# Simplifying and generalising Murphy's Brier score decomposition

Stefan Siegert

*Exeter Climate Systems, University of Exeter, EX4 4QE, United Kingdom*

**The decomposition of the Brier score into Reliability, Resolution and Uncertainty has become a standard method in forecast verification. In this note a very simple derivation of the familiar Brier score decomposition is presented. The Reliability and Resolution terms can be calculated as average Brier score differences between the issued forecast, the recalibrated forecast and the climatological reference forecast. The result suggests a simple way to calculate similar decompositions for arbitrary verification scores, and that recalibration methods and reference forecasts can be chosen more flexibly than is generally appreciated. A new decomposition of the continuous ranked probability score (CRPS) is proposed.**

*Key Words:* forecast verification; reliability; resolution; uncertainty; recalibration; decomposition; scoring rules

*Received ...*

## 1. Introduction

Brier (1950) proposed a metric to evaluate a forecaster by comparing a number of $N$ past forecast probabilities $p_1, \ldots, p_N$ to their verifying observations $y_1, \ldots, y_N$. The binary observation is $y_t = 1$ if an event occured at time $t$, and $y_t = 0$ otherwise. The empirical average Brier score (also simply referred to as the "Brier score") is given by the squared difference between forecasts and observations averaged over time:

$$B(p) = \frac{1}{N}\sum_{t=1}^{N}(y_t - p_t)^2. \quad (1)$$

The perfect Brier score of zero is obtained by a forecaster who always issues $p_t = 1$ when $y_t = 1$ and $p_t = 0$ whenever $y_t = 0$. The higher the Brier score, the worse the forecast.

Murphy (1973) proposed a decomposition of the empirical Brier score into the sum of three terms. The issued forecasts $p_t$ are assumed to have only $K$ distinct values, that is $p_t \in \{P_1, \ldots, P_K\}$ for all $t$. Denote by $n_k$ the number of times the $k$th forecast value was issued, and by $o_k$ the total number of events that have occurred when the $k$th forecast value was issued. The average event frequency for the $k$th forecast value is given by $o_k/n_k$. Denote by $\bar{o} = 1/N \sum_{t=1}^{N} y_t$ the climatological event frequency. It will be useful to denote by $k(t)$ the index of the forecast value that was issued at time $t$, that is $p_t = P_{k(t)}$. The empirical Brier score of the forecasts $p_1, \ldots, p_N$ can be decomposed into three components called Reliability (REL), Resolution (RES), and Uncertainty (UNC), that characterise different attributes of the forecast. In particular

$$B(p) = REL - RES + UNC \quad (2)$$

© 0000 Royal Meteorological Society

where

$$REL = \sum_{k=1}^{K} \frac{n_k}{N} \left( \frac{o_k}{n_k} - P_k \right)^2, \quad (3)$$

$$RES = \sum_{k=1}^{K} \frac{n_k}{N} \left( \frac{o_k}{n_k} - \bar{o} \right)^2, \text{ and} \quad (4)$$

$$UNC = \bar{o}(1 - \bar{o}). \quad (5)$$

A reliable forecaster or forecasting system should issue probabilities that are equal to average event frequencies, that is $P_k = o_k/n_k$ for all $k$. A reliable forecaster thus has $REL = 0$, and any violations of reliability lead to $REL > 0$, thus increasing the Brier score. If the conditional event frequencies $o_k/n_k$ were the same for all categories, the different forecast values cannot distinguish events that are more or less likely than average; such an "uninformed" forecaster has $RES = 0$. If conditional event frequencies are different for different forecast values, the forecaster has $RES > 0$ which decreases the Brier score. Reliability and Resolution thus have intuitive interpretations as weighted squared distances in the reliability diagram (Toth et al. 2003).

Similar decompositions have been derived for different verification scores, e.g. the continuous ranked probability score (CRPS, Hersbach 2000), the discrete ranked probability score (Candille and Talagrand 2005), the logarithmic score (Weijs et al. 2010), the quantile score (Bentzien and Friederichs 2014), and the error-spread score (Christensen 2015). Bröcker (2009) has shown that every proper verification score can be decomposed into non-negative components that characterise uncertainty, reliability and resolution.

Murphy (1973) presented a lengthy derivation of the original decomposition, repeatedly solving quadratic equations, and using geometrical arguments. Bröcker (2009) derived the general result using advanced probability calculus. For some practitioners and forecast users these derivations might be difficult to follow. In this note a much simpler derivation of Murphy's original result is presented. The derivation suggests that decomposing arbitrary verification scores similar to the Brier score poses no mathematical difficulties. A number of previous results about score decompositions are discussed. Decomposing the Brier Score and CRPS of seasonal temperature forecasts serves as an illustration.

