Decomposing proper scores into conditional uncertainty, resolution, and reliability terms

Sam Allen
University of Bern
Oeschger Centre for Climate Change Research

Abstract

Probabilistic forecasts are often assessed and compared using proper scoring rules. While proper scoring rules return a single measure of forecast accuracy, it is well-known that the expected score can be decomposed into terms that quantify the uncertainty, resolution, and reliability of the forecasts. Allen *et al.* (2023) extend this by introducing decompositions of proper scores into uncertainty, resolution, and reliability components conditional on certain events, or states, having occurred. This vignette reproduces the results presented therein, and demonstrates how the conditional score decompositions can be implemented in practice for the Brier score.

Keywords: probabilistic forecast evaluation, proper scoring rules, score decompositions, R.

1. Introduction

Scoring rules are functions that take a probabilistic forecast F and an observation y as inputs, and output a real (possibly non-finite) value, or score. The score quantifies the distance between the forecast and the observation, and a lower score is therefore preferable.

A scoring rule S is said to be proper with respect to a class of probability distributions $\mathcal{F}$ if

$$\mathsf{E}_{Y\sim G}S(G,Y) \leq \mathsf{E}_{Y\sim G}S(F,Y), \tag{1}$$

for all $F, G \in \mathcal{F}$, where E_Y denotes the expectation with respect to the random variable Y. S is strictly proper with respect to $\mathcal{F}$ if the above inequality is strict.

For concision, we write the expected score for the forecast F as

$$\bar{S}(F,G) := \mathsf{E}_{Y\sim G}S(F,Y). \tag{2}$$

The score entropy of a distribution $G \in \mathcal{F}$ is then defined as

$$e(G) := \bar{S}(G,G), \tag{3}$$

and the corresponding score divergence between two distributions $F, G \in \mathcal{F}$ as

$$d(F,G) := \bar{S}(F,G) - \bar{S}(G,G). \tag{4}$$

By definition, the score divergence is a non-negative function if it is generated from a scoring rule that is proper.

While proper scoring rules return a single measure of the distance between the forecast and the observation, it is well-known that the expected score can be decomposed into terms that quantify the uncertainty, resolution, and reliability of the forecasts. In particular, we have that

$$\mathsf{E}_{F,Y} S(F,Y) = \underbrace{e(G)}_{=\mathrm{UNC}_Y} - \underbrace{\mathsf{E}_F d(G, G_F)}_{=\mathrm{RES}_F} + \underbrace{\mathsf{E}_F d(F, G_F)}_{=\mathrm{REL}_F} \tag{5}$$

where G denotes the unconditional distribution of Y, and G_F the conditional distribution of Y given the forecast F (Bröcker 2009). Note that both the outcome Y and the forecast F are treated as random variables in this case.

The first term of the decomposition, UNC_Y, expresses the inherent variability of the predictand, and is therefore referred to as the forecast *uncertainty*. The second component, RES_F, the forecast *resolution*, is the expected divergence between the unconditional and conditional distributions of Y, thereby measuring the discrimination ability of the forecasts. Finally, the *reliability* component, REL_F, is the expected divergence between F and G_F; a forecast is said to be reliable (or auto-calibrated) if $F = G_F$ almost surely, leading to $\mathrm{REL}_F = 0$, and the reliability term can therefore be interpreted as the extent to which the forecasts are miscalibrated, or unreliable. Since RES_F and REL_F are defined as expectations of score divergence, these terms will be non-negative whenever a proper scoring rule is used to evaluate the forecasts.

However, it could be the case that a reliable forecast is miscalibrated conditional on certain events having occurred; the miscalibrations could cancel each other out, leading to the REL_F term being zero despite these conditional errors. To address this, Allen *et al.* (2023) introduce a conditional decomposition of proper scoring rules that assesses forecast performance conditionally on some auxiliary variable A. For example, A could represent the occurrence of a particular season or weather regime, or could denote that forecast performance is being assessed at a particular location. The decomposition of the expected score into forecast uncertainty, resolution and reliability terms can easily be applied conditionally on A:

$$\mathsf{E}_{F,Y|A} S(F,Y) = e(G_A) - \mathsf{E}_{F|A} d(G_A, G_{F,A}) + \mathsf{E}_{F|A} d(F, G_{F,A}) \tag{6}$$

