A bias-corrected decomposition of the Brier score

C. A. T. Ferro[ab*] and T. E. Fricker[b]

[a] *National Centre for Atmospheric Science*

[b] *College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK*

[*] Correspondence to: C. A. T. Ferro, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QF, UK. E-mail: c.a.t.ferro@exeter.ac.uk

The Brier score is a widely used measure of performance for probabilistic forecasts of event occurrences, and it is often decomposed additively into three terms that quantify the reliability and resolution of the forecasts, and the uncertainty of the forecasted events. The standard decomposition yields biased estimates of the large-sample values of these three quantities: reliability is overestimated and uncertainty is underestimated, while resolution may be either overestimated or underestimated. An unbiased decomposition is shown to be unattainable but a new decomposition is proposed that has smaller biases and therefore provides a more accurate measure of forecast performance. The implications for the Brier skill score and the attributes diagram are discussed, and results are illustrated with seasonal forecasts of sea surface temperatures.

Key Words: attributes diagram; calibration; forecast verification; refinement; reliability; resolution

Received . . .

Citation: . . .

1. The Brier score and its decomposition

Suppose that probabilities $p_1, \ldots, p_n$ are forecasts for the occurrence of n events, and let $x_1, \ldots, x_n$ indicate whether or not the n events occur, so that $x_i = 1$ if the ith event occurs and $x_i = 0$ if the ith event fails to occur. The Brier score (Brier, 1950) for these forecasts is

$$B = \frac{1}{n}\sum_{i=1}^{n}(p_i - x_i)^2$$

and takes values in the interval $[0, 1]$ with smaller values indicating better forecasts.

Suppose now that each forecast can take one of only K distinct values, $\pi_1, \ldots, \pi_K$. Let $I_k = \{i : p_i = \pi_k\}$ be the set of indices for those occasions on which π_k is forecast and let n_k be the number of such occasions. For those k for which $n_k > 0$, define the conditional relative frequency,

$$\bar{x}_k = \frac{1}{n_k} \sum_{i \in I_k} x_i,$$

to be the proportion of events that occur out of the n_k occasions on which π_k is forecast. Also define

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

to be the overall proportion of occasions on which the event occurs. Then the Brier score can be decomposed (Murphy, 1973) as

$$B = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC}, \qquad (1)$$

where

$$\mathrm{REL} = \sum_{k \in K_0} \frac{n_k}{n} (\pi_k - \bar{x}_k)^2, \qquad (2)$$

$$\mathrm{RES} = \sum_{k \in K_0} \frac{n_k}{n} (\bar{x}_k - \bar{x})^2, \qquad (3)$$

$$\mathrm{UNC} = \bar{x}(1 - \bar{x}) \qquad (4)$$

and $K_0 = \{k : n_k > 0\}$ so that the sums are over those k for which n_k exceeds zero. The first term (REL) in the decomposition is a weighted average of the squared differences between the conditional relative frequencies and the corresponding forecasts, and measures the reliability of the forecasts. The best score for the reliability is zero, which is obtained if the conditional relative frequencies are equal to their corresponding forecasts. The second term (RES) is a weighted variance of the conditional relative frequencies and measures the resolution of the forecasts. The worst score for the resolution is zero, which is obtained if the conditional relative frequencies are the same for all forecasts. The third term (UNC) is a measure of uncertainty or climatological variation in the event occurrence. Very rare or very common events have low uncertainty.

Bröcker (2011) showed that the three terms in the Brier score decomposition (1) are biased. This means that the expected value of each term is typically different from its true value, defined to be the value that would be obtained were the sample size, n, increased to infinity. The reliability is systematically overestimated, the uncertainty is systematically underestimated, while the resolution may be either overestimated or underestimated. Therefore, evaluating this standard decomposition for finite samples can give a misleading impression of forecast quality. We show that an unbiased decomposition of the Brier score is unattainable but propose a new decomposition that has smaller biases than the standard decomposition and therefore provides a more accurate measure of forecast performance. We discuss the implications of the bias for the Brier skill score and the attributes diagram, and illustrate the new decomposition with seasonal forecasts of sea surface temperatures.

2. The bias and a bias-corrected decomposition

We show that the standard decomposition of the Brier score is biased and derive our results under the assumption that the forecast-verification pairs $\{(p_i, x_i) : i = 1, \ldots, n\}$ are independent and identically distributed random variables. Extensions to dependent random variables are discussed in section 5. Define the long-run relative frequency with which the event occurs to be the expected value

$$\mu = E(x_i)$$

for all i, and define the long-run relative frequency with which the event occurs amongst those occasions on which the forecast equals π_k to be

$$\mu_k = E(x_i \mid p_i = \pi_k)$$

for all i and each k. Also define the expected frequency with which π_k is forecast in a sequence of n forecasts to be

$$E(n_k) = n\phi_k$$

for each k, where $\phi_k > 0$. The weak law of large numbers tells us that $\bar{x} \to \mu$, $\bar{x}_k \to \mu_k$ and $n_k/n \to \phi_k$ for each k as $n \to \infty$. Substituting these limits into the decomposition (1) of the Brier score yields the following limits for the reliability, resolution and uncertainty:

$$\begin{aligned} \mathrm{REL}_\infty &= \sum_{k=1}^{K} \phi_k(\pi_k - \mu_k)^2, &(5)\\ \mathrm{RES}_\infty &= \sum_{k=1}^{K} \phi_k(\mu_k - \mu)^2, \\ \mathrm{UNC}_\infty &= \mu(1-\mu). &(6)\end{aligned}$$

