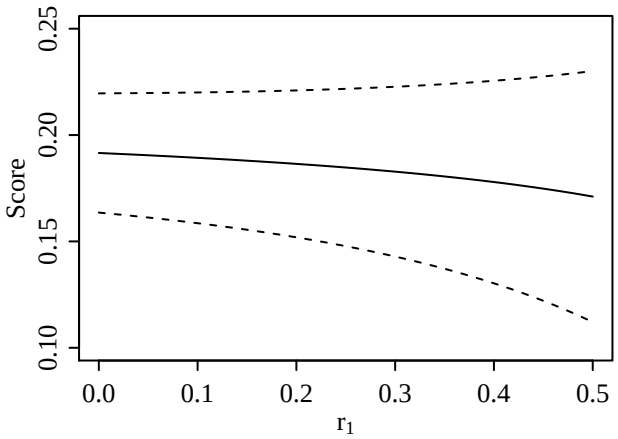


Figure 5. Graph of the mean error-corrected quadratic score (solid line) with approximate 95% confidence intervals (dashed lines) against the observation error probability, r_1, for the tornado forecasts.

in the period. Elsner *et al.* (2013) estimate the probability to be less than 0.4 in the decade around 1997–98. With these values for the probability of observing individual tornadoes, and noting that more than one tornado could occur during any watch, we might expect r_1 to be somewhat less than 0.4 for our data set.

We would prefer to have a more precise estimate of r_1, which could even be different for each observation. The studies just cited testify to the difficulty of forming such estimates, however, and our purpose is not to establish definitive estimates but to illustrate the potential impact of observation error on forecast scores. We show in Figure 5, therefore, how the mean error-corrected quadratic score varies as r_1 increases from 0 to 0.5. This shows how our choice for the value of r_1 would affect our estimate of the quadratic score that would be obtained if we had perfect observations. The score improves slightly from 0.19 when $r_1 = 0$ to 0.17 when $r_1 = 0.5$, but the change is small relative to the sampling variation and so we find that observation error has little impact in this example.

Our second example uses data from an experiment reported by Brown *et al.* (1999). Probability forecasts were made for the occurrence of aircraft icing conditions in six regions of the US during the winters of 1996–97 and 1997–98. Pilot reports (PIREPs) of icing were used as the observations. The 1242 forecasts and observations are shown in Table III and are available in the `verification`

Table III. The frequency with which each of 13 probabilities was issued as a forecast of icing occurrence, and the frequency with which icing was subsequently observed.

Probability (%)	2	5	10	20	30
Forecast frequency	120	101	139	159	156
Icing frequency	4	7	14	28	39
Probability (%)	40	50	60	70	80
Forecast frequency	158	152	109	84	50
Icing frequency	66	73	78	61	43
Probability (%)	90	95	98		
Forecast frequency	11	2	1		
Icing frequency	9	2	1		

package (Gilleland, 2015) of the R statistical programming environment (R Core Team, 2015). If we assume that there is no observation error then the mean quadratic score for these forecasts is 0.16 with standard error 0.005.

Observation errors are common in icing PIREPs (e.g. Briggs *et al.*, 2005) and so the six regions of the study were centred on large cities in an attempt to reduce the chance of such errors arising. We shall, nonetheless, examine how the mean score varies with r_0 and r_1 for illustrative purposes. As before, it is difficult to obtain good estimates of typical error probabilities but Briggs *et al.* (2005) estimate both r_0 and r_1 to be about 0.2 in another study of icing forecasts. We allow both r_0 and r_1 to vary between 0 and 0.5. Figure 6 shows that the mean error-corrected quadratic score improves as either r_0 or r_1 increases, and changes more quickly with r_0 than with r_1. The changes are large compared to the standard errors, which are about 0.01, and so observation error could have a large impact in this example. (The greater sensitivity of the score to r_0 than to r_1 is due to the term r_y in the error-corrected score (13) and to the statistics of the forecasts and observations. There are more observations of $y = 0$ than of $y = 1$ and the magnitude of the term $s_0(f, y) - s_0(f, 1 - y)$ tends to be bigger when $y = 0$ than when $y = 1$.)