## 2. Key result

Suppose the issued forecasts $p_t$ are uncalibrated, that is the forecast values $P_k$ are not equal to average event frequencies $o_k/n_k$. A simple and straightforward recalibration method would be to replace the forecast $p_t$ by the conditional event frequency for that forecast value. The recalibrated forecasts form a new set of forecasts, denoted $q_1, \ldots, q_N$, with values $q_t = o_{k(t)}/n_{k(t)}$. Furthermore, it is common practice to benchmark the issued forecasts $p_t$ against easily available reference forecasts $r_1, \ldots, r_N$. The most commonly used benchmark forecast is the constant climatological event frequency, that is $r_t = \bar{o}$ for all $t$. The issued forecasts $p_t$, as well as the newly constructed forecasts $q_t$ and $r_t$ have average Brier scores $B(p)$, $B(q)$, and $B(r)$ when compared to the observations $y_1, \ldots, y_N$. By adding and subtracting identical terms, the Brier Score $B(p)$ can trivially be written as

$$B(p) = [B(p) - B(q)] - [B(r) - B(q)] + [B(r)]. \quad (6)$$

The key result of this note is that the three terms in squared brackets are identical to the components of the original Murphy (1973) Brier score decomposition, that is

$$REL = B(p) - B(q), \quad (7)$$

$$RES = B(r) - B(q), \text{ and} \quad (8)$$

$$UNC = B(r). \quad (9)$$

Proofs are in the appendix.

## 3. Discussion

Equation 6 is true for arbitrary choices of the function $B$, not just the Brier score. The decomposition therefore extends to arbitrary verification scores. The decomposition only requires the calculation of the recalibrated forecasts $q_1, \ldots, q_N$ and the reference forecasts $r_1, \ldots, r_N$. After calculating average scores of

© 0000 Royal Meteorological Society 

$q_t$ and $r_t$ over the observations $y_1, \ldots, y_N$, the decomposition into Uncertainty, Reliability, and Resolution can be calculated using equations 7–9. The decomposition into Reliability, Resolution and Uncertainty is thus not a special feature of the Brier score (which has long been known). The remarkable fact about the Brier score is, however, that average score differences can be rewritten in the forms of equations 3 and 4, such that these terms become interpretable as distances in the reliability diagram, between the calibration curve of the forecast, the diagonal, and the horizontal no-skill line (Toth et al. 2003).

The decomposition written as in eq. 6 is implied in the results of Bröcker (2009), who showed that every proper verification score gives rise to a non-negative divergence function to measure the distance between probability forecasts. The divergence function $d(p, q)$ defined by the verification score $S(p, y)$ is the expected score difference $E_y[S(p, y) - S(q, y)]$, where $q$ is assumed to be the true distribution of $y$, and $E_y$ denotes expectation over the random variable $y$ (Gneiting and Raftery 2007). If $S$ is a proper score, $d$ is non-negative because the best (lowest) expected score is achieved by forecasting the true distribution of $y$. The expected score of the climatology $E_y S(\bar{o}, y)$ is called the entropy, and is used as a measure of lack of information. Bröcker (2009) showed that a verification score can be decomposed into Reliability, Resolution, and Uncertainty using its corresponding entropy and divergence function. By expanding the expected score of $p$ as in equation 6, and defining $q$ to be the conditional distribution of $y$ given $p$, we get

$$\begin{aligned} E_{p,y}[S(p,y)] =& E_{p,y}\{[S(p,y) - S(q,y)] \\ & - [S(\bar{o},y) - S(q,y)] + S(\bar{o},y)\} \\ =& E_p d(p,q) - E_p d(\bar{o},q) + E_y S(\bar{o},y) \end{aligned} \quad (10)$$

which is one of the key results of Bröcker (2009). The divergence $E_y[S(p, y) - S(q, y)]$ can be calculated analytically for some verification scores (Bröcker 2008); for example, $d(p, q) = (p - q)^2$ for the Brier score. If the divergence function is not available in closed form, it can be estimated by the average score difference, using a suitable recalibration method to estimate $q$, the true distribution of the observation. Bentzien and Friederichs (2014), for example, used average score differences to calculate the divergence function of the quantile score.

Equation 6 holds true for arbitrary choices of the forecasts $q_t$ and $r_t$. However, in order to interpret score differences as measures of reliability and resolution, the choice of $q_t$ and $r_t$ must be justified. The forecast $q_t$ should be a recalibration of $p_t$, that is, a function of $p_t$ chosen in such a way to remove its systematic violations of reliability. The within-category frequency $o_{k(t)}/n_{k(t)}$ used for the original decomposition is but one possible method. In Bröcker (2012) it is shown that estimating the calibration using the within-category frequency can lead to problems if the number of forecast categories is large, such that $K \approx N$. Alternative methods exist to recalibrate probability forecasts for binary events, e.g. kernel density estimation (Bröcker 2008) or logistic regression (Wilks 2011, ch. 7), which avoid problems due to over-fitting.

The reference forecast $r_t$ should be an easily available "fallback" forecast that would be issued if the forecasts $p_t$ were unavailable. Comparing the quality of $p_t$ and $r_t$ quantifies the additional value of $p_t$, as is often done by calculating skill scores (Wilks 2011, ch. 8). The average climatological frequency $\bar{o}$ is often a poor choice as a reference forecast. If the observations exhibit time-series features such as persistence, trends, or seasonality, these patterns can be (and should be) exploited to construct the fallback forecast $r_t$. Using a more skilful reference forecast than the climatology ensures that the added value of $p_t$ is not overestimated.