where G_A denotes the conditional distribution of Y given A, and $G_{F,A}$ is the conditional distribution of Y given both the forecast F and the auxiliary variable A. By the tower law of conditional expectations, the total expected score is recovered by taking the expectation with respect to A:

$$\mathsf{E}_{F,Y} S(F,Y) = \underbrace{\mathsf{E}_A e(G_A)}_{=\mathrm{UNC}_{Y|A}} - \underbrace{\mathsf{E}_A \mathsf{E}_{F|A} d(G_A, G_{F,A})}_{=\mathrm{RES}_{F|A}} + \underbrace{\mathsf{E}_A \mathsf{E}_{F|A} d(F, G_{F,A})}_{=\mathrm{REL}_{F|A}}. \tag{7}$$

The terms $\mathrm{UNC}_{Y|A}$, $\mathrm{RES}_{F|A}$, and $\mathrm{REL}_{F|A}$ represent the expected uncertainty, resolution, and reliability given the auxiliary variable A.

This constitutes a second, distinct decomposition of the expected score into uncertainty, resolution, and reliability components, where the terms depend on A. Although these terms are not equivalent to those in Equation 5, Allen *et al.* (2023) demonstrate that an alternative representation of the expected score exists that amalgamates the two decompositions, thereby possessing the benefits of both:

$$\mathsf{E}_{F,Y} S(F,Y) = \{\mathrm{UNC}_{Y|A} + \mathrm{RES}_A\} - \{\mathrm{RES}_A + \mathrm{RES}_{F|A} - \mathrm{RES}_{A|F}\} + \{\mathrm{REL}_{F|A} - \mathrm{RES}_{A|F}\}. \tag{8}$$

The first, second and third sets of curly braces are equivalent, respectively, to the uncertainty, resolution and reliability components in Equation 5.

This extended decomposition contains two terms, RES_A and $\text{RES}_{A|F}$, that are both added and subtracted from the decomposition, and therefore have no effect on the overall score:

$$\begin{aligned} \text{RES}_A &:= \mathsf{E}_A d(G, G_A), \\ \text{RES}_{A|F} &:= \mathsf{E}_{F,A} d(G_F, G_{F,A}). \end{aligned} \tag{9}$$

Although these extra terms contribute nothing to the expected score, they are themselves useful, in that they convey supplementary information to the forecaster regarding the behaviour of the forecasts and observations conditional on the chosen states: RES_A is the expected divergence between the unconditional distribution of Y and the conditional distribution of Y given A, and it therefore measures the amount of information contained in the variable A; similarly, $\text{RES}_{A|F}$ is the expected divergence between the conditional distribution of Y given F and the conditional distribution of Y given both F and A, thereby quantifying the additional information contained in the variable A that is not already present in the forecast.

Having re-introduced these decompositions, the goal of this vignette is to demonstrate how the accompanying R package allows users to implement the decompositions in practical applications. We then discuss some potential extensions to the existing functionality that should be explored in the future.

2. Brier score

Score decompositions have been studied in most detail using the Brier score. Hence, the package currently only contains functionality to decompose the Brier score into conditional uncertainty, resolution, and reliability terms. Nonetheless, it should be straightforward to apply the conditional decomposition to other scoring rules for binary outcomes, such as the logarithmic score, which could be added to the package in the future.

The Brier score is defined as the squared difference between a probability forecast $p \in [0, 1]$ and the corresponding binary outcome $y \in \{0, 1\}$:

$$\text{BS}(p, y) = (p - y)^2. \tag{10}$$

Given a sample of n forecasts $p_1, \ldots, p_n$ and corresponding observations $y_1, \ldots, y_n$, forecasters are typically assessed using their average score

$$\frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2. \tag{11}$$

Of course, this is an unbiased estimator for the expected score $\mathsf{E}_{p,Y}\text{BS}(p, Y)$.