These are the values that would be obtained were the sample size infinite. For finite n, however, a special case of a result obtained by Bröcker (2011) shows that the expected values of the reliability, resolution and uncertainty terms in the standard decomposition (1) are as follows:

$$\begin{aligned} E(\mathrm{REL}) &= \mathrm{REL}_\infty + \frac{1}{n}\sum_{k=1}^{K} \nu_{k,n}\mu_k(1-\mu_k), \\ E(\mathrm{RES}) &= \mathrm{RES}_\infty + \frac{1}{n}\sum_{k=1}^{K} \nu_{k,n}\mu_k(1-\mu_k) - \frac{\mu(1-\mu)}{n}, \\ E(\mathrm{UNC}) &= \mathrm{UNC}_\infty - \frac{\mu(1-\mu)}{n}, &(7)\end{aligned}$$

where $\nu_{k,n}$ is the probability that n_k exceeds zero. A special case of these expressions (in which members of an ensemble predict the event independently with probability μ, the forecast is the proportion of ensemble members that predict the event, and $\mu_k = \mu$ for all k) was obtained by Ferro *et al.* (2008) in their investigation of the effect of ensemble size on the Brier score, but they did not comment on the dependence of these expected values on the sample size, n. The differences between the expected and limiting values above are the biases. The bias in the reliability,

$$\begin{aligned} \mathrm{bias}(\mathrm{REL}) &= E(\mathrm{REL}) - \mathrm{REL}_\infty \\ &= \frac{1}{n}\sum_{k=1}^{K} \nu_{k,n}\mu_k(1-\mu_k), &(8)\end{aligned}$$

is non-negative and decreases monotonically to zero as n increases. In other words, REL tends to overestimate REL_∞ and the reliability of the forecasts will tend to appear poorer than it would do were a larger sample available. The bias in the uncertainty,

$$\mathrm{bias}(\mathrm{UNC}) = -\frac{\mu(1-\mu)}{n}, \qquad (9)$$

is non-positive and increases monotonically to zero as n increases. So, the uncertainty will tend to appear smaller than it would do were a larger sample available. The bias in the resolution can be positive or negative, but also converges to zero as n increases. In practice, however, the bias in the resolution is often positive because $\mu(1-\mu)$ is often small compared to $\sum_{k=1}^{K} \nu_{k,n}\mu_k(1-\mu_k)$, in which case the resolution of the forecasts will tend to appear better than it would do were a larger sample available.

We prove in the appendix that unbiased estimators for the reliability and resolution are unattainable. Nonetheless, we propose a new decomposition of the Brier score in which the estimate of uncertainty is unbiased and the estimates of reliability and resolution have smaller biases than in the standard decomposition. This new decomposition is

$$B = \mathrm{REL}' - \mathrm{RES}' + \mathrm{UNC}', \qquad (10)$$

where

$$\mathrm{REL}' = \mathrm{REL} - \frac{1}{n}\sum_{k\in K_1}\frac{n_k}{n_k-1}\bar{x}_k(1-\bar{x}_k), \tag{11}$$

$$\mathrm{RES}' = \mathrm{RES} - \frac{1}{n}\sum_{k\in K_1}\frac{n_k}{n_k-1}\bar{x}_k(1-\bar{x}_k) + \frac{\bar{x}(1-\bar{x})}{n-1}, \tag{12}$$

$$\mathrm{UNC}' = \mathrm{UNC} + \frac{\bar{x}(1-\bar{x})}{n-1} \tag{13}$$

and $K_1 = \{k : n_k > 1\}$ so that the sums are over those k for which n_k exceeds 1. Usually all n_k exceed 1 because small n_k are often eradicated by relabelling distinct forecasts with a common forecast value (e.g. Bröcker and Smith, 2007), although this will typically change the limiting values, REL_∞ and RES_∞, being estimated. Whether or not forecasts are pooled in this way, the new decomposition yields more accurate estimates than the standard decomposition. We prove in the appendix that UNC' is unbiased and that the biases of REL' and RES' decay to zero at a faster rate than the biases of REL and RES as the sample size, n, increases.

The new decomposition has one complication: REL' and RES' can be negative. In such cases, we recommend replacing the sum in the definitions of REL' (11) and RES' (12) by the largest value for which both terms are non-negative. This is equivalent to replacing REL' with $\max\{\mathrm{REL}', \mathrm{REL}' - \mathrm{RES}', 0\}$ and replacing RES' with $\max\{\mathrm{RES}', \mathrm{RES}' - \mathrm{REL}', 0\}$. This ensures that the three terms in the decomposition still combine to equal B.

