METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **24**: 276–278 (2017)
Published online 20 February 2017 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/met.1626

RMetS
Royal Meteorological Society

Probability forecasts with observation error: what should be forecast?

I. T. Jolliffe*
Exeter Climate Systems, University of Exeter, UK

ABSTRACT: When probability forecasts are made of a binary event, a commonly used measure for assessing the forecasts is the Brier score. One of its properties is that it is proper, meaning that its expected value cannot be improved by the forecaster issuing a probability other than his/her true belief. This property assumes that the occurrence or otherwise of the forecast event is recorded without error. This note investigates what forecast should be made in order to minimize the expected value of the Brier score when errors are present in the observations. Should it still be the forecaster's true belief or should it be something else, implying that the forecaster should hedge his/her forecast? The answer is that it depends on whether the forecaster can model the error mechanism or whether the error mechanism is unknown. It is shown that in the former case the forecaster's true belief of the probability of the event should still be forecast. However, in the case of an unknown error mechanism, the forecaster should attempt to forecast the probability that the erroneous observation indicates that the event has occurred, rather than the true probability of the event.

KEY WORDS binary probability forecasts; Brier score; hedging; observation error; proper scores

Received 15 June 2016; Revised 25 August 2016; Accepted 6 September 2016

1. Introduction

Suppose that probability forecasts are to be made for a set of binary events, such as whether the temperature at a station will be below a threshold or whether 'severe weather' will occur somewhere in a specified geographical region. A commonly used measure of the quality of a set of such forecasts is the Brier score (Brier, 1950). As well as being intuitive, the Brier score has the desirable property that it is proper (Winkler and Murphy, 1968). This means that the forecaster cannot improve the expected value of the score by hedging, i.e. by giving a forecast probability different from his/her 'true belief'. In demonstrating that the Brier score is proper it is assumed that the occurrence or otherwise of the event of interest is observed without error. This is not always the case.

For example, suppose that the forecast is for the occurrence of a severe weather event somewhere in a large geographical area. If the event of interest is localized and the observing network is sparse, it is then possible that a true occurrence of the event is missed. Another example is when the event of interest is that some measured variable is above/below a threshold, e.g. wind speed above 50 knots or temperature below 0 °C. The measuring instrument used will have some measurement error, which can lead to a false declaration that the event has/has not occurred. These two examples are examined in detail later but, more generally, observation error can take many forms including reporting errors and practices such as rounding, as well as measurement error. The latter can vary greatly in magnitude, ranging from small values for directly observed temperatures to large discrepancies for rainfall radar.

* Correspondence: I. T. Jolliffe, 30 Woodvale Road, Gurnard, Cowes, Isle of Wight PO31 8EG, UK. E-mail: i.t.jolliffe@exeter.ac.uk; ian@sandloch.fsnet.co.uk

This note investigates what forecast should be made (the forecaster's true belief or otherwise) when there is possible error associated with observation of the forecast event. Section 2 discusses what the optimal forecast is when the error mechanism is not known or is known. Two examples of the latter case are described in Section 3 and Section 4 gives some final remarks.

2. The Brier score, optimal forecasts and propriety

Suppose that (f_i, o_i), $i = 1,2, \ldots, n$, is a set of n probability forecasts of an event and the corresponding observations. Here, f_i can take any value between 0 and 1, while o_i can only take the values 1 or 0, depending on whether the forecast event does or does not occur. The Brier score is defined as:

$$B = \frac{1}{n}\sum_{i=1}^{n}\left(f_i - o_i\right)^2 \tag{1}$$

If the observations are without error, it is relatively straightforward to show that the Brier score is proper, meaning that its expected value cannot be improved by forecasting something other than the forecaster's true belief. Consider one of the terms in the summation defining B in Equation (1). Dropping its subscript for notational convenience, this is $(f-o)^2$. Suppose that the forecaster believes that there is a probability p_{h} (honest belief) of the event occurring on this occasion. The expected value of a binary random variable is the sum over its two values of the product of each value and the probability of that value occurring. The random variable o takes the two values 0 and 1, with probabilities $1-p_{\mathrm{h}}$ and p_{h} respectively, so the expected value of $(f-o)^2$ is:

$$E\left\{(f-o)^2\right\} = f^2\left(1-p_{\mathrm{h}}\right) + (f-1)^2 p_{\mathrm{h}} \tag{2}$$

Differentiating this expression with respect to f and equating its derivative to zero shows that it is uniquely minimized when $f = p_{\mathrm{h}}$. Minimizing each term in B in Equation (1) will minimize their sum, so the forecaster's optimum forecast is his/her

true/honest belief on each forecast occasion. However, what happens when the observations are made with error?

