Variance estimation for Brier Score decomposition[*]

Stefan Siegert[†]

Draft version: July 19, 2013

Abstract

The Brier Score is a widely-used criterion to assess the quality of probabilistic predictions of binary events. The expectation value of the Brier Score can be decomposed into the sum of three components called reliability, resolution, and uncertainty which characterize different forecast attributes. Given a dataset of forecast probabilities and corresponding binary verifications, these three components can be estimated empirically. Here, propagation of uncertainty is used to derive expressions that approximate the sampling variances of the estimated components. Variance estimates are provided for both the traditional estimators, as well as for refined estimators that include a bias correction. Applications of the derived variance estimates to artificial data illustrate their validity, and application to a meteorological prediction problem illustrates a possible use case. The observed increase of variance of the bias-corrected estimators is discussed.

1 Introduction

The basis of the following discussion is a data set of forecast probabilities $\{p_n\}_{n=1}^N$, and corresponding verifications $\{y_n\}_{n=1}^N$. A binary prediction setting is assumed, that is, the verification at instance n, y_n, is either one if the event happens, or zero if it does not happen. The forecast probability p_n is a probabilistic prediction for the event $y_n = 1$. The empirical Brier Score (Brier, 1950) assigned to the set of forecasts $\{p_n\}$ is given by

$$\mathrm{Br} = \frac{1}{N}\sum_{n=1}^{N}(p_n - y_n)^2. \tag{1}$$

The Brier Score is negatively oriented, assigning lower values to better forecasts. The Brier Score further has the property of being *proper*, which means that a forecaster cannot improve his expected Brier Score by issuing forecasts q that differ from his best estimates p of the actual event probabilities. In fact, any such deviance from p will increase his expected Brier Score, which makes the Brier Score a *strictly proper* scoring rule (DeGroot and Fienberg, 1983).

It has been shown by Murphy (1973) that the Brier Score can be decomposed additively into three non-negative terms, called reliability, resolution, and uncertainty:

$$\mathrm{Br} = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC}. \tag{2}$$

A qualitative interpretation of the individual components is given next; mathematical details follow below. The reliability term quantifies how far the forecast probabilities p_n differ from the corresponding conditional event probabilities $\mathbb{P}(y_n = 1 \mid p_n)$. Ideally, it should always hold that $p_n = \mathbb{P}(y_n = 1 \mid p_n)$; in this case the reliability component vanishes. A systematic difference between the two terms is penalized by a positive reliability component. The resolution component rewards variations of the forecast probabilities that are consistent with varying event probabilities. A forecasting scheme that constantly issues the same probabilities has zero resolution. Any meaningful variability of the forecast leads to a positive resolution term which improves the Brier Score. The uncertainty component is equal to the Brier Score of the average (climatological) probability. It thus serves as a benchmark to which the Brier Score of the forecast under consideration can be compared. A 'useful' forecast should have a Brier Score that is lower than its uncertainty component, or in other words, the resolution should be larger than the reliability.

Consider the forecast probability p and the corresponding verification y as two (dependent) random quantities. Then the *calibration function* $\pi(p)$ and the *climatology* $\bar{\pi}$ are defined as

$$\pi(p) = \mathbb{P}(y = 1 \mid p), \text{ and} \tag{3}$$
$$\bar{\pi} = \mathbb{P}(y = 1). \tag{4}$$

Using these definitions, the three components of the

[*] Submitted to *Quarterly Journal of the Royal Meteorological Society*

[†] Corresponding author's address: University of Exeter, Harrison Building, North Park Road, Exeter, EX4 4QF, United Kingdom. Email: s.siegert@exeter.ac.uk

Brier Score decomposition are formally given by

$$\mathrm{REL}^* = \mathbb{E}\left[p - \pi(p)\right]^2, \tag{5}$$

$$\mathrm{RES}^* = \mathbb{E}\left[\pi(p) - \bar{\pi}\right]^2, \text{ and} \tag{6}$$

$$\mathrm{UNC}^* = \bar{\pi}(1 - \bar{\pi}), \tag{7}$$

where $\mathbb{E}$ denotes the mathematical expectation value (Bröcker, 2009). The star (*) is used to differentiate the exact analytical expressions from their empirical estimators, which are discussed below.

In practice, the three components of the Brier Score decomposition must be estimated empirically from the set of forecast probabilities and corresponding verifications $\{p_n, y_n\}$. Such estimators are derived in Murphy (1973); they are presented below in a somewhat different notation, which is suitable for variance estimation by propagation of uncertainty (see Sec. 2).

First of all, the observed forecast probabilities $\{p_n\}$ are binned into D mutually exclusive and collectively exhaustive bins $p^d_{\sqcup}$, where $d = 1, \cdots, D$. Here, bins of equal width which are half-open to the left are used, except the first bin which is closed (but the theory also applies to variable bin widths). As an example, if $D = 3$ one would have $\{p^d_{\sqcup}\}^3_{d=1} = \{[0, 1/3], (1/3, 2/3], (2/3, 1]\}$. Using this binning of the forecast probabilities, the following matrices are defined:

$$A \in \{0,1\}^{N\times D} : A_{nd} = \mathbb{I}(p_n \in p^d_{\sqcup}), \tag{8}$$

$$B \in \{0,1\}^{N\times D} : B_{nd} = \mathbb{I}(p_n \in p^d_{\sqcup})\ y_n, \tag{9}$$

$$C \in [0,1]^{N\times D} : C_{nd} = \mathbb{I}(p_n \in p^d_{\sqcup})\ p_n, \tag{10}$$

$$Y \in \{0,1\}^{N\times 1} : Y_n = y_n, \tag{11}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Summation over a column or row of a matrix is abbreviated by a bullet ($\bullet$), for example

$$A_{\bullet d} = \sum_{n=1}^{N} A_{nd}. \tag{12}$$

A bullet without a second index always refers to the row vector of column sums of a matrix, as in

$$A_{\bullet} = \mathbf{1}^T A \tag{13}$$

where $\mathbf{1}$ is the $N \times 1$ column vector with all elements equal to one.