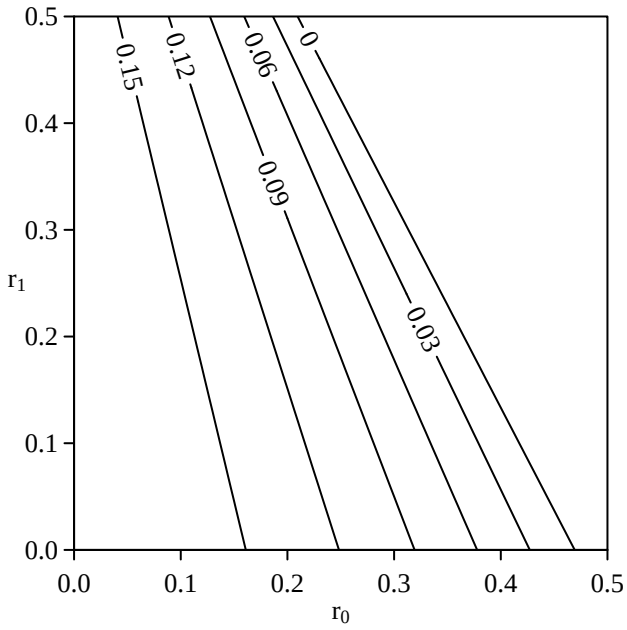


Figure 6. Contours of the mean error-corrected quadratic score against the observation error probabilities, r_0 and r_1, for the icing forecasts.

6. Summary and discussion

We showed that, in the presence of observation error, proper scoring rules favour good forecasts of the observations rather than of the truth, and yield scores that vary with the quality of the observations. We introduced error-corrected proper scoring rules to overcome these problems. We provided a general method for constructing these scoring rules in the case of categorical predictands and proposed error-corrected versions of the Dawid-Sebastiani scoring rule in the case of numerical predictands.

We noted that there are situations in which error-corrected scoring rules do not exist. In such circumstances, we can fall back on existing approaches to observation error. If we need a scoring rule that is proper (but not unbiased) in the presence of observation error, for example if our primary goal is to encourage forecasters to issue as their forecast their honest belief about the truth, then we can use the error-convolved scoring rules reviewed in section 3.3. If we want to estimate the score that would be obtained if we had perfect observations, but do not require a scoring rule that is proper or unbiased, then we can use the deconvolved observation approach reviewed in section 3.2.

This paper has proposed a framework for handling observation error when measuring the performance of forecasts but more work is needed. A rigorous treatment of the theory and a fuller characterisation of the new scoring rules, including the classes of distributions relative

to which they are proper and unbiased, would be valuable. Further investigation of the sensitivity of scoring rules to observation error, for example along the lines of the analysis in the appendix, might help to identify situations in which the effects of observation errors are negligible. Some extensions are also possible. For example, if we have more than one observation of the predictand, as mentioned in section 3.1, then we may evaluate an error-corrected proper scoring rule for each observation and average the resulting scores. This mean score will also be proper and unbiased. If we have an ensemble forecast rather than a probability forecast then we can form error-corrected versions of the fair scoring rules of Ferro (2014).

Our data examples illustrated the potential benefits of accounting for observation error, both when measuring forecast performance and when forming probability forecasts by post-processing ensembles. Accounting appropriately for observation error helps to produce better forecasts of the truth and to ensure that we favour forecasters who issue better forecasts of the truth. Failing to account for the effects of observation error when deciding between two forecasting systems, for example, could lead to the wrong choice and a high opportunity cost. Estimating the distribution of observation errors is difficult, but, as the gap between forecast error and observation error narrows, the value of good error estimates increases. The benefits of the framework outlined in this paper motivate continued efforts to improve estimates of observation errors.

Acknowledgements

This paper has benefitted from comments from two anonymous referees, one of whom inspired the analysis in the appendix.

Appendix

We have seen the impacts that observation errors can have on proper scoring rules, but we might wonder how large the errors need to be for their impacts to have practical significance. Answers depend on the details of the forecasts, observations, scoring rule and purpose of the evaluation exercise, so the following analysis, which focuses on the major impact of forecasts being ranked incorrectly, is intended to be indicative rather than comprehensive.