2.1. Sample estimators for the decomposition terms

For simplicity, assume that the forecast p can only take one of a finite number of values, $p \in \{P_1, ..., P_K\}$. Murphy (1973) showed that in such circumstances, the average Brier score over n forecast instances can be divided into three components:

$$\frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2 = \widehat{\text{UNC}}_Y - \widehat{\text{RES}}_F + \widehat{\text{REL}}_F, \tag{12}$$

where

$$\begin{aligned}\widehat{\mathrm{UNC}}_Y &= \bar{y}(1-\bar{y}),\\ \widehat{\mathrm{RES}}_F &= \sum_{k=1}^{K} \frac{n_{k\bullet}}{n}(\bar{y}_{k\bullet}-\bar{y})^2,\\ \widehat{\mathrm{REL}}_F &= \sum_{k=1}^{K} \frac{n_{k\bullet}}{n}(P_k-\bar{y}_{k\bullet})^2.\end{aligned} \tag{13}$$

Here, we define $\bar{y} = \sum_{i=1}^{n} y_i/n$, and $\bar{y}_{k\bullet} = \sum_{I_{k\bullet}} y_i/n_{k\bullet}$, with $I_{k\bullet} = \{i : p_i = P_k\}$ the set of all instances where P_k was issued as the forecast, and $n_{k\bullet} = |I_{k\bullet}|$ the number of such instances.

The terms $\widehat{\mathrm{UNC}}_Y$, $\widehat{\mathrm{RES}}_F$, and $\widehat{\mathrm{REL}}_F$ provide sample estimators for the forecast uncertainty, resolution, and reliability. However, although the mean score is an unbiased estimator for the expected score, $\widehat{\mathrm{UNC}}_Y$, $\widehat{\mathrm{RES}}_F$, and $\widehat{\mathrm{REL}}_F$ are generally biased estimators for the corresponding population terms UNC_Y, RES_F, and REL_F (Bröcker 2012).

To address this, Ferro and Fricker (2012) introduced bias corrections to the above sample estimators:

$$\begin{aligned}\widetilde{\mathrm{UNC}}_Y &= \widehat{\mathrm{UNC}}_Y + \frac{\bar{y}(1-\bar{y})}{n-1},\\ \widetilde{\mathrm{RES}}_F &= \widehat{\mathrm{RES}}_F + \frac{\bar{y}(1-\bar{y})}{n-1} - \frac{1}{n}\sum_{k=1}^{K} \frac{n_{k\bullet}}{n_{k\bullet}-1}\bar{y}_{k\bullet}(1-\bar{y}_{k\bullet}),\\ \widetilde{\mathrm{REL}}_F &= \widehat{\mathrm{REL}}_F - \frac{1}{n}\sum_{k=1}^{K} \frac{n_{k\bullet}}{n_{k\bullet}-1}\bar{y}_{k\bullet}(1-\bar{y}_{k\bullet}),\end{aligned} \tag{14}$$

where it is assumed that $n_{k\bullet} \neq 1$ for all $k = 1, \ldots, K$. The $\widetilde{\mathrm{UNC}}_Y$ term is an unbiased estimator for UNC_Y, whereas $\widetilde{\mathrm{REL}}_F$ and $\widetilde{\mathrm{RES}}_F$ remain biased, but with a bias that decays at a much faster rate than that of $\widehat{\mathrm{REL}}_F$ and $\widehat{\mathrm{RES}}_F$. Note, however, that these bias corrections result in estimators for RES_F and REL_F that may be negative.

Finally, Dimitriadis *et al.* (2021) proposed an approach to estimate the score decomposition terms using isotonic regression:

$$\begin{aligned}\mathrm{UNC}^*_Y &= \frac{1}{n}\sum_{i=1}^{n}(\bar{y}-y_i)^2,\\ \mathrm{RES}^*_F &= \frac{1}{n}\sum_{i=1}^{n}(\bar{y}-y_i)^2 - \frac{1}{n}\sum_{i=1}^{n}(p^*_i-y_i)^2,\\ \mathrm{REL}^*_F &= \frac{1}{n}\sum_{i=1}^{n}(p_i-y_i)^2 - \frac{1}{n}\sum_{i=1}^{n}(p^*_i-y_i)^2,\end{aligned} \tag{15}$$

where $p^*_1, \ldots, p^*_n$ are probability forecasts obtained by applying an isotonic regression-based re-calibration scheme to $p_1, \ldots, p_n$. These forecasts are reliable by construction (in sample), and contain the same information as the original forecasts; the reliability of the original forecasts is thus estimated by considering the score difference between these forecasts and the re-calibrated forecasts.