In Murphy's original decomposition, Reliability and Resolution are guaranteed to be non-negative. By writing the components of the original decomposition as score differences, we have shown that the average Brier score of the recalibrated forecasts $q_t = o_{k(t)}/n_{k(t)}$ is always at least as good as the average Brier scores of $p_t$ and $r_t$. But if we allow for arbitrary recalibrations and reference forecasts, and calculate Reliability and Resolution as score differences, the terms are not a priori guaranteed to be non-negative. However, a simple argument can be used to ensure non-negativity nonetheless. Suppose we make a choice for estimating the recalibrated forecasts $q_t$ that has a worse score than the original forecast $p_t$, that is we get $REL < 0$ from equation 7. This means that there is no benefit in recalibrating $p_t$, and we should

© 0000 Royal Meteorological Society 

choose $q_t = p_t$ for all $t$. Then we get $REL = 0$, an indication that recalibration is unnecessary, which implies that the forecast is already calibrated. Similarly, if we make a choice for $r_t$ that has a better score than our chosen recalibration $q_t$, this means we can further improve the recalibration by setting $q_t = r_t$. Then we have $RES = 0$, which is often taken as an indication that after recalibration the issued forecasts $p_t$ offer no improvement over the reference forecast. In summary, non-negativity of REL and RES can always be guaranteed by allowing both $q_t = p_t$ and $q_t = r_t$ (for all $t$) as possible recalibration schemes.

Forecasts $p_t$ are often issued as continuous quantities, such that every value of $p_t$ occurs only once. To calculate the Brier score decomposition it is common practice to bin the forecast probabilities into a finite number of categories to calculate the recalibrated forecasts by average event frequencies per bin (Bröcker 2012). To calculate the Reliability term, the within-bin averages of the forecast probabilities are substituted for $P_k$ in equation 3. Stephenson et al. (2008) have pointed out that the components of Murphy's original decomposition do not add up to $B(p)$ if continuous forecasts are binned. The mismatch can be understood by realising that the within-bin averages define a new set of forecasts $\bar{p}_1, \ldots, \bar{p}_N$ that obtain an average Brier score $B(\bar{p})$ when evaluated against $y_1, \ldots, y_N$. $B(\bar{p})$ is generally different from $B(p)$. By substituting the within-bin averages for $P_k$ in equation 3, we are effectively calculating the decomposition of $B(\bar{p})$ instead of $B(p)$. The Brier score $B(p)$ of the issued forecasts can be decomposed into score differences by introducing an additional term as follows:

$$B(p) = [B(\bar{p}) - B(q)] - [B(r) - B(q)] + [B(r)] + [B(p) - B(\bar{p})] \tag{11}$$

The first three terms in square brackets on the rhs are the Brier score components of $B(\bar{p})$. These components add up to $B(p)$ if the issued forecasts $p_t$ are equal to their within-bin averages, such that $B(p) = B(\bar{p})$. If there is any within-bin variability of the forecasts, the residual term $B(p) - B(\bar{p})$ can be non-zero, and the components of $B(\bar{p})$ do not add up to $B(p)$. Stephenson et al. (2008) showed that the residual can be further decomposed into two terms that depend on the within-bin variances of $p_t$ and the within-bin covariances between $p_t$ and $y_t$. Instead of introducing extra terms, one can use score differences and equation 6 to decompose $B(p)$ into components that add up exactly – replacing the original forecasts by their within-bin averages becomes unnecessary. Note that, if score differences are used, the residual term $B(p) - B(\bar{p})$ is merged into the Reliability component of the Brier score decomposition of $B(\bar{p})$: With score differences, we get $REL = B(p) - B(q) = [B(\bar{p}) - B(q)] + [B(p) - B(\bar{p})]$, where the first term $B(\bar{p}) - B(q)$ is the Reliability term of the decomposition of $B(\bar{p})$. Stephenson et al. (2008), on the other hand, suggested to merge the residual into the Resolution term.

The decomposition of the CRPS for ensemble forecasts proposed by Hersbach (2000) was not derived using average score differences. The Hersbach (2000) decomposition uses the CRPS of the climatological distribution to define the Uncertainty term, and defines the Reliability in terms of deviations from flatness of the rank histogram. The Resolution term is defined as the remainder which completes the decomposition. This decomposition of the CRPS is unsatisfactory for a number of reasons. Firstly, the Resolution component is defined "somewhat artificially" (Hersbach 2000), and can even be negative. Secondly, it is known (Hamill 2001) that unreliable ensembles can produce flat rank histograms, and thus appear reliable under the Hersbach (2000) CRPS decomposition. Lastly, the decomposition gives no guidance how to recalibrate the ensemble in order to achieve the "potential" CRPS. The general framework for score decomposition outlined in this note can solve these problems. A suitable reference forecast for the continuous observation (such as the climatological distribution) must be chosen. A suitable recalibration method for the ensemble forecasts (such as non-homogeneous Gaussian regression (NGR, Gneiting et al. 2005)) is applied, to flatten the rank histogram, and to correct known conditional biases. Using a parametric recalibration method such as NGR also avoids the curse of dimensionality encountered when ensemble forecasts are recalibrated by binning the space of possible forecasts (Candille and Talagrand 2005). After defining the reference forecasts and the recalibrated forecasts, the decomposition follows by taking average CRPS differences. This would make the recalibration method explicit and produce a more interpretable Resolution component. On the other hand, the explicit relationship between the CRPS Reliability and the