Consider the case of binary predictands from section 4.1, where observation errors are defined by the misclassification probabilities, r_0 and r_1, and the forecasts, f, are probabilities for the event $\{x = 1\}$. Let y and f be conditionally independent given x, as in section 2.3, and suppose that the forecasts are calibrated so that $\Pr(x = 1 \mid f) = f$. It follows that

$$\begin{aligned}\Pr(y = 1 \mid f) &= \Pr(y = 1 \mid x = 0)\Pr(x = 0 \mid f) \\ &\quad + \Pr(y = 1 \mid x = 1)\Pr(x = 1 \mid f) \\ &= r_0 + (1 - r_0 - r_1)f.\end{aligned}$$

With no observation error, the expected score is

$$\begin{aligned}\mathrm{E}_{f,x}\{s(f,x)\} &= \mathrm{E}_f[\mathrm{E}_{x|f}\{s(f,x) \mid f\}] \\ &= \mathrm{E}_f\{fs(f,1) + (1-f)s(f,0)\},\end{aligned}$$

which we denote by S_f. With observation error, the expected score is

$$\mathrm{E}_{f,y}\{s(f,y)\} = r_0 S_{f1} + r_1 S_{f0} + (1 - r_0 - r_1)S_f,$$

where $S_{f1} = \mathrm{E}_f\{s(f,1)\}$ and $S_{f0} = \mathrm{E}_f\{s(f,0)\}$.

Let g denote some other calibrated forecasts with corresponding quantities S_g, S_{g1} and S_{g0}, and suppose that $S_g > S_f$ so that f scores better than g when there is no observation error. This order is reversed in the presence of observation error if

$$\mathrm{E}_{g,y}\{s(g,y)\} \leq \mathrm{E}_{f,y}\{s(f,y)\}.$$

When $r_0 = r_1$, for example, this inequality holds if and only if

$$D + r_0(D_0 + D_1 - 2D) \leq 0,$$

where $D = S_g - S_f$, $D_1 = S_{g1} - S_{f1}$ and $D_0 = S_{g0} - S_{f0}$. For the quadratic scoring rule, $s(f, x) = (f - x)^2$, we have $D_0 + D_1 = -2D$ and the inequality becomes $r_0 \geq 1/4$. Thus, the quadratic scoring rule will rank calibrated forecasts incorrectly (on average) if the misclassification probabilities are equal and exceed 0.25. Remarkably, this threshold (which is approximately the size of the misclassification probabilities in the aircraft icing example in section 5.2) does not depend on the true score difference, D. Such reverses are able to occur because the error in the score, given by the second term of the error-corrected scoring rule (13), is bigger for better forecasts.

For numerical predictands, simplify section 4.2 by considering additive observation errors with $\mathrm{E}(y \mid x) = x$ and $\mathrm{var}(y \mid x) = c^2$, and let μ and σ denote the mean and standard deviation of the forecasts, f. As above, let y and f be conditionally independent given x, and let the forecasts be calibrated so that $\mathrm{E}(x \mid f) = \mu$ and $\mathrm{var}(x \mid f) = \sigma^2$. Suppose also that, whereas μ may vary from forecast to forecast, σ is constant. Consider the Dawid-Sebastiani scoring rule,

$$s(f, x) = \log \sigma + \frac{(x - \mu)^2}{2\sigma^2}.$$

With no observation error, the expected score is

$$\begin{aligned} \mathrm{E}_{f,x}\{s(f, x)\} &= \mathrm{E}_f[\mathrm{E}_{x|f}\{s(f, x) \mid f\}] \\ &= \log \sigma + \frac{1}{2} \end{aligned}$$

since $\mathrm{E}_{x|f}\{(x - \mu)^2 \mid f\} = \sigma^2$. With observation error, the expected score is

$$\mathrm{E}_{f,y}\{s(f, y)\} = \log \sigma + \frac{1}{2} + \frac{c^2}{2\sigma^2}.$$

Suppose that there are two such forecasts, f and g, with means μ_f and μ_g, and standard deviations σ_f and σ_g. Let both forecasts be calibrated but let $\sigma_g > \sigma_f$ so that f scores better than g when there is no observation error. This order is reversed, and the forecasts are ranked incorrectly, in the presence of observation error if the score difference,

$$\begin{aligned} &\mathrm{E}_{g,y}\{s(g, y)\} - \mathrm{E}_{f,y}\{s(f, y)\} \\ &\qquad = \log\left(\frac{\sigma_g}{\sigma_f}\right) - \frac{c^2(\sigma_g^2 - \sigma_f^2)}{2\sigma_f^2\sigma_g^2}, \end{aligned}$$

is negative. This occurs if

$$\lambda \geq -\theta^{-1} \log(1 - \theta), \tag{23}$$

where $\lambda = c^2/\sigma_f^2$ measures the size of the observation error variance relative to the smaller forecast variance, and $\theta = (\sigma_g^2 - \sigma_f^2)/\sigma_g^2$ measures the difference in quality of the two forecasts. As above, such reverses are able to occur because the error in the score due to observation error is bigger for better forecasts.