Using these definitions, $A_{\bullet d}$ is equal to the total number of cases where $p_n \in p^d_{\sqcup}$. $B_{\bullet d}$ is equal to the number of cases where $p_n \in p^d_{\sqcup}$ and at the same time $y_n = 1$. Therefore, a binned estimator for the calibration function is given by

$$\pi(p) \approx \pi_d = \frac{B_{\bullet d}}{A_{\bullet d}}, \tag{14}$$

where $p \in p^d_{\sqcup}$. The climatology is estimated by

$$\bar{\pi} \approx \frac{Y_{\bullet}}{N}. \tag{15}$$

Furthermore, $C_{\bullet d}/A_{\bullet d}$ is equal to the average forecast probability in the d-th bin. $Y_{\bullet}$ is equal to the total number of events that have occurred. Lastly, note that $B_{\bullet\bullet} = Y_{\bullet}$, and $A_{\bullet\bullet} = N$.

The above notation is closely related to the *contingency table* N_{kd}, defined by

$$N_{kd} = \sum_{n=1}^{N} \mathbb{I}(y_n = k)\mathbb{I}(p_n \in p^d_{\sqcup}), \tag{16}$$

which is commonly used to estimate Brier Score decompositions of binned probabilistic forecasts. The matrices A, B, C, and Y were introduced to facilitate variance estimation.

Using the matrices A, B, C, and Y, the estimators for the three components of the Brier Score decomposition originally proposed by Murphy (1973) are given by

$$\begin{aligned}\mathrm{REL} &= \mathrm{REL}(A_{\bullet}, B_{\bullet}, C_{\bullet}) \\ &= \frac{1}{N} \sum_{d\in\mathbb{D}_0} \frac{1}{A_{\bullet d}} \left(B_{\bullet d} - C_{\bullet d}\right)^2,\end{aligned} \tag{17}$$

$$\begin{aligned}\mathrm{RES} &= \mathrm{RES}(A_{\bullet}, B_{\bullet}, Y_{\bullet}) \\ &= \frac{1}{N} \sum_{d\in\mathbb{D}_0} A_{\bullet d} \left(\frac{B_{\bullet d}}{A_{\bullet d}} - \frac{Y_{\bullet}}{N}\right)^2\end{aligned} \tag{18}$$

$$\begin{aligned}\mathrm{UNC} &= \mathrm{UNC}(Y_{\bullet}) \\ &= \frac{Y_{\bullet}(N - Y_{\bullet})}{N^2},\end{aligned} \tag{19}$$

where $\mathbb{D}_0 = \{d : A_{\bullet d} > 0\}$. In the following REL, RES, and UNC are referred to as the *traditional estimators* of the components of Brier Score decomposition.

In Ferro and Fricker (2012) it is shown that the traditional estimators are biased. They show that the bias can be corrected to some extent, although never perfectly eliminated. Using the above notation, the estimators proposed by Ferro and Fricker (2012) are given by

$$\begin{aligned}&\mathrm{REL}'(A_{\bullet}, B_{\bullet}, C_{\bullet}) \\ &= \mathrm{REL} - \frac{1}{N} \sum_{d\in\mathbb{D}_1} \left\{\frac{B_{\bullet d}(A_{\bullet d} - B_{\bullet d})}{A_{\bullet d}(A_{\bullet d} - 1)}\right\},\end{aligned} \tag{20}$$

$$\begin{aligned}&\mathrm{RES}'(A_{\bullet}, B_{\bullet}, Y_{\bullet}) \\ &= \mathrm{RES} - \frac{1}{N} \sum_{d\in\mathbb{D}_1} \left\{\frac{B_{\bullet d}(A_{\bullet d} - B_{\bullet d})}{A_{\bullet d}(A_{\bullet d} - 1)}\right\} \\ &\quad + \frac{Y_{\bullet}(N - Y_{\bullet})}{N^2(N-1)},\end{aligned} \tag{21}$$

and

$$\begin{aligned}\mathrm{UNC}'(Y_\bullet) &= \mathrm{UNC} + \frac{Y_\bullet(N-Y_\bullet)}{N^2(N-1)} \\ &= \frac{Y_\bullet(N-Y_\bullet)}{N(N-1)},\end{aligned} \tag{22}$$

where $\mathbb{D}_1 = \{d : A_{\bullet d} > 1\}$. REL′, RES′, and UNC′ are referred to as the *bias-corrected estimators*.

Due to the analytical expressions Eq. (5) – Eq. (7), it holds that $\mathrm{REL}^* \in [0,1]$, $\mathrm{RES}^* \in [0,1]$ and $\mathrm{UNC}^* \in [0,0.25]$. One could argue that estimators for the individual components should be confined to these intervals as well. While the traditional estimators always satisfy this restriction, the bias-corrected estimators do not. Ferro and Fricker (2012) acknowledge the possibilities $\mathrm{REL}' < 0$ and $\mathrm{RES}' < 0$, and recommend a suitable modification to their bias correction. Unfortunately, this modification does not account for the possibilities $\mathrm{UNC}' > 0.25$ and $\mathrm{RES}' > 1$. In Appendix B a modification of the bias-corrected estimators is suggested which avoids all possible inconsistencies.