© 0000 Royal Meteorological Society 

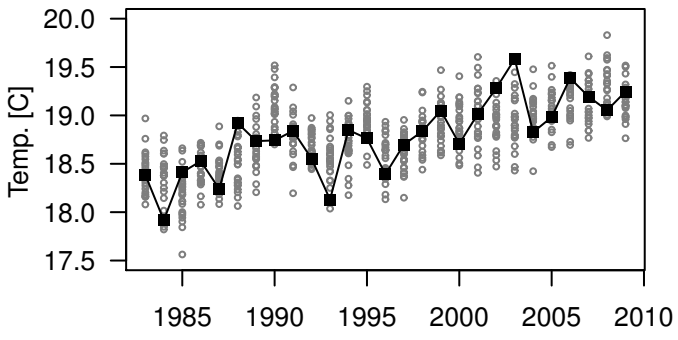


**Figure 1.** Reanalysis data (black line) and 3-months ahead ensemble forecasts (gray markers) of average European summer temperatures.

rank histogram would be lost, and the decomposition would be sensitive to the choice of the recalibration method.

## 4. Application

To illustrate and explore the ideas developed in the previous section, we analyse ensemble forecasts of European summer temperatures produced by the NCEP climate forecast system version 2 (Saha et al. 2014). Data from the NCEP climate forecast system reanalysis (Saha et al. 2010) were taken as verifying observations. Ensemble forecasts of 24 members were issued each year from 1983–2009 ($N = 27$), with initialisation dates between 11 April and 6 May. The forecast target is near-surface air temperature over Europe, averaged spatially over the rectangular region $30°$N, $75°$N, $12.5°$W, $42.5°$E, and averaged temporally over the summer months June, July, August. The forecast lead time is thus 1–3 months. A warm bias of $0.64$K was removed from all ensemble members. The bias was removed because it is often considered to be a trivial source of unreliability that can easily be eliminated. We shall be interested in violations of Reliability beyond the mean bias, and treat the mean-debiased ensemble as the "raw ensemble" from now on. Time series of ensemble members and observations are plotted in Figure 1.

The goal of this section is to illustrate ideas, rather than a thorough analysis of the forecast skill of this particular system. To keep things simple, we ignore the important aspects of out-of-sample evaluation and uncertainty assessment. We first apply the CRPS decomposition proposed in the previous section to the ensemble forecasts, and compare the results to the decomposition by Hersbach (2000). Then we transform the continuous temperature data into a binary prediction problem by thresholding, to study different approaches to decomposing the Brier Score.

The continuous ranked probability score (CRPS; Matheson and Winkler 1976) of a series of cumulative forecast distributions $F_1, \ldots, F_N$ and real-valued observation $y_1, \ldots, y_N$ is given by

$$CRPS = \frac{1}{N} \sum_{t=1}^{N} \int_{-\infty}^{\infty} dx \, |F_t(x) - H(x - y_t)|^2, \qquad (12)$$

where $H(x)$ is the Heaviside step function. The CRPS has the same units as the prediction target (i.e. Kelvin for temperature forecasts). The CRPS of a series of ensemble forecasts $x_1, \ldots, x_N$, each with members $x_t = (x_{t,1}, \ldots, x_{t,R})$ and verifying observations $y_1, \ldots, y_N$, is calculated by

$$\frac{1}{N} \sum_{t=1}^{N} \left( \frac{1}{R} \sum_{r=1}^{R} |x_{t,r} - y_t| - \frac{1}{2R^2} \sum_{r,r'=1}^{R} |x_{t,r} - x_{t,r'}| \right) \qquad (13)$$

(Gneiting and Raftery 2007), which is equivalent to the expression used by Hersbach (2000). Gneiting and Raftery (2007) also show that the CRPS of forecasts issued as a Normal distributions $\mathcal{N}(\mu_t, \sigma_t^2)$, with means $\mu_1, \ldots, \mu_N$ and variances $\sigma_1^2, \ldots, \sigma_N^2$ is given by

$$\frac{1}{N} \sum_{t=1}^{N} \sigma_t \left[ z_t (2\Phi(z_t) - 1) + 2\varphi(z_t) - \pi^{-\frac{1}{2}} \right], \qquad (14)$$

where $z_t = \frac{y_t - \mu_t}{\sigma_t}$, and $\varphi(x)$ and $\Phi(x)$ are the standard Normal density function and distribution function, respectively.