Independent work by Bröcker (2011) proposed a different decomposition:

$$\mathrm{REL}'' = \mathrm{REL} - \frac{1}{n}\sum_{k\in K_0}\bar{x}_k(1-\bar{x}_k), \tag{14}$$

$$\mathrm{RES}'' = \mathrm{RES} - \frac{1}{n}\sum_{k\in K_0}\bar{x}_k(1-\bar{x}_k) + \frac{\bar{x}(1-\bar{x})}{n}, \tag{15}$$

$$\mathrm{UNC}'' = \mathrm{UNC} + \frac{\bar{x}(1-\bar{x})}{n}. \tag{16}$$

We prove in the appendix that the biases of these estimates all decay to zero more slowly than the biases of our decomposition. We also show in the appendix that the biases of the uncertainty and reliability terms in these three decompositions satisfy the following orderings for all n:

$$\mathrm{bias}(\mathrm{UNC}) \le \mathrm{bias}(\mathrm{UNC}'') \le \mathrm{bias}(\mathrm{UNC}') = 0 \tag{17}$$

and

$$0 \le \mathrm{bias}(\mathrm{REL}') \le \mathrm{bias}(\mathrm{REL}'') \le \mathrm{bias}(\mathrm{REL}). \tag{18}$$

The ordering on the biases of the resolution terms can depend on n.

3. The Brier skill score and attributes diagram

The Brier skill score (e.g. Glahn and Jorgensen, 1970) is defined as $\mathrm{BSS} = 1 - B/B_{\mathrm{ref}}$, where B_{ref} is the Brier score for some set of reference forecasts. If the reference forecasts are always equal to the in-sample climatology, $\bar{x}$, then $B_{\mathrm{ref}} = \mathrm{UNC}$ and $\mathrm{BSS} = (\mathrm{RES} - \mathrm{REL})/\mathrm{UNC}$ (Murphy, 1996). The preceding calculations show that both the numerator and denominator of this BSS are systematically underestimated, but to say anything about the bias of the ratio would require further analysis. We do find, however, that the BSS based on the new decomposition is larger than the BSS based on the standard decomposition: since $\mathrm{UNC}' \ge \mathrm{UNC}$, we have $\mathrm{BSS}' = 1 - B/\mathrm{UNC}' \ge 1 - B/\mathrm{UNC} = \mathrm{BSS}$ with equality if and only if $B = 0$.

Our new decomposition of the Brier score also has implications for the attributes diagram of Hsu and Murphy (1986). The attributes diagram augments the reliability diagram, which comprises the points $\{(\pi_k, \bar{x}_k) : k = 1, \ldots, K\}$, with three lines: the horizontal line at height $\bar{x}$ representing climatology, the 45° line through the origin representing perfect reliability ($\mathrm{REL} = 0$), and the no-skill line $\bar{x}_k = (\pi_k + \bar{x})/2$. The positions of the points and the first two lines in the diagram are unaffected because the forecast values π_k are fixed, the quantities $\bar{x}_k$ and $\bar{x}$ are unbiased estimators of the corresponding long-run quantities, and if all points lie on the 45° line then $\mathrm{REL}' =$

0. The no-skill line, however, is affected. This line is derived by recalling that reference forecasts equal to the in-sample climatology, $\bar{x}$, yield a Brier score equal to UNC. Setting $B = \mathrm{UNC}$ implies $\mathrm{REL} = \mathrm{RES}$, and the no-skill line is obtained by equating the summands in the definitions of REL (2) and RES (3). However, UNC is a biased estimator for the expected Brier score achieved by the long-run climatological reference forecast, μ. An unbiased estimator is UNC' and setting $B = \mathrm{UNC}'$ implies $\mathrm{REL}' = \mathrm{RES}'$. A new, no-skill curve is obtained, therefore, by equating the summands in the definitions of REL' (11) and RES' (12). Rewriting the last term in the definition of RES' as the sum $\sum_{k=1}^{K} n_k\{\bar{x}_k(1-\bar{x}_k) + (\bar{x}_k - \bar{x})^2\}/\{n(n-1)\}$ shows that this new curve is defined by

$$\bar{x}_k = \frac{\pi_k^2 - \alpha}{2\pi_k - \beta},$$

where $\alpha = n\bar{x}^2/(n-1)$ and $\beta = (2n\bar{x}-1)/(n-1)$. This curve is a hyperbola with asymptotes $\pi_k = \beta/2$ and $\bar{x}_k = (2\pi_k + \beta)/4$. In the region between the two branches of the hyperbola, the REL' summand is less than the corresponding RES' summand and so this region represents forecasts that make a positive contribution to skill.