Note that binning of continuous forecast probabilities for Brier Score decomposition introduces an estimation error which can lead to the effect that the estimated components do not add up to the average Brier Score calculated by Eq. (1). Two additional components related to within-bin variance and within-bin covariance have been introduced to account for this effect (Stephenson et al., 2008). Furthermore, binning of forecast probabilities is not the only possible method to estimate the calibration function. Bröcker (2008) uses kernel density estimation and chooses the kernel bandwidth by leave-one-out Brier Score minimization. The latter method can also be applied to choose the number of bins in the binning-and-counting approach. In order to maintain a certain level of simplicity, the above refinements of Brier Score decomposition are not considered in the present paper.

A note on terminology: In order to limit confusion due to repeated use of the word *estimate*, we shall always use the term *estimator* to refer to the components of the Brier Score decomposition estimated by Eq. (17) – Eq. (22), and the term *variance estimates* to refer to the approximated variance of these components.

In Sec. 2 of this article it is shown how propagation of uncertainty can be applied to calculate variance estimates for the estimators of a Brier Score decomposition. The variance estimates are validated in an artificial prediction setting in Sec. 3. Application to a meteorological prediction problem in Sec. 4 illustrates a possible use case. In Sec. 5 the simplifying assumptions, validity of the new variance estimates, and variance increase of the bias-corrected estimators are discussed. Section 6 concludes the article. The article is complemented with Supplementary Online Material which includes source code written in the R programming environment (R Core Team, 2012) to reproduce all calculations. A library for the R environment (Siegert, 2013) is available to apply the results of this study in practice.

2 Variance estimation by propagation of uncertainty

The general setting is now that we have scalar estimators F for the components of a Brier Score decomposition, which depend nonlinearly on the column sums $\mathbf{x}$ of a matrix X. For example if $F = \mathrm{REL}$ we have

$$X = [A|B|C] \in \mathbb{R}^{N\times 3D} \tag{23}$$

$$\mathbf{x} = \mathbf{1}^T X = [A_\bullet|B_\bullet|C_\bullet] \in \mathbb{R}^{1\times 3D}. \tag{24}$$

It is possible to apply *propagation of uncertainty* (e. g. Mood et al., 1974) to estimate the variance of $F(\mathbf{x})$ as a function of the covariances of its arguments. The first-order Taylor expansion of F around $\bar{\mathbf{x}}$ (the expectation value of $\mathbf{x}$) is given by

$$F(\mathbf{x}) \approx F(\bar{\mathbf{x}}) + \frac{\partial F(\bar{\mathbf{x}})}{\partial \mathbf{x}}(\mathbf{x}-\bar{\mathbf{x}})^T, \tag{25}$$

where $\partial F(\bar{\mathbf{x}})/\partial \mathbf{x}$ is shorthand for the Jacobian of $F(\mathbf{x})$ evaluated at $\bar{\mathbf{x}}$. Under this approximation, the variance of $F(\mathbf{x})$ is given by

$$\begin{aligned}\mathbb{V}[F(\mathbf{x})] &= \mathbb{E}[F(\mathbf{x}) - \mathbb{E}F(\mathbf{x})]^2 &(26)\\ &= \frac{\partial F(\bar{\mathbf{x}})}{\partial \mathbf{x}}\mathbb{C}\mathrm{ov}(\mathbf{x})\frac{\partial F(\bar{\mathbf{x}})}{\partial \mathbf{x}^T}, &(27)\end{aligned}$$

where $\mathbb{C}\mathrm{ov}(\mathbf{x}) = \mathbb{E}[(\mathbf{x}-\bar{\mathbf{x}})^T(\mathbf{x}-\bar{\mathbf{x}})]$. Recall that the i-th element of $\mathbf{x}$ is the sum over the i-th column of X. Under the assumption that the rows of X are independent and identically distributed, it can be shown that

$$\mathbb{C}\mathrm{ov}(\mathbf{x}) \approx X^T\left(\mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\right)X, \tag{28}$$

where $\mathbf{I}$ denotes the identity matrix. Equation (28) is derived using the fact that $\mathbb{C}\mathrm{ov}(\mathbf{x}_i, \mathbf{x}_j)$ is equal to N times the covariance between the elements of the i-th and j-th column of the matrix X.

Equation (27) combined with Eq. (28) provides a recipe to estimate the variances of the estimators REL, RES, and UNC, as well as their bias-corrected counterparts. All data that is necessary to estimate the variances has already been calculated for the estimators themselves. The only tedious bit is the calculation of the derivatives of the estimators with respect to the individual column sums for the Jacobian. These derivatives are given in Appendix A.

3 Application to artificial data

In order to illustrate their validity, the variance estimates are applied to Brier Score decomposition in an artificial prediction setting, for which the components of the decomposition are known analytically. The results are discussed in Sec. 5. The code to reproduce the numerical computations of this article is available in the Supplementary Online Material.

In the artificial example, assume that the event $y \in \{0, 1\}$ is an independent realization of a Bernoulli trial with success probability q. The value $y = 1$ means that 'the event occurs'. In the present example, the event probability q is itself a random variable that is equally likely to assume one of 6 possible values, namely $q \in \{q_d\}_{d=1}^{6} = \{0.05, 0.15, \cdots, 0.55\}$. A forecasting scheme for the event y which has nonvanishing resolution and nonvanishing reliability is constructed as follows: The forecast probability p corresponds to the actual event probability q whenever $q \neq 0.55$. But whenever the event probability $q = 0.55$, the forecast probability is equal to $p = 1$. That is, $p \in \{p_d\}_{d=1}^{6} = \{q_1, \cdots, q_5, 1\}$, with equal probability of $\frac{1}{6}$.