The average CRPS of the raw (uncalibrated) ensemble forecasts, calculated by eq. 13, is $0.138$K. To decompose the CRPS of the ensemble forecast, denoted $CRPS(p)$, we apply eq. 6 to the CRPS, that is, we define Reliability, Resolution, and Uncertainty by

$$\begin{aligned} REL &= CRPS(p) - CRPS(q) \\ RES &= CRPS(r) - CRPS(q) \\ UNC &= CRPS(r) \end{aligned} \qquad (15)$$

where $p$, $q$, and $r$ are the raw ensemble forecast, the recalibrated forecast, and the reference forecast, respectively. The components trivially satisfy $CRPS(p) = REL - RES + UNC$. To uniquely define the decomposition, a recalibration method and a reference

© 0000 Royal Meteorological Society 

Table 1. Average CRPS of the raw ensemble, the recalibrated forecast ($q_{\mathrm{NGR}}$), the climatological forecast ($r_{\mathrm{clim}}$) and the persistence forecast ($q_{\mathrm{pers}}$).

| forecast | raw | $q_{\mathrm{NGR}}$ | $r_{\mathrm{clim}}$ | $r_{\mathrm{pers}}$ |
|---|---|---|---|---|
| CRPS [K] | 0.138 | 0.136 | 0.22 | 0.18 |

forecast have to be specified. For recalibration of the ensemble forecasts, we use non-homogeneous Gaussian regression (NGR; Gneiting et al. 2005). NGR assumes that the forecast at time $t$ is a Normal distribution, whose mean and variance depend linearly on the ensemble mean $m_t$ and ensemble variance $v_t$, respectively. The parameters are estimated by numerical minimisation of the CRPS, and so the recalibrated forecast $q_{\mathrm{NGR}}$ at time $t$ is given by

$$q_{\mathrm{NGR},t} = \mathcal{N}\left(-0.53K + 1.03m_t, -0.04K^2 + 2.10v_t\right). \quad (16)$$

The average CRPS of the NGR-recalibrated forecast is $0.136$K, a slight improvement over the CRPS of the raw ensemble.

We further calculate two possible reference forecasts. Firstly, we calculate the climatological distribution $r_{\mathrm{clim}}$ by the empirical distribution function over all observations. The CRPS of $r_{\mathrm{clim}}$ is calculated using eq. 13, assuming that the same ensemble forecast $x_t = (y_1, \ldots, y_N)$ is issued for all $t$; we obtain $CRPS(r_{\mathrm{clim}}) = 0.22K$. The previous section mentioned that the climatological distribution can be a poor choice as a reference forecast if the observations exhibit obvious time-series features. To account for the clear trend of the temperature data (see Fig. 1), we consider persistence as an alternative reference forecast. Specifically, we fit a first-order auto-regressive model to the observations, that is the forecast at time $t$ is a Normal distribution whose mean depends linearly on the observation at time $t-1$. The forecast variance is constant. Using minimum CRPS parameter estimation, we obtain the persistence forecast for time $t$ by

$$r_{\mathrm{pers},t} = \mathcal{N}\left(8.55K + 0.55y_{t-1}, 0.11K^2\right). \quad (17)$$

The CRPS of $r_{\mathrm{pers}}$ is $0.18K$. The CRPS values of all the different temperature forecasts are summarised in Table 1.

We decompose the CRPS of the ensemble forecasts into Reliability, Resolution and Uncertainty using different methods. To calculate the Hersbach (2000) decomposition we use the corresponding function in the `R`-package `verification` (NCAR - Research Applications Laboratory 2015). We further use the CRPS values given in Table 1 to calculate two different decompositions based on score differences as in eq. 15. The first decomposition uses climatology as a reference forecast, $r = r_{\mathrm{clim}}$, and the second decomposition uses persistence as a reference forecast, $r = r_{\mathrm{pers}}$. The components of the three decompositions are summarised in Table 2. We note the following:

Table 2. Three decompositions of the CRPS of the temperature ensemble forecast. The notation $a(-b)$ stands for $a \times 10^{-b}$.

| Decomposition | REL [K] | RES [K] | UNC [K] |
|---|---|---|---|
| Hersbach (2000) | $3.07(-3)$ | $8.01(-2)$ | $2.15(-1)$ |
| Eq. 15 ($q_{\mathrm{NGR}}$, $r_{\mathrm{clim}}$) | $1.61(-3)$ | $7.87(-2)$ | $2.15(-1)$ |
| Eq. 15 ($q_{\mathrm{NGR}}$, $r_{\mathrm{pers}}$) | $1.61(-3)$ | $4.29(-2)$ | $1.79(-1)$ |

1. The numerical values of the Reliability and Resolution components of the Hersbach (2000) decomposition, and of the decomposition into score differences with $r = r_{\mathrm{clim}}$ and $q = q_{\mathrm{NGR}}$ differ, but not by much. The similarity is surprising, since the decompositions are motivated differently. Larger differences might occur in different data sets.
2. The effect of recalibrating the forecast is small, leading to an improvement of CRPS only on the order of $10^{-3}$K. This is reflected in a relatively small Reliability term – the ensemble appears to be well-calibrated.
3. The Resolution terms are larger than the Reliability terms in all decompositions, indicating that the ensemble forecast is more skilful than either reference forecast.
4. Using persistence as a reference forecast in the decomposition ($r = r_{\mathrm{pers}}$) reduces both the Uncertainty and the Resolution term compared to the decomposition with $r = r_{\mathrm{clim}}$. Smaller Uncertainty indicates higher inherent predictability of the observations due to the presence of a trend. Smaller Resolution implies smaller forecast skill of the recalibrated forecast compared to the reference forecast. That is, using a more skilful reference forecast reduces the perceived difficulty in forecasting the observations, and the perceived value of the forecasting system whose score is decomposed. The Reliability term does not depend on the choice of the reference forecast.