4. A numerical illustration

We illustrate the standard (1) and bias-corrected (10) decompositions of the Brier score using 4928 probabilistic, seasonal forecasts of equatorial Pacific monthly mean sea surface temperature (SST) anomalies constructed previously by Stephenson *et al.* (2005). The forecasts are for the event that the anomaly is positive and are verified against the ERA-40 reanalysis (Uppala *et al.*, 2005). Further information about the forecasts and verifications may be found in Stephenson *et al.* (2008). We categorize the forecast probabilities into ten, non-overlapping bins of width 0.1 and replace each forecast by its corresponding bin mean so that there are ten distinct forecast probabilities. Qualitatively similar results were obtained for other bin widths.

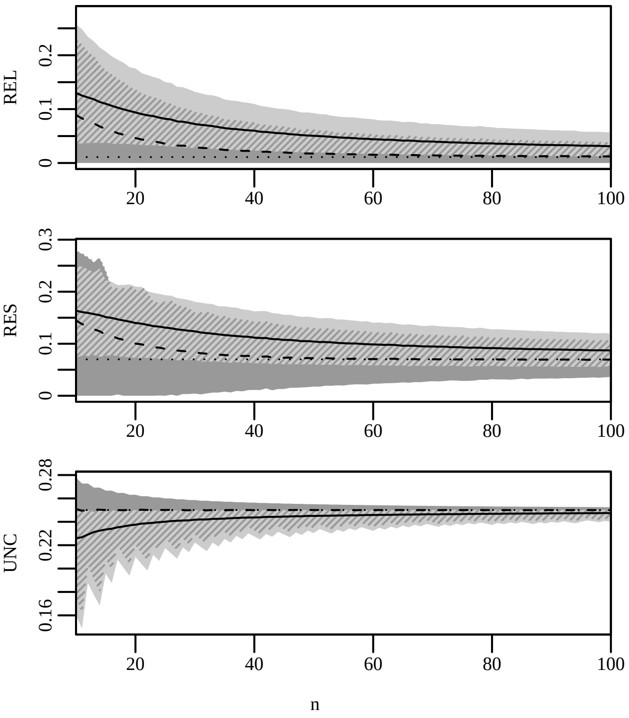

Figure 1. Expected values of reliability, resolution and uncertainty against sample size, n, for the SST forecasts: standard decomposition (solid lines), bias-corrected decomposition (dashed lines) and true, long-run values (dotted lines). Pointwise 5–95% intervals of the sampling distributions are superimposed: standard decomposition (light grey regions), bias-corrected decomposition (dark grey regions) and their overlap (hashed).

We use the following procedure to illustrate how the biases in the Brier score components depend on the sample size, n. First, we calculate the bias-corrected reliability, resolution and uncertainty components of the Brier score using all 4928 forecast-verification pairs, and take these values as approximations to the true, long-run values REL_∞, RES_∞ and UNC_∞. Then, for each $n < 4928$, we form 10 000 samples of n forecast-verification pairs by sub-sampling at random from the full data set, and compute the standard and bias-corrected decompositions for each sample. Thus, for each n, we obtain 10 000 values of REL and REL', RES and RES', and UNC and UNC'. The means of these values approximate the corresponding expected values and are plotted in Figure 1 for $10 \leq n \leq 100$. The 5% and 95% quantiles of the 10 000 values are also plotted to illustrate the sampling variation.

As expected, the standard Brier score decomposition yields large biases. The expected values of REL and RES exceed REL_∞ and RES_∞ while the expected value of

UNC lies below UNC_∞. The magnitudes of the biases are considerable when n is small. For example, the expected value of REL is at least five times greater than REL_∞ when n is less than 40. When the bias-corrected decomposition is used, the bias of UNC' is zero for all n while the biases of REL' and RES' are smaller and decay more rapidly than the biases of REL and RES. The biases of REL' and RES' are negligible when n is greater than about 60, an accuracy achieved by REL and RES only once n exceeds 300 (not shown).

The 5–95% intervals defined by the quantiles of the sampling distributions are wider for RES' than for RES, slightly wider for UNC' than for UNC, and slightly narrower for REL' than for REL. The sampling variation is greater for UNC' than for UNC because $\mathrm{UNC}'/\mathrm{UNC} = n/(n-1) > 1$. We do not know if the sampling variation for REL' is always less than for REL, or if the sampling variation for RES' is always greater than for RES. For individual data sets, standard errors and confidence intervals for the three components might be estimated using ideas similar to those employed by Ferro (2007).

The SST data used in Figure 1 exhibit significant temporal dependence up to lags of three months and, therefore, violate the independence assumption that was used to derive the biases of the decompositions in section 2. The resampling scheme employed to construct Figure 1, however, destroys the time order of the data, and so the results are indicative of how the decompositions perform when there is no temporal dependence. The performance of the decompositions in the presence of temporal dependence is discussed in section 5.