For the above scheme, the climatological probability is equal to

$$\bar{\pi} = \frac{1}{6}\sum_{d=1}^{6} q_d = \frac{3}{10}. \tag{29}$$

The true uncertainty of this forecasting scheme is thus given by

$$\mathrm{UNC}^* = \bar{\pi}(1-\bar{\pi}) = \frac{21}{100}. \tag{30}$$

Furthermore, since the calibration function in this setting is given by

$$\pi(p_d) = q_d, \tag{31}$$

the true reliability component of the Brier Score of the forecast p is calculated as

$$\mathrm{REL}^* = \frac{1}{6}\sum_{d=1}^{6}(p_d - q_d)^2 = \frac{27}{800}, \tag{32}$$

and the true resolution of the forecast is given by

$$\mathrm{RES}^* = \frac{1}{6}\sum_{d=1}^{6}(q_d - \bar{\pi})^2 = \frac{7}{240}. \tag{33}$$

Note that in this example $\mathrm{REL}^* > \mathrm{RES}^*$, and therefore the forecast is 'useless' in the sense that the constant climatological probability $\bar{\pi}$ achieves a better Brier Score (which is equal to UNC^*) than the forecast probabilities p.

A single numerical experiment consists of $N = 250$ forecast probabilities p_n, and corresponding event indicators y_n, independently sampled as outlined above. Each such experiment results in a data set of forecasts and verifications $\{p_n, y_n\}_{n=1}^{N}$, and a Brier Score decomposition is estimated for this data set. The forecast probabilities assume only 6 discrete values, and therefore no artificial binning has to be applied. For infinitely many forecast instances the estimators converge to the true components, without discrepancies introduced by the binning. The resulting estimators REL, RES, and UNC, as well as their bias-corrected counterparts RES', REL', and UNC' are calculated for this data, together with their corresponding variance estimates derived in Sec. 2. This whole experiment is repeated 100 times, each time with a new realization of forecast probabilities p_n and corresponding event indicators y_n.

The results of these 100 trials are illustrated in Fig. 1. For each trial, the traditional (left) and bias-corrected (right) estimators for reliability, resolution, and uncertainty are shown, augmented with error bars with a half width of two estimated standard deviations.

In Table 1, the outcome of the experiment is further quantified by statistical summary measures. To make the calculation of these summary measures precise, consider as an example the estimator REL. Define REL_i to be the estimator REL obtained on the i-th trial and $\mathbb{V}\mathrm{REL}_i$ the corresponding estimated variance. Furthermore define $\overline{\mathrm{REL}} = \frac{1}{100}\sum_{i=1}^{100}\mathrm{REL}_i$. The sample variance (first column) was calculated by $\frac{1}{100}\sum_{i=1}^{100}(\mathrm{REL}_i - \overline{\mathrm{REL}})^2$, the average estimated variance (second column) was calculated by $\frac{1}{100}\sum_{i=1}^{100}\mathbb{V}\mathrm{REL}_i$, the average squared error (third column) was calculated by $\frac{1}{100}\sum_{i=1}^{100}(\mathrm{REL}_i - \mathrm{REL}^*)^2$, and the average bias (fourth column) was calculated by $\frac{1}{100}\sum_{i=1}^{100}(\mathrm{REL}_i - \mathrm{REL}^*)$. Summary measures for the other components were calculated accordingly.

The artificial example of the present section was used to check convergence rates of the variance estimates. The average absolute difference between sample variance and estimated variance decays proportional to $N^{-3/2}$. This result holds for all components of the Brier Score decomposition. However, it was only obtained empirically, and might not generalize to different prediction settings. A rigorous