© 0000 Royal Meteorological Society

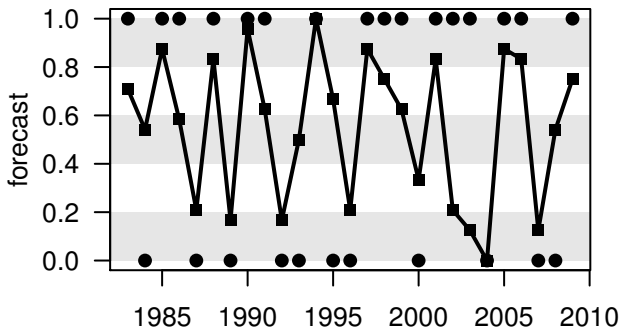


**Figure 2.** Probability forecasts $p_t$ (solid line and squares) and corresponding binary observations $y_t$ (filled circles) for 1983–2009. The gray and white bands indicate the 5 bins used to calculate the Murphy (1973) Brier Score decomposition.

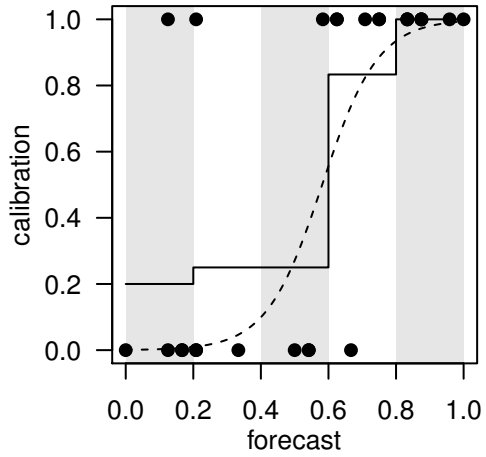


**Figure 3.** Calibration curves estimated by binning and counting ($q_{\rm bin}$, solid line) and logistic regression ($q_{\rm lr}$, dashed line). Circles are observations $y_t$ plotted over the issued forecast probabilities $p_t$.

We next transform the ensemble forecasts to binary forecasts by asking "Will this year's temperature exceed last year's temperature"? The binary observation $y_t$ equals one (zero) if the observed temperature at time $t$ is larger (smaller) than at time $t-1$. Probability forecasts $p_1, \ldots, p_N$ are generated from the ensemble by taking the fraction of ensemble members at time $t$ that exceed the observed temperature of the previous year $t-1$. The binary observations $y_1, \ldots, y_N$ and probability forecasts $p_1, \ldots, p_N$ are shown in Figure 2. To calculate the Murphy (1973) Brier Score decomposition, we bin the forecast probabilities into 5 bins of equal width (indicated by gray and white bands in Figure 2). To calculate the Murphy (1973) decomposition, the within-bin averages of the forecast probabilities in the five bins $(0.12, 0.24, 0.54, 0.69, 0.89)$ are substituted for $P_1, \ldots, P_5$ in eq. 3.

Figure 3 shows two different calibration curves: The solid line is the calibration curve based on within-bin average event frequencies used for the Murphy (1973) Brier Score decomposition in eqs. 3 and 4. The calibrated forecast $q_{\rm bin}$ at time $t$ is given by

$$q_{{\rm bin},t} = \frac{o_{k(t)}}{n_{k(t)}}, \tag{18}$$

where $o_1, \ldots, o_5 = (1, 1, 1, 5, 8)$ and $n_1, \ldots, n_5 = (5, 4, 4, 6, 8)$. The average Brier score of $q_{\rm bin}$ is $0.116$. The dashed line in Figure 3 corresponds to recalibration by logistic regression (Wilks 2011, ch. 7), where the conditional probability of $y_t = 1$ is modelled by a logistic function of the uncalibrated forecast $p_t$. Using minimum Brier Score parameter estimation, we obtain the calibrated forecast $q_{\rm lr}$ at time $t$ by

$$q_{{\rm lr},t} = \left[1 + \exp(7.07 - 12.15 p_t)\right]^{-1}. \tag{19}$$

The average Brier score of $q_{\rm lr}$ is $0.128$, which is worse than the Brier Score of $q_{\rm bin}$. (Note that maximum likelihood estimation yields parameter values $2.81$ and $-6.05$, and an average Brier Score of $0.138$.) We use the climatological event frequency as the reference forecast, that is, $r_{\rm clim} = \bar{o} = 0.59$. The Brier Score of $r_{\rm clim}$ is $0.241$. The Brier scores of the different forecasts are summarised in Table 3.