It is clear that the average Brier score is recovered by combining these estimators, and this approach can readily be implemented with any other scoring rule for binary outcomes. In contrast to the two approaches described beforehand, this does not require the assumption that the forecast can only take one of a finite number of values, and is guaranteed to yield non-negative estimators for all components. We therefore generally advocate the use of this approach in practice.

2.2. Sample estimators for the conditional decomposition terms

Similar sample estimators can also be obtained for the conditional decomposition terms of Equation 8. In this case, we assume that we have access to the values $a_1, \dots, a_n$ of the auxiliary variable corresponding to the forecasts and observations introduced above, and we additionally assume that there are only a finite number of possible values, i.e. $a_i \in \{A_1, ..., A_J\}$ for $i = 1, \dots, n$. In the following, we refer to these possible values of A as *states*.

As before, we let $I_{\bullet j} = \{i : a_i = A_j\}$ denote the set of all instances where state A_j occurred, and let $I_{kj} = \{i : p_i = P_k, a_i = A_j\}$ denote the set of instances where P_k was issued as the forecast and state A_j occurred, with $n_{\bullet j} = |I_{\bullet j}|$ and $n_{kj} = |I_{kj}|$ the number of instances in these sets. We then define $\bar{y}_{\bullet j} = \sum_{I_{\bullet j}} y_i / n_{\bullet j}$ and $\bar{y}_{kj} = \sum_{I_{kj}} y_i / n_{kj}$.

Using this, the average Brier score can be decomposed into

$$\frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2 = \left\{\widehat{\mathrm{UNC}}_{Y|A} + \widehat{\mathrm{RES}}_A\right\} - \left\{\widehat{\mathrm{RES}}_A + \widehat{\mathrm{RES}}_{F|A} - \widehat{\mathrm{RES}}_{A|F}\right\} + \left\{\widehat{\mathrm{REL}}_{F|A} - \widehat{\mathrm{RES}}_{A|F}\right\}, \tag{16}$$

where

$$\begin{aligned}
\widehat{\mathrm{UNC}}_{Y|A} &= \sum_{j=1}^{J} \frac{n_{\bullet j}}{n} \bar{y}_{\bullet j}(1 - \bar{y}_{\bullet j}), \\
\widehat{\mathrm{RES}}_A &= \sum_{j=1}^{J} \frac{n_{\bullet j}}{n} (\bar{y}_{\bullet j} - \bar{y})^2, \\
\widehat{\mathrm{RES}}_{F|A} &= \sum_{j,k=1}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{\bullet j} - \bar{y}_{kj})^2, \\
\widehat{\mathrm{RES}}_{A|F} &= \sum_{j,k=1}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{k\bullet} - \bar{y}_{kj})^2, \\
\widehat{\mathrm{REL}}_{F|A} &= \sum_{j,k=1}^{J,K} \frac{n_{kj}}{n} (P_k - \bar{y}_{kj})^2.
\end{aligned} \tag{17}$$

Bias corrections to these estimators can be derived in a similar vein to before:

$$
\begin{aligned}
\widetilde{\mathrm{UNC}}_{Y|A} &= \widehat{\mathrm{UNC}}_{Y|A} + \frac{1}{n}\sum_{j=1}^{J} \frac{n_{\bullet j}}{n_{\bullet j}-1}\bar{y}_{\bullet j}(1-\bar{y}_{\bullet j}), \\
\widetilde{\mathrm{RES}}_{A} &= \widehat{\mathrm{RES}}_{A} + \frac{\bar{y}(1-\bar{y})}{n-1} - \frac{1}{n}\sum_{j=1}^{J} \frac{n_{\bullet j}}{n_{\bullet j}-1}\bar{y}_{\bullet j}(1-\bar{y}_{\bullet j}), \\
\widetilde{\mathrm{RES}}_{F|A} &= \widehat{\mathrm{RES}}_{F|A} + \frac{1}{n}\sum_{j=1}^{J} \frac{n_{\bullet j}}{n_{\bullet j}-1}\bar{y}_{\bullet j}(1-\bar{y}_{\bullet j}) - \frac{1}{n}\sum_{j,k=1}^{J,K} \frac{n_{kj}}{n_{kj}-1}\bar{y}_{kj}(1-\bar{y}_{kj}), \\
\widetilde{\mathrm{RES}}_{A|F} &= \widehat{\mathrm{RES}}_{A|F} - \frac{1}{n}\sum_{j,k=1}^{J,K} \frac{n_{kj}}{n_{kj}-1}\bar{y}_{kj}(1-\bar{y}_{kj}) + \frac{1}{n}\sum_{k=1}^{K} \frac{n_{k\bullet}}{n_{k\bullet}-1}\bar{y}_{k\bullet}(1-\bar{y}_{k\bullet}), \\
\widetilde{\mathrm{REL}}_{F|A} &= \widehat{\mathrm{REL}}_{F|A} - \frac{1}{n}\sum_{j,k=1}^{J,K} \frac{n_{kj}}{n_{kj}-1}\bar{y}_{kj}(1-\bar{y}_{kj}),
\end{aligned}
\tag{18}
$$