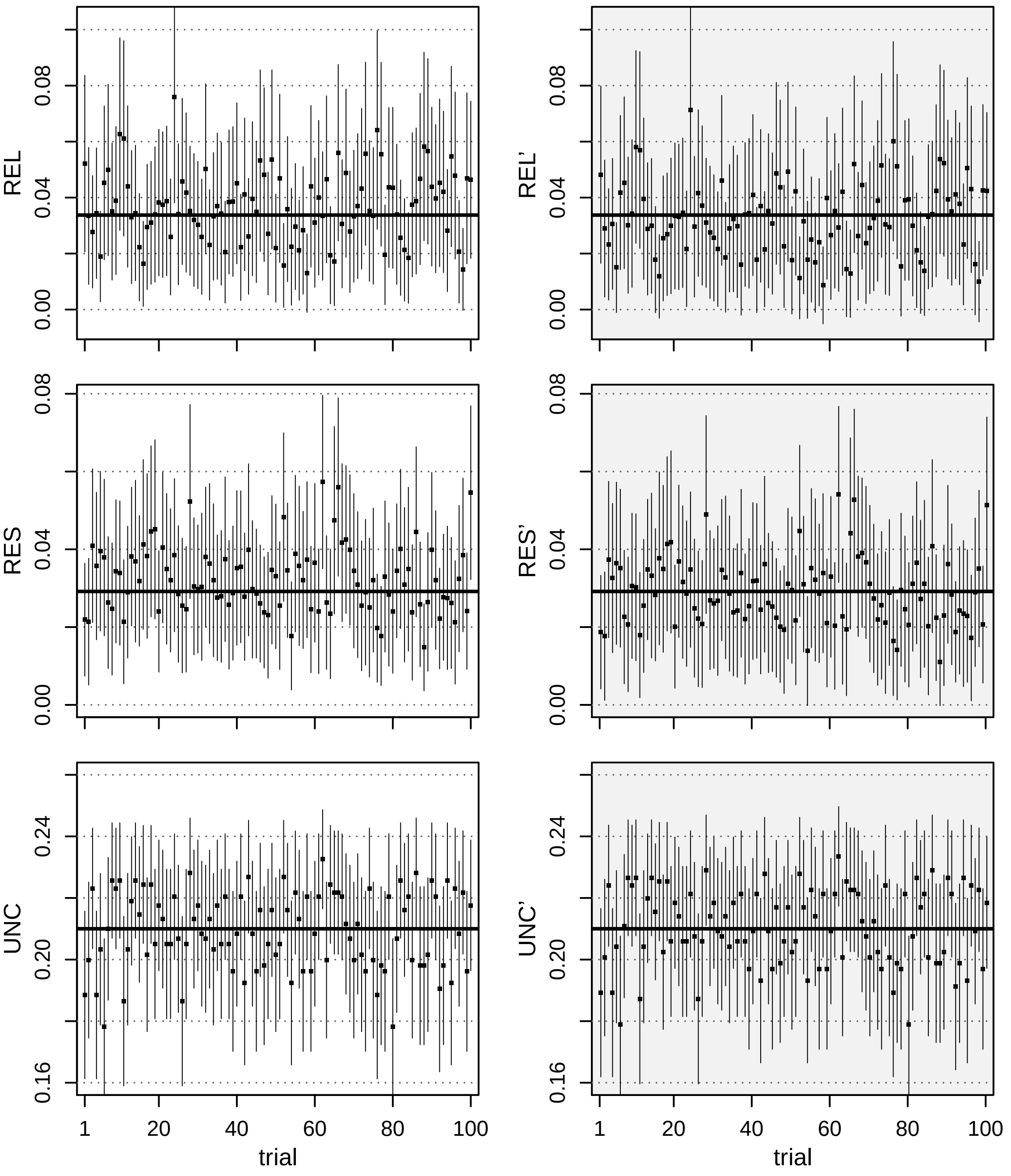

Figure 1: Illustration of the experiment with artificial data. For each trial of the experiment, the traditional and bias-corrected estimators of the Brier Score components are shown, augmented with error bars with a half-width of 2 estimated standard deviations. The bold black line indicates the true value. All abscissae are as in the lower panels.

Table 1: Summary of the artificial example. All averages are taken over the 100 trials of Fig. 1. The first column shows the sample variance of the estimators. The second column shows the average of the estimated variances. The third column shows the average squared difference between the estimator and the true value. The fourth column shows the average bias, that is the average difference between the estimated value and the true value.

	sample variance	avg. est. variance	avg. squared error	avg. bias
REL	1.540×10^{-4}	1.665×10^{-4}	1.641×10^{-4}	3.182×10^{-3}
REL′	1.548×10^{-4}	1.663×10^{-4}	1.563×10^{-4}	-1.184×10^{-3}
RES	7.062×10^{-5}	8.521×10^{-5}	8.093×10^{-5}	32.101×10^{-4}
RES′	7.220×10^{-5}	8.776×10^{-5}	7.230×10^{-5}	-3.155×10^{-4}
UNC	1.561×10^{-4}	1.336×10^{-4}	1.565×10^{-4}	-6.195×10^{-4}
UNC′	1.573×10^{-4}	1.347×10^{-4}	1.574×10^{-4}	2.214×10^{-4}

proof of estimator convergence rates is outside the scope of this paper.

4 Meteorological application

In this section, Brier Score decomposition is applied to real forecast data and the variance estimates are used to quantify the variability of the components of the decomposition. We use daily maximum temperature observations measured at Dresden/Germany (WMO no. 10488) between 1 January 1980 and 31 December 1999 (Deutscher Wetterdienst, 2012). The (binary) prediction target is the exceedance of a certain threshold one day in the future.

The data between 1 January 1980 and 31 December 1989 is used as training data. Denote this data by T'_n, where n is an integer that indicates 'days since 1 January 1970'. We omit the unit of T'_n and remember that it is measured in °C. The *seasonal cycle* c_n is obtained by fitting a second order trigonometric polynomial to the observations:

$$\begin{aligned} c_n = \beta_0 &+ \beta_1 \cos(\omega n) + \beta_2 \sin(\omega n) \\ &+ \beta_3 \cos(2\omega n) + \beta_4 \sin(2\omega n), \end{aligned} \tag{34}$$

where $\omega = 2\pi/(365.2425 \text{ days})$ and the coefficients were fitted by minimizing the sum of squared differences between c_n and T'_n using ordinary linear regression. For the data at hand, we obtain

$$\{\beta_0, \cdots, \beta_4\} = \{13.2, -10.7, -3.1, -0.6, 0.03\}$$

over the training period. Using the seasonal cycle, the *anomalies* T_n are defined by

$$T_n = T'_n - c_n. \tag{35}$$

The anomalies in the training period have zero mean (by construction) and a standard deviation of 4.7. Next, a first-order autoregressive (AR1) model is fitted to the anomalies, using the `R` function `ar` provided by the `stats` package (R Core Team, 2012). That is, the temperature anomaly T_{n+1}, conditional on the anomaly T_n is modeled by

$$T_{n+1} = \alpha T_n + \sigma \epsilon_n, \tag{36}$$

where α is the AR1 parameter which quantifies the serial dependence of successive temperature anomalies, σ^2 is the variance of the residuals, and ϵ_n is a realization of unit variance Gaussian white noise. The values $\alpha = 0.77$ and $\sigma = 2.97$ are obtained in the training data.