A number of authors have considered the effect of observational errors on verification measures, e.g. Candille and Talagrand (2008), Pappenberger *et al.* (2009), Santos and Ghelli (2012) and Mittermaier and Stephenson (2015). Of these, Candille and Talagrand (2008) and Santos and Ghelli (2012) specifically consider the Brier score when there is observational uncertainty, but from a different perspective to that given here. They look at the effect of the errors on components of the score corresponding to reliability and resolution and on the Brier skill score, rather than what should be the optimal forecast.

Now, suppose that it is known that the observation may be in error, but there is no knowledge of the mechanism by which these errors occur. For a given forecast, let o_e denote the error-prone observation and let $q = \Pr(o_e = 1)$. If these error-prone observations are all that are available, it is not possible to evaluate the expected value $E\{(f-o)^2\}$ of the terms in the Brier score. Instead it seems natural to minimize:

$$E\left\{\left(f-o_e\right)^2\right\} = f^2(1-q) + (f-1)^2 q \tag{3}$$

which is minimized when $f = q$. Thus, the forecaster should forecast his/her belief that the error-prone observation will indicate that the event has occurred, which will typically be different from his/her belief that the event actually will occur.

If the observations are error-prone it would be hoped that there is at least some knowledge of the error mechanism. Examples of this are given in the next section but for the moment simply suppose that it is possible to give beliefs for q and for the conditional probabilities $c_1 = \Pr(o = 1|o_e = 1)$, $c_0 = \Pr(o = 1|o_e = 0)$. Then:

$$\Pr(o=1) = p = c_1 q + c_0(1-q) \tag{4}$$

$$\Pr(o=0) = 1-p = \left(1-c_1\right)q + \left(1-c_0\right)(1-q) \tag{5}$$

Replacing p_h and q in Equations (2) and (3), respectively, by the expression in Equation (4) gives the expected value of the term in the Brier score when the forecast f is made to be:

$$E\left\{(f-o)^2\right\} = (1-f)^2\left\{c_1 q + c_0(1-q)\right\} + f^2\left\{\left(1-c_1\right)q + \left(1-c_0\right)(1-q)\right\}$$

Differentiating with respect to f and equating to zero, there is considerable simplification and it turns out that the expected value of the Brier score is minimized for:

$$f = c_1 q + \left(1-c_0\right)(1-q) = p$$

Hence, given this knowledge/belief about the relationship between the true and error-prone forecasts, the forecaster should again forecast his/her true p_h belief regarding p, given his/her knowledge/belief regarding c_0, c_1 and q.

3. Two scenarios for error-prone observations

3.1. Spatially missed events

Suppose that the forecast is for the occurrence of a severe weather event somewhere in a large geographical area. If the event of interest is localized and the observing network is sparse, it is then possible that a true occurrence of the event is missed, but it is likely that when an occurrence is said to have occurred then this is correct. This means that $c_1 \sim 1$ but $c_0 > 0$ and $p \sim q + c_0(1-q) > q$. As an example, suppose that $c_0 = 0.2$; so, on the 20% of occasions that the observation said 'no event' there was, in fact, a missed event. If $q = 0.75$, then $p = 0.80$, and if $q = 0.20$, then $p = 0.36$, for example. One apparently strange aspect of this scenario is that there is a lower bound on p, namely c_0. With $c_0 = 0.2$, for example, it is impossible to have $p < 0.2$; otherwise $q < 0$. It ought to be possible to take p to be any value between 0 and 1 and indeed it is. The restriction is really one on c_0. If p, q are very small, then the proportion of 'observed non-events' that correspond to missed events is itself very small – hence the bound.

3.2. Temperature below/above threshold measured with error

Suppose that the event of interest is whether or not the surface temperature is below or above a threshold. As an example, 0 °C is an obvious threshold. The observed temperature may be subject to measurement error due to the instrument, the exposure, the measuring process or any combination of these. Other factors may also lead to errors such as a reporting process in which there is a mismatch between the time of observation and the time for which the forecast is made. Let T be the observed temperature and τ the true temperature and assume that $T|\tau \sim N\left(\tau, \sigma_T^2\right)$ and $\tau \sim N\left(\mu, \sigma_\tau^2\right)$. From these assumptions, if μ, ${\sigma_T}^2$ and ${\sigma_\tau}^2$ are specified, then the joint distribution of T and τ can be found, as can the conditional distribution of τ, given T. The joint distribution of T and τ is bivariate normal with $E(T) = E(\tau) = \mu$, $\mathrm{var}(\tau) = {\sigma_\tau}^2$, $\mathrm{var}(T) = {\sigma_\tau}^2 + {\sigma_T}^2$ and $\mathrm{cov}(T, \tau) = {\sigma_\tau}^2$.