We calculate Brier score decompositions using the original method proposed by Murphy (1973), and using score differences as in eq. 6, with either $q = q_{\rm bin}$ or $q = q_{\rm lr}$. The decompositions are summarised in Table 4. We note the following:

1. The components of the Murphy (1973) Brier score decomposition do not add up exactly to $B(p)$. The small discrepancy of $-7 \times 10^{-5}$ and can be further

Table 3. Average Brier Score of the raw probabilities ($p$), the two recalibrated forecasts ($q_{\rm bin}$ and $q_{\rm lr}$) and the climatological forecast ($r_{\rm clim}$).

| forecast | $p$ | $q_{\rm bin}$ | $q_{\rm lr}$ | $r_{\rm clim}$ |
|---|---|---|---|---|
| Brier Score | 0.139 | 0.116 | 0.128 | 0.241 |

Table 4. Reliability and Resolution terms or Brier Score decompositions using the established method by Murphy (1973), and using score differences as in eq. 6. Within-bin event frequencies ($q = q_{\rm bin}$) and logistic regression estimates ($q = q_{\rm lr}$) were used as the recalibrated forecasts.

| Decomposition | REL | RES | UNC |
|---|---|---|---|
| Murphy (1973) | 0.02252 | 0.125 | 0.241 |
| Eq. 6 ($q_{\rm bin}$, $r_{\rm clim}$) | 0.02245 | 0.125 | 0.241 |
| Eq. 6 ($q_{\rm lr}$, $r_{\rm clim}$) | 0.010 | 0.113 | 0.241 |

© 0000 Royal Meteorological Society

decomposed into the difference between the within-bin-variance ($2.86 \times 10^{-3}$) and within-bin covariance ($2.93 \times 10^{-3}$); see Stephenson et al. (2008) for definitions of these terms.

2. The Resolution and Uncertainty components obtained by the Murphy (1973) decomposition and by score differences with $q = q_{\mathrm{bin}}$ are identical. The only difference is in the Reliability terms.
3. The components of the decompositions into score differences by eq. 6 add up to the Brier Score of $p_t$ exactly – unlike the Murphy (1973) decomposition with binned forecasts.
4. The Reliability and Resolution terms change if logistic regression is used for forecast recalibration instead of the usual binning approach. Logistic regression does not improve the score of the uncalibrated forecasts $p_t$ as much as the binning approach does. Therefore, the Reliability term with $q = q_{\mathrm{lr}}$, is closer to zero – The forecasts appear more reliable than with $q = q_{\mathrm{bin}}$. At the same time, the Resolution terms decrease, indicating a smaller potential improvement over the reference forecast.

## 5. Conclusion

A simple derivation of the popular Brier score decomposition into Reliability, Resolution, and Uncertainty, originally due to Murphy (1973), has been presented. The components of the decomposition can be calculated by taking average score differences. Other than being simpler than the original derivation, it also simplifies the interpretation of the components, sheds new light on existing results on decomposition of verification scores, and allows for a straightforward generalisation to arbitrary verification scores, arbitrary recalibration methods, and arbitrary reference forecasts. There is some evidence that previous authors have used the proposed strategy to calculate score decompositions, but it was not formulated explicitly, and does not seem to be widely known. The simple method of motivating score decomposition can be used to resolve the non-additivity of the Brier score components of binned forecasts, and gives directions for a new method to decompose the CRPS. An application to seasonal temperature forecasts shows that the proposed methodology yields similar results as traditional decompositions, while being easier to calculate and more flexible in the choice of recalibration method and reference forecast.

In summary, for score decomposition it is sufficient to calculate recalibrated forecasts $q_1, \ldots, q_N$, and suitable reference forecasts $r_1, \ldots, r_N$. The decomposition of the average score of the forecast $p$ follows (rather trivially) by calculating average scores and score differences of $p$, $q$ and $r$.

## Acknowledgments

Forecast and observation data were downloaded through the ECOMS user data gateway R-interface (Santander Meteorology Group 2015). Data and analyses (R-code) can be requested via email to s.siegert@exeter.ac.uk. I benefited from fruitful discussions with David Stephenson, Chris Ferro, Keith Mitchell, Theo Economou, Phil Sansom, Louis-Philippe Caron, Rob Williams and Jochen Bröcker. Comments from the editor and two anonymous reviewers greatly helped to improve the quality of the manuscript. Funding from the European Union Programme FP7/2007-13 under grant agreement 3038378 (SPECS) is gratefully acknowledged.

## Appendix: Proofs

The Brier score difference between the original forecast $p_t$ and the recalibrated forecast $q_t = o_{k(t)}/n_{k(t)}$ is equal to Murphy's original Reliability term:

$$B(p) - B(q) = \frac{1}{N}\sum_{t=1}^{N}\left[(y_t - p_t)^2 - \left(y_t - \frac{o_{k(t)}}{n_{k(t)}}\right)^2\right] \tag{20}$$

$$= \frac{1}{N}\sum_{t=1}^{N}\left[p_t^2 - 2y_t p_t - \frac{o_{k(t)}^2}{n_{k(t)}^2} + 2y_t\frac{o_{k(t)}}{n_{k(t)}}\right] \tag{21}$$

$$= \frac{1}{N}\sum_{k=1}^{K}\left[n_k P_k^2 - 2o_k P_k - n_k\frac{o_k^2}{n_k^2} + 2o_k\frac{o_k}{n_k}\right] \tag{22}$$

$$= \sum_{k=1}^{K}\frac{n_k}{N}\left(\frac{o_k}{n_k} - P_k\right)^2. \tag{23}$$

The Brier score difference between the climatological forecast $r_t = \bar{o}$ and the recalibrated forecast $q_t = o_{k(t)}/n_{k(t)}$ is equal to Murphy's original Resolution term:

© 0000 Royal Meteorological Society

$$B(r) - B(q) = \frac{1}{N}\sum_{t=1}^{N}\left[(y_t - \bar{o})^2 - \left(y_t - \frac{o_{k(t)}}{n_{k(t)}}\right)^2\right] \quad (24)$$

$$= \frac{1}{N}\sum_{t=1}^{N}\left[\bar{o}^2 - 2y_t\bar{o} - \frac{o_{k(t)}^2}{n_{k(t)}^2} + 2y_t\frac{o_{k(t)}}{n_{k(t)}}\right] \quad (25)$$

$$= \frac{1}{N}\sum_{k=1}^{K}\left[n_k\bar{o}^2 - 2o_k\bar{o} - n_k\frac{o_k^2}{n_k^2} + 2o_k\frac{o_k}{n_k}\right] \quad (26)$$

$$= \sum_{k=1}^{K}\frac{n_k}{N}\left(\frac{o_k}{n_k} - \bar{o}\right)^2. \quad (27)$$

For completeness, we derive the well-known result that the Brier score of the climatological forecast $r_t = \bar{o}$ is equal to Murphy's original Uncertainty term:

$$B(r) = \frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{o})^2 \quad (28)$$

$$= \frac{1}{N}\sum_{t=1}^{N}\left(y_t - 2y_t\bar{o} - \bar{o}^2\right) \quad (29)$$

$$= \bar{o} - 2\bar{o}^2 + \bar{o}^2 = \bar{o}(1-\bar{o}). \quad (30)$$

## References


S. Bentzien and P. Friederichs. Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140 (683):1924–1934, 2014. doi:10.1002/qj.2284.

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

J. Bröcker. Some remarks on the reliability of categorical probability forecasts. *Monthly Weather Review*, 136(11):4488–4502, 2008. doi:10.1175/2008MWR2329.1.

J. Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, 2009. doi:10.1002/qj.456.

J. Bröcker. Probability forecasts. In I. T. Jolliffe and D. B. Stephenson, editors, *Forecast Verification: A Practitioner's Guide in Atmospheric Science (2nd ed.)*, chapter 7. John Wiley & Sons, Ltd, 2012.

G. Candille and O. Talagrand. Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, 131(609):2131–2150, 2005. doi:10.1256/qj.04.71.

H. M. Christensen. Decomposition of a new proper score for verification of ensemble forecasts. *Monthly Weather Review*, 143(5):1517–1532, 2015. doi:10.1175/MWR-D-14-00150.1.

T. Gneiting, A. E. Raftery, A. H. Westveld, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005. doi:10.1175/MWR2904.1.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, Mar 2007. doi:10.1198/016214506000001437.

T. M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560, 2001. doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

H. Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000. doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, Jun 1976. doi:10.1287/mnsc.22.10.1087.

A.H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12:595–600, 1973. doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

NCAR - Research Applications Laboratory. *verification: Forecast verification utilities.*, 2015. URL `https://CRAN.R-project.org/package=verification`. R package version 1.42.

Suranjana Saha, Shrinivas Moorthi, Hua-Lu Pan, Xingren Wu, Jiande Wang, Sudhir Nadiga, Patrick Tripp, Robert Kistler, John Woollen, David Behringer, and Coauthors. The NCEP Climate Forecast System Reanalysis. *Bulletin of the American Meteorological Society*, 91(8):1015–1057, Aug 2010. doi:10.1175/2010bams3001.1.

Suranjana Saha, Shrinivas Moorthi, Xingren Wu, Jiande Wang, Sudhir Nadiga, Patrick Tripp, David Behringer, Yu-Tai Hou, Hui-ya Chuang, Mark Iredell, and Coauthors. The NCEP Climate Forecast System Version 2. *J. Climate*, 27(6):2185–2208, Mar 2014. doi:10.1175/jcli-d-12-00823.1.

Santander Meteorology Group. *ecomsUDG.Raccess: R interface to the ECOMS User Data Gateway*, 2015. URL `http://meteo.unican.es/trac/wiki/udg/ecoms`. R package version 4.2-0.

D. B. Stephenson, C. A. S. Coelho, and I. T. Jolliffe. Two extra components in the Brier score decomposition. *Weather and Forecasting*, 23(4):752–757, 2008. doi:10.1175/2007WAF2006116.1.

Z. Toth, O. Talagrand, G. Candille, and Y. Zhu. Probability and ensemble forecasts. In I. T. Jolliffe and D. B. Stephenson, editors, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, chapter 7. John Wiley & Sons, Ltd, 2003.

S. V. Weijs, R. van Nooijen, and N. van de Giesen. Kullback-Leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition. *Monthly Weather Review*, 138(9):3387–3399, 2010. doi:10.1175/2010MWR3229.1.

D. S. Wilks. *Statistical methods in the atmospheric sciences (3rd ed.)*. Academic press, 2011.


© 0000 Royal Meteorological Society 