The prediction target is whether the temperature anomaly at time n exceeds a threshold τ°C on the next day, that is $y_n = \mathbb{I}(T_n > \tau)$. Using the autoregressive model, a probabilistic 24h exceedance forecast is produced using the formula

$$p_n = \mathbb{P}(T_n > \tau \mid T_{n-1} = t) = 1 - \Phi_{\alpha t, \sigma}(\tau), \tag{37}$$

where $\Phi_{\mu,\sigma}(x)$ is the cumulative Gaussian distribution function with mean μ and variance σ^2, evaluated at x. Using Eq. (37) and the parameters obtained from the training data, daily forecasts are produced for the time between 1 January 1990 and 31 December 1999. The forecast probabilities p_n for the targets y_n are analyzed by decomposition of the Brier Score, with the number of equidistant bins set to $D = 10$.

The result of the analysis is presented in Table 2 for the choice of the threshold $\tau = 5$. Estimators of the three components REL, RES, and UNC, in the traditional and the bias-corrected version are given in the first row. Both, the traditional and bias-corrected estimators add up to REL−RES+UNC = 0.0875. The empirical Brier Score calculated by Eq. (1) is equal to Br = 0.0868. The difference between Br and REL − RES + UNC is due to the binning of the continuous AR1 forecast probabilities, which was applied to estimate the decomposition. In the second row of Table 2, the corresponding variance estimates are shown.

Table 2: Summary of Brier Score decomposition of 10 years' worth of temperature anomaly exceedance forecasts (1 day lead time, threshold 5°C) by an autoregressive model.

	REL	RES	UNC	REL$'$	RES$'$	UNC$'$
estimate	9.060×10^{-4}	0.0542	0.1408	4.130×10^{-4}	0.0537	0.1408
variance	2.096×10^{-7}	1.157×10^{-5}	1.684×10^{-5}	2.056×10^{-7}	1.164×10^{-5}	1.685×10^{-5}

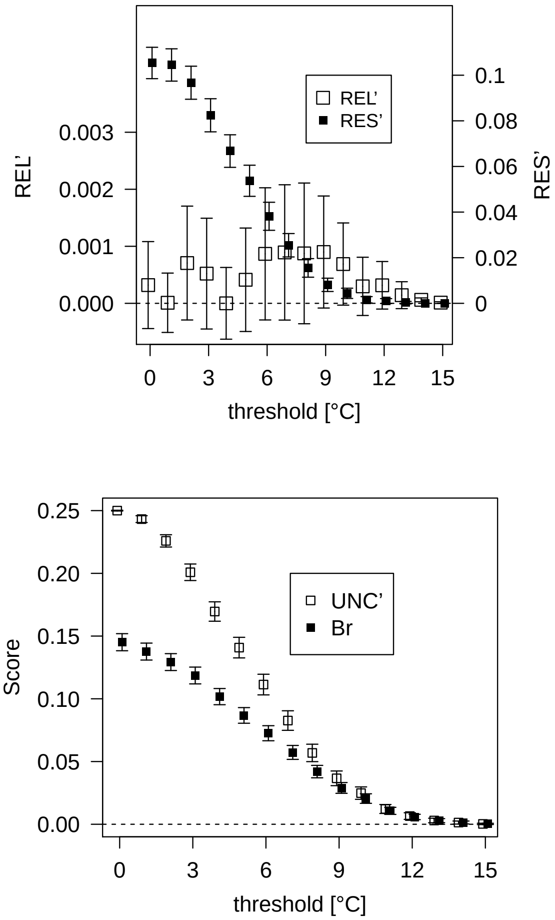

Figure 2: Brier Score decomposition of the temperature anomaly exceedance forecasts by an autoregressive model. Upper panel: REL$'$ and RES$'$ as a function of the threshold which defines the exceedance event, augmented with errorbars of half width two estimated standard deviations. Please note the different y-scales for REL$'$ and RES$'$. Lower panel: Same as above for UNC$'$ and Br. In this plot the y-scale is the same for both quantities.

In Fig. 2, the bias-corrected components of the autoregressive exceedance forecast and the empirical Brier Score are shown as functions of the threshold. The error bars of half widths two standard deviations provide an estimate of the sampling variability of the components. The error bars for the Brier Score were estimated by the standard error of the mean, i.e. $\sqrt{\mathbb{V}(y-p)^2}/\sqrt{N}$ (Bradley et al., 2008).

5 Discussion

The assumptions and simplifications that entered the derivation of the variance estimates must be discussed. The first simplification of the problem was the first order Taylor expansion in Eq. (25). Its validity relies on the assumption, that the difference between the observed values of the arguments and their expectation values is small enough that quadratic terms can be ignored. Furthermore, the covariance matrix that appears in Eq. (27) is estimated by the sample covariance. Both assumptions can lead to errors, especially if N, the number of forecasts and verifications is small. In the light of these criticisms we should expect that more accurate variance estimates than the ones presented here ought to exist. Nonetheless, Fig. 1 suggests that we obtain reasonable variance estimates despite all the simplifying assumptions. A further simplifying assumption is serial independence: To estimate the covariance matrix by Eq. (28), we made use of the assumption that the pairs of forecast probabilities and event indicators $\{p_n, y_n\}$ are independent for different n. This assumption might not hold in meteorological applications because the probability of rain on day $n+1$, for example, is often similar to the probability of rain on day n. The possible effects serial dependence were not considered at all in this paper.

Figure 1 further shows that the two-standard deviations confidence intervals cover the true value between 91 and 97 out of 100 trials. For estimators that are Gaussian and unbiased, the expected value would be close to 95. In the artificial example the biases are about one order of magnitude smaller than the overall variability of the estimators, and the variations of the estimators appear symmetric around their mean and without large deviations.