A graph of the threshold (23) is plotted in Figure 7. The graph shows that the forecasts are ranked correctly if the observation error variance is less than the smaller forecast variance ($\lambda < 1$). If the observation error variance exceeds the smaller forecast variance, however, the forecasts might be ranked incorrectly. When the forecasts are similar (as is often the case when we are comparing two sets of skilful

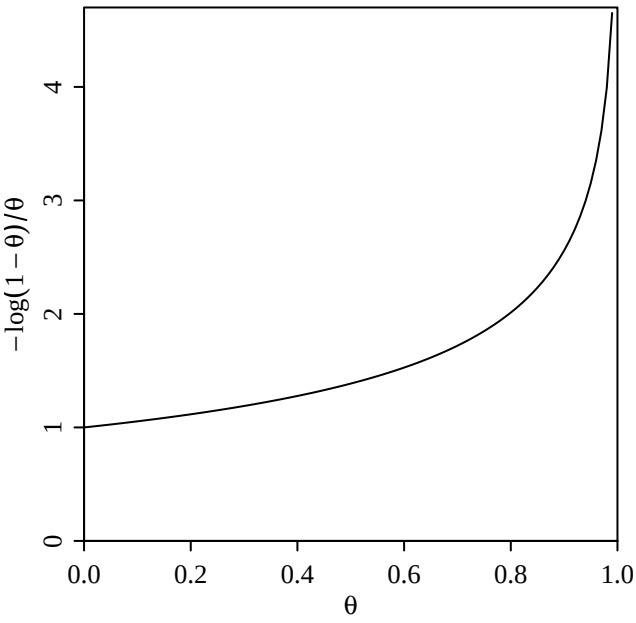


Figure 7. A threshold on the relative observation error variance, λ, above which forecasts are ranked incorrectly, plotted against the difference in forecast quality, θ.

forecasts), θ is small and the observation error variance needs to exceed the forecast variance by only a small amount in order to give the wrong ranking. For example, if $\theta \leq 0.1$ (so that σ_f^2 is at most 10% smaller than σ_g^2), the ranking is wrong if the observation error variance exceeds the forecast variance by about 5%.

References

Anderson CJ, Wikle CK, Zhou Q, Royle JA. 2007. Population influences on tornado reports in the United States. *Weather and Forecasting* **22**: 571–579, doi: 10.1175/WAF997.1.

Anderson JL. 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Clim.* **9**: 1518–1530, doi: 10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

Bowler NE. 2006. Explicitly accounting for observation error in categorical verification of forecasts. *Mon. Weather Rev.* **134**: 1600–1606, doi: 10.1175/MWR3138.1.

Bowler NE. 2008. Accounting for the effect of observation errors on verification of MOGREPS. *Meteorol. Appl.* **15**: 199–205, doi: 10.1002/met.64.

Bowler NA, Cullen MJP, Piccolo C. 2015. Verification against perturbed analyses and observations. *Nonlinear Process. Geophys.* **22**: 403–411, doi: 10.5194/npg-22-403-2015.

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**: 1–3, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Briggs W, Pocernich M, Ruppert D. 2005. Incorporating misclassification error in skill assessment. *Mon. Weather Rev.* **133**: 3382–3392, doi: 10.1175/MWR3032.1.

Bröcker J, Smith LA. 2007. Scoring probabilistic forecasts: the importance of being proper. *Weather and Forecasting* **22**: 382–388, doi: 10.1175/WAF966.1.

Broecker J. 2012. Probability forecasts. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (2nd edn), Jolliffe IT, Stephenson DB. (eds.): 119–139, John Wiley and Sons: Chichester.

Brown BG, Bernstein BC, McDonough F, Bernstein TAO. 1999. Probability forecasts of in-flight icing conditions. *8th Conference on Aviation, Range, and Aerospace Meteorology, Dallas, TX, 10–15 Jan 1999.* American Meteorological Association.

Candille G, Côté C, Houtekamer PL, Pellerin G. 2007. Verification of an ensemble prediction system against observations. *Mon. Weather Rev.* **135**: 2688–2699, doi: 10.1175/MWR3414.1.

Candille G, Talagrand O. 2008. Impact of observational error on the validation of ensemble prediction systems. *Q. J. R. Meteorol. Soc.* **134**: 959–971, doi: 10.1002/qj.268.