where it is similarly assumed that $n_{k\bullet}, n_{\bullet j}, n_{kj} \neq 1$ for all $k = 1, \ldots, K$, $j = 1, \ldots, J$. As for the bias corrections for the classical decomposition terms, the biases of these terms are not zero, but they decay to zero at a faster rate than the biases of the uncorrected estimators.

Finally, the isotonic regression-based approach proposed by Dimitriadis *et al.* (2021) can be leveraged to construct estimators for the conditional decomposition terms. This can be achieved by extending the general framework to construct score decompositions discussed by Siegert (2017), among others. If $r_1, \ldots, r_n$ represent uninformative reference forecasts, and $q_1, \ldots, q_n$ represent re-calibrated versions of the original forecasts $p_1, \ldots, p_n$, then the average score for these original forecasts can be written as

$$
\left\{\frac{1}{n}\sum_{i=1}^{n} S(r_i, y_i)\right\} - \left\{\frac{1}{n}\sum_{i=1}^{n} S(r_i, y_i) - \frac{1}{n}\sum_{i=1}^{n} S(q_i, y_i)\right\} + \left\{\frac{1}{n}\sum_{i=1}^{n} S(p_i, y_i) - \frac{1}{n}\sum_{i=1}^{n} S(q_i, y_i)\right\}. \tag{19}
$$

The terms in square brackets represent estimators for the uncertainty, resolution, and reliability components of the expected score, respectively. The canonical choice for the reference forecasts is the unconditional observed mean, i.e. $r_i = \bar{y}$ for $i = 1, \ldots, n$, while Dimitriadis *et al.* (2021) argue that isotonic regression provides a canonical re-calibration scheme when the outcomes are binary. This leads to the estimators presented in Equation 15.

A similar, general decomposition can be constructed in the conditional case. In particular, let $r_1^A, \ldots, r_n^A$ represent forecasts that are uninformative given the auxiliary information A, and let $q_1^A, \ldots, q_n^A$ be conditionally re-calibrated versions of the original forecasts. Then, we

can estimate the terms of the conditional decomposition using

$$\begin{aligned}
\mathrm{UNC}^*_{Y|A} &= \frac{1}{n}\sum_{i=1}^{n} S(r_i^A, y_i), \\
\mathrm{RES}^*_{A} &= \frac{1}{n}\sum_{i=1}^{n} S(r_i, y_i) - \frac{1}{n}\sum_{i=1}^{n} S(r_i^A, y_i), \\
\mathrm{RES}^*_{F|A} &= \frac{1}{n}\sum_{i=1}^{n} S(r_i^A, y_i) - \frac{1}{n}\sum_{i=1}^{n} S(q_i^A, y_i), \\
\mathrm{RES}^*_{A|F} &= \frac{1}{n}\sum_{i=1}^{n} S(q_i, y_i) - \frac{1}{n}\sum_{i=1}^{n} S(q_i^A, y_i), \\
\mathrm{REL}^*_{F|A} &= \frac{1}{n}\sum_{i=1}^{n} S(p_i, y_i) - \frac{1}{n}\sum_{i=1}^{n} S(q_i^A, y_i).
\end{aligned} \tag{20}$$

As before, the conditional reference forecasts could be the conditional means corresponding to each event $\bar{y}_{\bullet j}$, while the conditionally re-calibrated forecasts could be obtained by fitting separate re-calibration schemes to the forecasts in each state. In the binary case, for example, this could be achieved by fitting separate isotonic regression models to the forecasts in each state, or by using A as a predictor in the model; this latter approach would additionally permit the decomposition to be constructed when A represents some continuous information, rather than a discrete event. Of course, estimators for the conditional Brier score decomposition terms are obtained by choosing S to be the Brier score.