To illustrate our proposed adjustment to the no-skill curve in the attributes diagram, we consider a subset of 88 forecasts from a single gridpoint, at 150°W in the central equatorial Pacific. Using data from a single gridpoint helps to highlight the differences between the standard and bias-corrected no-skill curves because the two are visually indistinguishable when n is large. The diagram is shown in Figure 2. We see that the bias-corrected curve results in a larger positive skill region. The Brier score for these data is $B = 0.131$ and the standard decomposition yields $\mathrm{REL} = 0.018$, $\mathrm{RES} = 0.137$ and $\mathrm{UNC} = 0.250$ with $\mathrm{BSS} = 0.475$, while the bias-corrected decomposition yields $\mathrm{REL}' = 0.009$, $\mathrm{RES}' = 0.129$ and $\mathrm{UNC}' = 0.251$ with $\mathrm{BSS}' = 0.478$.

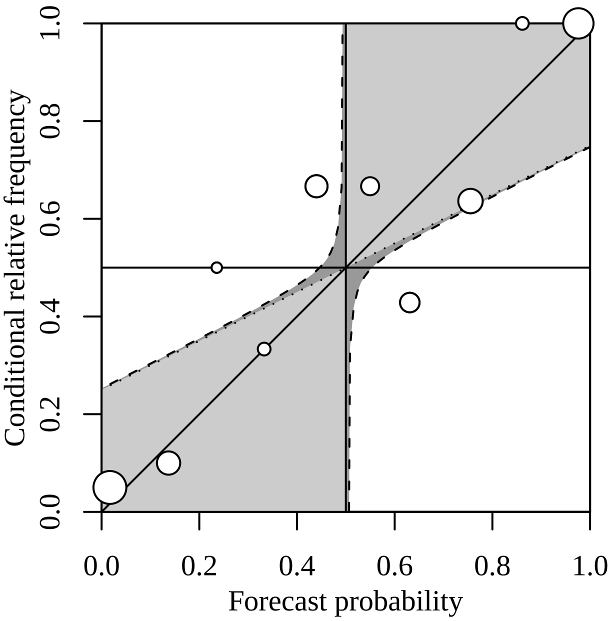

Figure 2. Attributes diagram for the SST forecasts. The circles are centred on the points $(\pi_k, \bar{x}_k)$ and their areas are proportional to the number, n_k, of contributing data. The light grey region is the positive-skill region given by the standard no-skill line (dotted line). The dark grey region is the area added to the positive-skill region by using the bias-corrected no-skill curve (dashed curve). The solid horizontal and vertical lines represent the observed climatology, $\bar{x}$.

5. Discussion

The reliability-resolution-uncertainty decomposition of the Brier score is obtained by conditioning on the forecasts (Murphy, 1973). An alternative decomposition is obtained by conditioning on the verifications (Murphy and Winkler, 1987) to yield three terms that Murphy (1996) refers to as the type 2 conditional bias, the discrimination and the variance of the forecasts. The standard version of this alternative decomposition yields biased estimates of these three quantities and a bias-corrected version can be obtained using calculations similar to those described above. Decompositions obtained by conditioning on either forecasts or verifications can be obtained for not only the Brier score, but for any score that takes the form of a mean squared error (Murphy, 1996) or weighted mean squared error (Young, 2010). Again, the standard decompositions are biased, but bias-corrected versions can be derived.

In fact, all proper scores can be decomposed into reliability, resolution and uncertainty terms (Bröcker, 2009). It would be useful to identify the bias of the decomposition for other scores and to construct bias-corrected decompositions where possible. Bröcker (2011) has considered the logarithmic (ignorance) score and the multi-category Brier score. We consider briefly the cases of the ranked probability score (RPS; Epstein, 1969) and the continuous ranked probability score (CRPS; Brown, 1974; Matheson and Winkler, 1976).

The RPS can be written as a sum of Brier scores corresponding to a nested sequence of events (e.g. Toth *et al.*, 2003) and, therefore, a decomposition of the RPS into reliability, resolution and uncertainty terms can be obtained by summing the corresponding terms of these Brier scores. Both standard and bias-corrected decompositions can be formed in this way. The CRPS can be written as an integral of Brier scores corresponding to a nested continuum of events (e.g. Hersbach, 2000) and so the CRPS can be decomposed in a similar manner, integrating the terms of the Brier score decompositions.

These decompositions of the RPS and CRPS, however, are unsatisfactory because they measure the average reliability and resolution of sets of forecasts for binary events instead of the reliability and resolution of the full probability distributions specified by the forecasts. Other decompositions based on the full distributions are preferable (Murphy, 1972; Candille and Talagrand, 2005). It appears to be possible to construct bias-corrected versions of these decompositions too. These alternative decompositions of the RPS and CRPS rely on each distinct forecast distribution being issued several times so that empirical distributions of the corresponding verifications can be constructed. Unless there are very many forecasts, it is therefore often necessary to group similar, rather than identical, forecast distributions (Candille and Talagrand, 2005). This is also often done for the Brier score when the issued forecast probabilities can take any value in the interval $[0, 1]$ instead of only K distinct values. When such grouping is used, Stephenson *et al.* (2008) show that the Brier score obtained by combining the reliability, resolution and uncertainty terms will typically differ from the value obtained by evaluating the Brier score directly from the ungrouped forecasts. In order to retrieve the Brier score for the ungrouped forecasts, it is necessary to generalize the resolution term in the decomposition to account for within-group variation. The same can be expected to be true for the decompositions of the RPS and CRPS when forecasts are grouped. The generalized resolution defined by Stephenson *et al.* (2008) is also biased but, again, a bias-corrected version can be derived. We expect that bias-corrected versions could also be obtained in the cases of the RPS and CRPS. Finally, other decompositions of the RPS and CRPS have been proposed that avoid the need to group forecasts (Hersbach, 2000; Candille and Talagrand, 2005). The bias of these decompositions could be investigated too.