Ciach GJ, Krajewski WF. 1999. On the estimation of radar rainfall error variance. *Adv. Water Resour.* **22**: 585–595, doi: 10.1016/S0309-1708(98)00043-8.

Dawid AP, Sebastiani P. 1999. Coherent dispersion criteria for optimal experimental design. *Ann. Stat.* **27**: 65–81, doi: 10.1214/aos/1018031101.

Elsner JB, Michaels LE, Scheitlin KN, Elsner IJ. 2013. The decreasing population bias in tornado reports across the Central Plains. *Weather Clim. Soc.* **5**: 221–232, doi: 10.1175/WCAS-D-12-00040.1.

Ferro CAT. 2014. Fair scores for ensemble forecasts. *Q. J. R. Meteorol. Soc.* **140**: 1917–1923, doi: 10.1002/qj.2270.

Friederichs P, Thorarinsdottir T. 2012. Forecast verification scores for extreme value distributions with an application to peak wind prediction. *Environmetrics* **23**: 579–594. doi: 10.1002/env.2176.

Gilleland E. 2015. verification: Weather Forecast Verification Utilities. R package version 1.42. URL `http://CRAN.R-project.org/package=verification`.

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. B* **69**: 243–268, doi: 10.1111/j.1467-9868.2007.00587.x.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.* **102**: 359–378, doi: 10.1198/016214506000001437.

Gneiting T, Raftery AE, Westveld AH, Goldman T. 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133**: 1098–1118, doi: 10.1175/MWR2904.1.

Good IJ. 1952. Rational decisions. *J. R. Stat. Soc. B* **14**: 107–114.

Gorgas T, Dorninger M. 2012. Quantifying verification uncertainty by reference data variation. *Meteorol. Z.* **21**: 259–277, doi: 10.1127/0941-2948/2012/0325.

Hamill TM. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Weather Rev.* **129**: 550–560, doi: 10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Mittermaier MP, Stephenson DB. 2015. Inherent bounds on forecast accuracy due to observation uncertainty caused by temporal sampling. *Mon. Weather Rev.* **143**: 4236–4243. doi: 10.1175/MWR-D-15-0173.1.

Pappenberger F, Ghelli A, Buizza R, Bódis K. 2009. The skill of probabilistic precipitation forecasts under observational uncertainties within the generalized likelihood uncertainty estimation framework for hydrological applications. *J. Hydrometeor.* **10**: 807–819, doi: 10.1175/2008JHM956.1.

Pinson P, Hagedorn R. 2012. Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. *Meteorol. Appl.* **19**: 484–500, doi: 10.1002/met.283.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Ray PS, Bieringer P, Niu X, Whissel B. 2003. An improved estimate of tornado occurrence in the Central Plains of the United States. *Mon. Weather Rev.* **131**: 1026–1031, doi: 10.1175/1520-0493(2003)131<1026:AIEOTO>2.0.CO;2.

Röpnack A, Hense A, Gebhardt C, Majewski D. 2013. Bayesian model verification of NWP ensemble forecasts. *Mon. Weather Rev.* **141**: 375–387, doi: 10.1175/MWR-D-11-00350.1.

Saetra Ø, Hersbach H, Bidlot J-R, Richardson DS. 2004. Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Weather Rev.* **132**: 1487–1501, doi: 10.1175/1520-0493(2004)132<1487:EOOEOT>2.0.CO;2.

Santos C, Ghelli A. 2012. Observational probability method to assess ensemble precipitation forecasts. *Q. J. R. Meteorol. Soc.* **138**: 209–221, doi: 10.1002/qj.895.

Thorarinsdottir TL, Gneiting T, Gissibl N. 2013. Using proper divergence functions to evaluate climate models. *J. Uncertain. Quantif.* **1**: 522–534, doi: 10.1137/130907550.

Vescio MD, Thompson RL. 2001. Subjective tornado probability forecasts in severe weather watches. *Weather and Forecasting* **16**: 192–195, doi: 10.1175/1520-0434(2001)016<0192:FSFSTP>2.0.CO;2.

Weijs SV, van de Giesen N. 2011. Accounting for observational uncertainty in forecast verification: an information-theoretical view on forecasts, observations and truth. *Mon. Weather Rev.* **139**: 2156–2162, doi: 10.1175/2011MWR3573.1.

Winkler RL. 1996. Scoring rules and the evaluation of probabilities. *Test* **5**: 1–60, doi: 10.1007/BF02562681.