Given this distribution and the conditional distribution of τ given T, the quantities p, q, c_0 and c_1 can be calculated. In the present context c_1, c_0 are probabilities that the true temperature is below the threshold, given that the observed temperature is below/above the threshold respectively. For illustration, suppose that $\sigma_\tau = 3.0$ °C and $\sigma_T = 0.5$ °C. This implies that the true temperatures are mostly within a range of about 12 °C and the observations are rarely in error by more than 1 °C. The values of σ_τ and σ_t are chosen subjectively, but the latter is not out of line with the conclusions of Mittermaier and Stephenson (2015). The correlation between T and τ is 0.986.

The probabilities p, q, c_0 and c_1 can be calculated for various values of μ. For a threshold of 0 °C, Table 1 gives values of p and q for selected values of μ.

Recall that p is the true value for the probability of a temperature below 0 °C, whereas q is the probability of this event implied by the erroneous observations. In the previous section it was suggested that for propriety a forecaster should forecast p rather than q. In this example, p and q are very similar because of the high correlation between T and τ, although their ratio diverges from 1 as μ increases. Differences between p and q are larger if σ_τ is much closer in value to σ_T, but this does not seem realistic.

Values for c_0 and c_1 for the same situation as above, for selected values of μ, are given in Table 2.

Recall that c_1 is the probability that the temperature really is below the threshold when the observed value says it is, and c_0 is the corresponding probability for the true temperature when the observation indicates that the temperature is not below the threshold. The probability c_1 is reasonably close to 1 unless the mean temperature is substantially above the 0 °C threshold. Similarly, c_0 is close to zero unless the mean temperature is well below the threshold.

Table 1. True and observed probabilities for a selection of values of μ.

Mean, μ	−4	−2	0	2	4	6	8
True probability p	0.909	0.748	0.500	0.252	0.091	0.023	0.004
Observed probability q	0.906	0.745	0.500	0.255	0.094	0.024	0.004
p/q	1.003	1.004	1.000	0.989	0.966	0.938	0.899

Table 2. Conditional probabilities c_1, c_0 for selected values of μ.

Mean, μ	−4	−2	0	2	4	6	8
c_1	0.990	0.974	0.947	0.912	0.867	0.817	0.761
c_0	0.133	0.088	0.053	0.026	0.011	0.003	0.001

4. Concluding remarks

The expected value of the Brier score for probability forecasts is minimized when the forecaster forecasts his/her true belief, meaning that the score is proper, provided that the observations are recorded without error. When error is present, with no knowledge of the error mechanism, the forecaster should forecast his/her belief regarding the probability of the erroneous observation indicating that the event of interest has occurred, rather than his/her belief regarding the true probability of the event. However, when the error mechanism can be modelled, the forecaster should revert to forecasting his/her belief regarding the true probability of the event. In the latter case, two simple examples have been examined. In these examples, comparisons can be made between the true probability of the event of interest and its probability as suggested by the erroneous observations. The reason for reversion to true belief is perhaps that with knowledge of the error mechanism it is possible to deduce true probabilities from those given by the erroneous observations, whereas this is not possible when the error mechanism is unknown.

This note assumes that propriety is the important property for binary probability forecasts. If this is not considered to be the case, strategies other than those recommended above may be deemed more appropriate.

Acknowledgement

The author is grateful to two anonymous reviewers whose helpful comments led to improvements in this note.

References

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**: 1–3.

Candille G, Talagrand O. 2008. Impact of observational error on the validation of ensemble prediction systems. *Q. J. R. Meteorolog. Soc.* **134**: 957–971.

Mittermaier MP, Stephenson DB. 2015. Inherent bounds on forecast accuracy due to observational uncertainty caused by temporal sampling. *Mon. Weather Rev.* **143**: 4236–4243.

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. Hydrometeorol.* **10**: 807–819.

Santos C, Ghelli A. 2012. Observational probability method to assess ensemble precipitation forecasts. *Q. J. R. Meteorolog. Soc.* **138**: 209–221.

Winkler RL, Murphy AH. 1968. 'Good' probability assessors. *J. Appl. Meteorol.* **7**: 751–758.