We have assumed throughout that the forecasts and verifications are independent and identically distributed random variables. Temporal dependence is likely to inflate biases and also to reduce the rates at which biases decay to zero. Analysing the biases in the presence of temporal dependence is complicated, however, because the verifications that contribute to the conditional relative frequencies, $\bar{x}_k$, are randomly spaced in time. Whichever decomposition is used, therefore, checking the convergence of the reliability, resolution and uncertainty estimates as the sample size increases is worthwhile. This can be done by plotting against n the estimates calculated from the first n data.

6. Summary

The standard decomposition of the Brier score is biased and we have proposed a simple, bias-corrected decomposition that provides a more accurate description of forecast reliability and resolution when the verification data can be described by independent and identically distributed random variables.

Acknowledgements

This work was funded by NERC Directed Grant NE/H003509/1. Caio Coelho provided the data. Expert comments from Jochen Bröcker, Ian Jolliffe and an anonymous referee helped us to improve the original manuscript.

Appendix

Proofs

There is no unbiased decomposition

If an unbiased estimator for REL_∞ (5) exists then it must be the sum of unbiased estimators for the summands of REL_∞ and these estimators could be subtracted from the summands of REL (2) to obtain unbiased estimators for the summands, $\nu_{k,n}\mu_k(1-\mu_k)$, of the bias (8) of REL, where $\nu_{k,n} = \Pr(n_k > 0) = 1 - (1-\phi_k)^n$ because the distribution of n_k is binomial with parameters n and ϕ_k. An unbiased estimator for $\nu_{k,n}\mu_k(1-\mu_k)$ must be a function of n_k and $\{x_i : i \in I_k\}$ but the order of the x_i carries no information about μ_k or ϕ_k and so we can require this estimator to be a function of n_k and s_k, where $s_k = \sum_{i\in I_k} x_i$ and the conditional distribution of s_k given n_k is binomial with parameters n_k and μ_k. Consider an estimator $g(n_k, s_k)$ with expectation

$$E\{g(n_k,s_k)\} = \sum_{m=0}^{n}\sum_{t=0}^{m} g(m,t)\Pr(s_k = t \mid n_k = m) \times \Pr(n_k = m) = \sum_{m=0}^{n}\binom{n}{m}\phi_k^m(1-\phi_k)^{n-m} \times \sum_{t=0}^{m} g(m,t)\binom{m}{t}\mu_k^t(1-\mu_k)^{m-t}.$$

This polynomial in μ_k and ϕ_k must equal the polynomial

$$\{1-(1-\phi_k)^n\}\mu_k(1-\mu_k) = \sum_{r=1}^{n}\binom{n}{r}(-1)^{r+1}\phi_k^r\mu_k(1-\mu_k)$$

for all μ_k and ϕ_k if $g(n_k, s_k)$ is to be an unbiased estimator for $\nu_{k,n}\mu_k(1-\mu_k)$. This can happen only if, for all $i = 0, 1, \ldots, n$ and $j = 0, 1, \ldots, n$, the coefficients of $\phi_k^i\mu_k^j$ in the two polynomials are equal . The latter polynomial, however, has a non-zero coefficient for $\phi_k\mu_k^2$ and the former polynomial contains no such term. Thus, there is no unbiased estimator for $\nu_{k,n}\mu_k(1-\mu_k)$ and hence no unbiased estimator for REL_∞. A similar argument shows that there is no unbiased estimator for RES_∞.

The bias of the new decomposition

We show first that the uncertainty term (13) in the new decomposition is unbiased. The definitions of UNC (4) and UNC' (13) yield $\mathrm{UNC}' = n\mathrm{UNC}/(n-1)$ while the expressions for UNC_∞ (6) and $E(\mathrm{UNC})$ (7) yield $E(\mathrm{UNC}) = (n-1)\mathrm{UNC}_\infty/n$ so that $E(\mathrm{UNC}') = nE(\mathrm{UNC})/(n-1) = \mathrm{UNC}_\infty$ and

$$\mathrm{bias}(\mathrm{UNC}') = 0. \tag{19}$$

To find the bias of the reliability term (11), write

$$\mathrm{REL}' = \mathrm{REL} - \frac{1}{n}\sum_{k=1}^{K} r_k,$$

where $r_k = n_k\bar{x}_k(1-\bar{x}_k)/(n_k-1)$ if $n_k > 1$ and $r_k = 0$ if $n_k \le 1$. If $n_k \le 1$ then $E(r_k \mid n_k) = 0$. If $n_k > 1$ then