Unbiasedness and Gaussianity thus seem to be reasonable first approximations to the statistical behavior of the estimators, and the observed coverage frequencies can be taken as further evidence for the quality of the derived variance estimates. However, coverage should not be overinterpreted as long as the exact statistical properties of the data are not known.

Table 1 illustrates the decrease of the biases by the estimators derived by Ferro and Fricker (2012). The magnitude of the average difference between the estimator and the true values is substantially lower for the bias-corrected estimators than for the traditional estimators. At the same time, however, the sample variances of these bias-corrected estimators are slightly larger than the sample variances of the traditional estimators. This is an example of the bias-variance tradeoff, regularly encountered in statistical estimation problems (e. g. Eldar, 2008). In fact, Table 1 shows that the reduction of the bias in the uncertainty, which comes at the cost of an increased variance, leads to a slight increase in the average squared error of this estimator. That is, even though the bias is reduced, the average squared difference between the estimator and the true value has increased. For the other two estimators, this is not the case - the increase in variance does not offset the bias-correction. It should be noted that the increase in sample variance of REL' is not reflected in the estimated variance which decreases very slightly. Whether or not any of these effects is systematic cannot be answered at this point and further investigation is required. The apparent increase in estimator variance is not a generic result, and should certainly not be taken as an argument against the use of bias-corrected estimators.

In Sec. 4 Brier Score decomposition has been applied to autoregressive forecasts of exceedance events of temperature anomalies. The Brier Score decomposition was applied to 10 years' worth of daily data. The two-standard-deviation error bars of all estimators are relatively wide, considering that the decomposition is based on more than 3000 data points. In evaluation studies of weather and climate forecasts, usually much less data is available and the variability of the estimators must be expected to be higher in these cases. Reliable estimates of the variability of the components of the Brier Score decomposition are required for an honest assessment of the significance of the results.

6 Summary and conclusions

The components of the Brier Score decomposition can be used to assess the forecast attributes reliability and resolution, as well as the inherent uncertainty of the underlying process. The decomposition thus provides insight that goes beyond quantifying the performance by calculating the average Brier Score. Variance estimates were derived for the traditional and bias-corrected estimators of the components of Brier Score decomposition. The variances were approximated by propagation of uncertainty. The validity of the variance estimates was illustrated using artificial data, where the true values of the components are known. An actual meteorological forecast setting illustrated a possible application. A discussion was provided about the implied assumptions, as well as the consequences of bias-correction.

In the cases considered, the variance estimates provided meaningful approximations as to the statistical variability of the components of Brier Score decomposition. Confidence intervals exhibited reasonable coverage frequencies, and estimated and empirical variances coincided, despite numerous simplifying assumptions. Furthermore, it was noted that bias-correction can come at the cost of an increased estimator variance. An example was shown where the bias-correction was not able to decrease the average squared difference of the estimator from its true value.

Forecasters who want to compare competing probabilistic forecasting schemes based on finite data will certainly find the competing Brier Score components to be different. Part of this difference is caused by sampling fluctuations. Using the variance estimates proposed here, the magnitude of these fluctuations can be quantified approximately. This allows for a more robust comparison in terms of true predictive skill, and helps to avoid overinterpretations of performance differences.

Acknowledgments

I thank Colm Mulhern for providing helpful comments on an earlier version of this text. I am grateful to Jochen Bröcker, Holger Kantz, and Christopher Ferro for fruitful discussions related to the present article, and I acknowledge expert comments from two reviewers which helped greatly to improve the quality of this article.

A Appendix: Derivatives

Note that some of the following derivatives can be undefined due to vanishing denominators. These derivatives must be set to zero.

A.1 REL

$$\frac{\partial \mathrm{REL}}{\partial A_{\bullet d}} = -\frac{(B_{\bullet d} - C_{\bullet d})^2}{NA_{\bullet d}^2} \tag{38}$$

$$\frac{\partial \mathrm{REL}}{\partial B_{\bullet d}} = \frac{2(B_{\bullet d} - C_{\bullet d})}{NA_{\bullet d}} \tag{39}$$

$$\frac{\partial \mathrm{REL}}{\partial C_{\bullet d}} = -\frac{2(B_{\bullet d} - C_{\bullet d})}{NA_{\bullet d}} \tag{40}$$

A.2 RES

$$\frac{\partial \mathrm{RES}}{\partial A_{\bullet d}} = -\frac{1}{N}\left(\frac{B_{\bullet d}}{A_{\bullet d}} - \frac{Y_\bullet}{N}\right)\left(\frac{B_{\bullet d}}{A_{\bullet d}} + \frac{Y_\bullet}{N}\right) \tag{41}$$

$$\frac{\partial \mathrm{RES}}{\partial B_{\bullet d}} = \frac{2}{N}\left(\frac{B_{\bullet d}}{A_{\bullet d}} - \frac{Y_\bullet}{N}\right) \tag{42}$$

$$\begin{aligned} \frac{\partial \mathrm{RES}}{\partial Y_\bullet} &= -\sum_{d \in \mathbb{D}_0} \frac{2A_{\bullet d}}{N^2}\left(\frac{B_{\bullet d}}{A_{\bullet d}} - \frac{Y_\bullet}{N}\right) \\ &= -\frac{2}{N^2}B_{\bullet\bullet} + \frac{2Y_\bullet}{N^3}A_{\bullet\bullet} = 0 \end{aligned} \tag{43}$$

A.3 UNC

$$\frac{\partial \mathrm{UNC}}{\partial Y_\bullet} = \frac{1}{N} - \frac{2Y_\bullet}{N^2} \tag{44}$$