$$\bar{x}_k(1-\bar{x}_k) = \frac{1}{n_k}\sum_{i\in I_k} x_i - \frac{1}{n_k^2}\left(\sum_{i\in I_k} x_i + \sum_{i\in I_k}\sum_{j\in I_k\setminus\{i\}} x_i x_j\right)$$

because $x_i^2 = x_i$ when $x_i = 0$ or 1, and so

$$\begin{aligned}
&E(r_k \mid n_k)\\
&= \frac{n_k}{n_k - 1}\left[\frac{1}{n_k}\sum_{i \in I_k} E(x_i \mid p_i = \pi_k)\right.\\
&\quad - \frac{1}{n_k^2}\left\{\sum_{i \in I_k} E(x_i \mid p_i = \pi_k)\right.\\
&\quad \left.\left. + \sum_{i \in I_k}\sum_{j \in I_k \setminus \{i\}} E(x_i \mid p_i = \pi_k) E(x_j \mid p_j = \pi_k)\right\}\right]\\
&= \frac{n_k}{n_k - 1}\left[\mu_k - \frac{1}{n_k^2}\left\{n_k \mu_k + n_k(n_k - 1)\mu_k^2\right\}\right]\\
&= \mu_k(1 - \mu_k).
\end{aligned}$$

Therefore,

$$\begin{aligned}
E(r_k) &= \mu_k(1 - \mu_k)\Pr(n_k > 1)\\
&= \mu_k(1 - \mu_k)\{\Pr(n_k > 0) - \Pr(n_k = 1)\}
\end{aligned}$$

and, using the bias (8) of REL,

$$\begin{aligned}
\mathrm{bias}(\mathrm{REL}') &= \mathrm{bias}(\mathrm{REL}) - \frac{1}{n}\sum_{k=1}^{K} E(r_k)\\
&= \frac{1}{n}\sum_{k=1}^{K} \Pr(n_k = 1)\mu_k(1 - \mu_k). \qquad (20)
\end{aligned}$$

The bias of RES' equals the bias of REL' because UNC' and the Brier score itself are unbiased: the expectation of the Brier score is independent of n.

Next, we calculate the rate at which the bias of REL', and hence of RES', decays as n increases. From the binomial distribution of n_k, we have

$$\Pr(n_k = 1) = n\phi_k(1 - \phi_k)^{n-1}$$

and, therefore, the bias (20) of REL' is

$$\mathrm{bias}(\mathrm{REL}') = \sum_{k=1}^{K} \phi_k(1 - \phi_k)^{n-1}\mu_k(1 - \mu_k),$$

which decays geometrically as n increases. The leading order terms in the biases for the standard decomposition decay at the much slower rate of $1/n$.

The bias of Bröcker's decomposition

Now we calculate the biases and their rates of decay for the decomposition (14)–(16) proposed by Bröcker (2011). Arguments similar to those above show that

$$\mathrm{bias}(\mathrm{UNC}'') = -\frac{\mu(1 - \mu)}{n^2} \qquad (21)$$

and

$$\mathrm{bias}(\mathrm{REL}'') = \frac{1}{n}\sum_{k=1}^{K} \mu_k(1 - \mu_k)\sum_{m=1}^{n}\frac{1}{m}\Pr(n_k = m) \qquad (22)$$

with $\mathrm{bias}(\mathrm{RES}'') = \mathrm{bias}(\mathrm{REL}'') + \mathrm{bias}(\mathrm{UNC}'')$. The biases of REL'', RES'' and UNC'' all decay to zero at rate $1/n^2$. This is immediate for UNC''. To see that it is true for REL'', and hence for RES'', note that

$$\begin{aligned}
\sum_{m=1}^{n}\frac{1}{m+1}\Pr(n_k = m) &\le \sum_{m=1}^{n}\frac{1}{m}\Pr(n_k = m)\\
&\le 2\sum_{m=1}^{n}\frac{1}{m+1}\Pr(n_k = m)
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{m=1}^{n}\frac{1}{m+1}\Pr(n_k = m)\\
&= \sum_{m=1}^{n}\frac{1}{m+1}\binom{n}{m}\phi_k^m(1 - \phi_k)^{n-m}\\
&= \frac{1/\phi_k}{n+1}\sum_{m=1}^{n}\binom{n+1}{m+1}\phi_k^{m+1}(1 - \phi_k)^{n-m}\\
&= \frac{1/\phi_k}{n+1}\sum_{m=2}^{n+1}\binom{n+1}{m}\phi_k^{m}(1 - \phi_k)^{n+1-m}\\
&= \frac{1/\phi_k}{n+1}\left\{1 - (1 - \phi_k)^{n+1} - (n+1)\phi_k(1 - \phi_k)^n\right\},
\end{aligned}$$

the leading order term of which decays at rate $1/n$. Therefore, $\sum_{m=1}^{n} m^{-1}\Pr(n_k = m)$ decays at rate $1/n$ and $\mathrm{bias}(\mathrm{REL}'')$ decays at rate $1/n^2$.