A.4 REL′

$$\begin{aligned} \frac{\partial \mathrm{REL}'}{\partial A_{\bullet d}} = -\frac{1}{NA_{\bullet d}^2}\Bigg[&(B_{\bullet d} - C_{\bullet d})^2 \\ &+ \frac{B_{\bullet d}^2}{A_{\bullet d} - 1} - \frac{A_{\bullet d}B_{\bullet d}(A_{\bullet d} - B_{\bullet d})}{(A_{\bullet d} - 1)^2}\Bigg] \end{aligned} \tag{45}$$

$$\frac{\partial \mathrm{REL}'}{\partial B_{\bullet d}} = \frac{2B_{\bullet d} - 1}{N(A_{\bullet d} - 1)} - \frac{2C_{\bullet d}}{NA_{\bullet d}} \tag{46}$$

$$\frac{\partial \mathrm{REL}'}{\partial C_{\bullet d}} = -\frac{2(B_{\bullet d} - C_{\bullet d})}{NA_{\bullet d}} \tag{47}$$

A.5 RES′

$$\begin{aligned} \frac{\partial \mathrm{RES}'}{\partial A_{\bullet d}} &= -\frac{1}{N}\left(\frac{B_{\bullet d}}{A_{\bullet d}} - \frac{Y_\bullet}{N}\right)\left(\frac{B_{\bullet d}}{A_{\bullet d}} + \frac{Y_\bullet}{N}\right) \\ &\quad + \frac{B_{\bullet d}}{NA_{\bullet d}^2(A_{\bullet d} - 1)^2}\left[(A_{\bullet d} - B_{\bullet d})^2 - B_{\bullet d}(B_{\bullet d} - 1)\right] \end{aligned} \tag{48}$$

$$\frac{\partial \mathrm{RES}'}{\partial B_{\bullet d}} = \frac{2}{N}\left(\frac{B_{\bullet d}}{A_{\bullet d}} - \frac{Y_\bullet}{N}\right) - \frac{A_{\bullet d} - 2B_{\bullet d}}{NA_{\bullet d}(A_{\bullet d} - 1)} \tag{49}$$

$$\frac{\partial \mathrm{RES}'}{\partial Y_\bullet} = \frac{N - 2Y_\bullet}{N^2(N - 1)} \tag{50}$$

A.6 UNC′

$$\frac{\partial \mathrm{UNC}'}{\partial Y_\bullet} = \frac{N - 2Y_\bullet}{N(N - 1)} \tag{51}$$

B Appendix: Avoiding inconsistencies due to the bias-correction

Let the variables S and T be defined by $\mathrm{REL}' = \mathrm{REL} - S$ (cf. Eq. (20)) and $\mathrm{UNC}' = \mathrm{UNC} + T$ (cf. Eq. (22)). The bias-correction proposed by Ferro and Fricker (2012) can be imagined as shifting the 3-vector $\mathbf{d} = (\mathrm{REL}, \mathrm{RES}, \mathrm{UNC})$ to a new point

$$\mathbf{d}' = (\mathrm{REL}', \mathrm{RES}', \mathrm{UNC}') = \mathbf{d} + \mathbf{c}, \tag{52}$$

where $\mathbf{c} = (-S, -S + T, T)$, and thus the shift takes place along a plane of constant Brier Score. Denote by $\mathcal{A} = [0, 1] \times [0, 1] \times [0, 0.25]$ the space of 'allowed' Brier Score decompositions. In order to avoid inconsistencies due to $\mathbf{d}' \notin \mathcal{A}$, a possible modification is to use the bias-correction

$$\mathbf{d}'' = (\mathrm{REL}'', \mathrm{RES}'', \mathrm{UNC}'') = \mathbf{d} + \gamma\mathbf{c}, \tag{53}$$

where γ is given by

$$\begin{aligned} \gamma = \min\Bigg\{&\frac{\mathrm{REL}}{S}, \max\left[\frac{\mathrm{RES}}{S - T}, \frac{\mathrm{RES} - 1}{S - T}\right], \\ &\frac{1 - 4\mathrm{UNC}}{4T}, 1\Bigg\}. \end{aligned} \tag{54}$$

The parameter γ is confined to the unit interval, and ensures that neither $\mathrm{REL}'' < 0$ nor $\mathrm{RES}'' < 0$ nor $\mathrm{RES}'' > 1$ nor $\mathrm{UNC}'' > 1/4$. Essentially γ ensures that the decomposition $\mathbf{d}$ is shifted linearly as far as possible to the bias-corrected decomposition $\mathbf{d}'$, but not too far as to carrying any of the components out of their allowed range.

References

A.A. Bradley, S.S. Schwartz, and T. Hashino. Sampling uncertainty and confidence intervals for the Brier score and Brier skill score. *Weather and Forecasting*, 23(5):992–1006, 2008. doi: 10.1175/2007WAF2007049.1.

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.

M.H. DeGroot and S.E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):12–22, 1983.

Deutscher Wetterdienst. Climatological Database, online `www.dwd.de` [accessed on 5 November 2012], 2012.

Y.C. Eldar. *Rethinking biased estimation*, volume 1. Now Publishers Inc, 2008.

C.A.T. Ferro and T.E. Fricker. A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society*, 138(668):1954–1960, 2012. doi: 10.1002/qj.1924.

A.M. Mood, F.A. Graybill, and D.C. Boes. *Introduction to the theory of statistics 3rd ed.* McGraw–Hill, New York, 1974.

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.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

S. Siegert. bride: Brier Score decomposition for probabilistic forecasts of binary events. R-package (version 1.3), 2013. URL `http://CRAN.R-project.org/package=bride`.

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