The ordering of the biases

The ordering (17) on the biases of the uncertainty terms follows immediately from the bias expressions (9), (19) and (21). The ordering (18) on the biases of the reliability terms follows from the bias expressions (8), (20) and (22) because

$$\begin{aligned}\Pr(n_k = 1) &\leq \sum_{m=1}^{n} \frac{1}{m} \Pr(n_k = m)\\ &\leq \sum_{m=1}^{n} \Pr(n_k = m)\\ &= \Pr(n_k > 0).\end{aligned}$$

References

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.

Bröcker J. 2009. Reliability, sufficiency, and the decomposition of proper scores. *Q. J. R. Meteorol. Soc.* **135**: 1512–1519. DOI: 10.1002/qj.456.

Bröcker J. 2011. Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Clim. Dyn.*. In press. DOI: 10.1007/s00382-011-1191-1.

Bröcker J, Smith LA. 2007. Increasing the reliability of reliability diagrams. *Weather and Forecasting* **22**: 651–661. DOI: 10.1175/WAF993.1.

Brown TA. 1974. '*Admissible scoring systems for continuous distributions*,' Technical Note P-5235, 27pp. The Rand Corporation: Santa Monica, California, USA.

Candille G, Talagrand O. 2005. Evaluation of probabilistic prediction systems for a scalar variable. *Q. J. R. Meteorol. Soc.* **131**: 2131–2150. DOI: 10.1256/qj.04.71.

Epstein ES. 1969. A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol.* **8**: 985–987. DOI: 10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

Ferro CAT. 2007. Comparing probabilistic forecasting systems with the Brier score. *Weather and Forecasting* **22**: 1076–1088. DOI: 10.1175/WAF1034.1.

Ferro CAT, Richardson DS, Weigel AP. 2008. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorol. Appl.* **15**: 19–24. DOI: 10.1002/met.45.

Glahn HR, Jorgensen DL. 1970. Climatological aspects of the Brier p-score. *Mon. Weather Rev.* **98**: 136–141. DOI: 10.1175/1520-0493(1970)098<0136:CAOTBP>2.3.CO;2.

Hersbach H. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting* **15**: 559–570. DOI: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hsu W, Murphy AH. 1986. The attributes diagram: a geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting* **2**: 285–293. DOI: 10.1016/0169-2070(86)90048-8.

Matheson JE, Winkler RL. 1976. Scoring rules for continuous probability distributions. *Management Science* **22**: 1087–1096. DOI: 10.1287/mnsc.22.10.1087.

Murphy AH. 1972. Scalar and vector partitions of the ranked probability score. *Mon. Weather Rev.* **100**: 701–708. DOI: 10.1175/1520-0493(1972)100<0701:SAVPOT>2.3.CO;2.

Murphy AH. 1973. A new vector partition of the probability score. *J. Appl. Meteorol.* **12**: 595–600. DOI: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Murphy AH. 1996. General decompositions of MSE-based skill scores: measures of some basic aspects of forecast quality. *Mon. Weather Rev.* **124**: 2353–2369. DOI: 10.1175/1520-0493(1996)124<2353:GDOMBS>2.0.CO;2.

Murphy AH, Winkler RL. 1987. A general framework for forecast verification. *Mon. Weather Rev.* **115**: 1330–1338. DOI: 10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.

Stephenson DB, Coelho CAS, Doblas-Reyes FJ, Balmaseda M. 2005. Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus* **57A**: 253–264. DOI: 10.1111/j.1600-0870.2005.00110.x.

Stephenson DB, Coelho CAS, Jolliffe IT. 2008. Two extra components in the Brier score decomposition. *Weather and Forecasting* **23**: 752–757. DOI: 10.1175/2007WAF2006116.1.

Toth Z, Talagrand O, Candille G, Zhu Y. 2003. Probability and ensemble forecasts. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Jolliffe IT, Stephenson DB. (eds.) John Wiley & Sons: Chichester. pp 137-163.

Uppala SM, Kållberg PW, Simmons AJ, Andrae U, da Costa Bechtold V, Fiorino M, Gibson JK, Haseler J, Hernandez A, Kelly GA, Li X, Onogi K, Saarinen S, Sokka N, Allan RP, Andersson E, Arpe K, Balmaseda MA, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Caires S, Chevallier F, Dethof A, Dragosavac M, Fisher M, Fuentes M, Hagemann S, Hólm E, Hoskins BJ, Isaksen L, Janssen PAEM, Jenne R, McNally AP, Mahfouf J-F, Morcrette J-J, Rayner NA, Saunders RW, Simon P, Sterl A, Trenberth KE, Untch A, Vasiljevic D, Viterbo P, Woollen J. 2005. The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131**: 2961–3012. DOI: 10.1256/qj.04.176.

Young RMB. 2010. Decomposition of the Brier score for weighted forecast-verification pairs. *Q. J. R. Meteorol. Soc.* **136**: 1364–1370. DOI: 10.1002/qj.641.