3. Application

3.1. Data

In this section, we demonstrate how the accompanying package allows these conditional decompositions to be applied in practice. We do so by reproducing the results presented in Allen *et al.* (2023), wherein the Brier score is used to evaluate forecast probabilities that the daily maximum temperature will exceed a chosen threshold. Forecasts are available from MeteoSwiss's COSMO-E ensemble prediction system at 146 synoptic weather stations across Switzerland, for the three summer months (JJA) between 2018 and 2020. The forecasts considered here have been issued three days in advance. The COSMO-E prediction system generates ensemble forecasts comprised of $M = 21$ members for the daily maximum temperature, and a probability that the temperature will exceed a chosen threshold is then extracted from the ensemble by considering the proportion of ensemble members that exceed the threshold. This results in 22 (i.e. $M + 1$) evenly spaced possible forecast values.

The COSMO-E forecasts are evaluated using the mean Brier score over all forecast cases, with daily maximum temperature measurements at the sites of interest used to verify the forecasts. The classical decomposition of the mean Brier score is used to assess the resolution and reliability of the forecasts, while the conditional decomposition is employed to calculate these terms conditionally on a set of states. For concision, we restrict attention here to one set of states, corresponding to a grouping of the stations based on their altitude. That is, each forecast-observation pair is assigned to a state depending on the altitude of the station

at which the forecast was issued: five groups are considered (i.e. $J = 5$), which correspond to whether the station altitude is lower than 500m, between 500m and 1000m, 1000m and 1500m, 1500m and 2000m, or above 2000m.

The COSMO-E ensemble prediction system is compared to four alternative forecast strategies: a climatological forecast, which is constant and equal to the relative frequency that the (out-of-sample) observations exceed the threshold of interest; a conditional climatological forecast that calculates this relative frequency for each state separately, and then issues the probability corresponding to the prevailing state as the forecast; a post-processing approach that uses logistic regression to re-calibrate the COSMO-E forecast; and an altitude-dependent post-processing model that fits a separate logistic regression model corresponding to each of the five states, and re-calibrates the COSMO-E forecasts depending on the prevailing state.

Note that the exact data used in Allen *et al.* (2023) is owned by MeteoSwiss and therefore not publicly available. To circumvent this, random noise has been added to the data. The results that we obtain here are therefore not identical to those presented in Allen *et al.* (2023), though the implementation of the conditional Brier score decompositions is the same. This noisy data is available in this package, and can be accessed using

```
R> data("noisy_data", package = "ConditionalScoreDecomp")
```

The result is a data frame containing the observation `Obs`, equal to zero or one depending on whether the observed temperature exceeds the threshold of interest; forecast probabilities for the COSMO-E (`Ens`), climatological (`Clim`), conditional climatological (`CondClim`), post-processed (`PP`), and conditional post-processed (`CondPP`) forecast strategies; the corresponding state `state`, with 1 denoting that the forecast was issued at a station with an altitude smaller than 500m, and 5 corresponding to a station with altitude greater than 2000m; and the temperature threshold `th`, ranging from 5 to 30, for which the observation and forecast probabilities were calculated.

3.2. Results

Firstly, consider a threshold of 20 degrees Celsius.

```
R> dat20 <- subset(noisy_data, th == 20)
```

We can calculate the classical Brier score decomposition using the `bs_decomp()` function

```
bs_decomp(o, p, bins = NULL, method = "isotonic")
```

This function takes a vector of binary outcomes `o` and a vector of corresponding probability forecasts `p` and outputs a vector containing the forecast uncertainty, resolution, and reliability, as well as the total Brier score. The `bins` argument is an integer that specifies how many bins the forecasts should be grouped into; if `bins = NULL` (the default), then it is assumed that `p` has already been binned. The decomposition can be performed using the three different estimators discussed in the previous section: `method = "classical"` returns the uncorrected decomposition introduced by Murphy (1973), `method = "bias-corrected"` returns Ferro and Fricker (2012)'s bias corrected terms, while `method = "isotonic"` employs isotonic regression to estimate the components, as in Dimitriadis *et al.* (2021). Any other

choice for `method` returns an error, and this is also the case if the observations are not binary, or if a probability forecast is specified that is not between zero and one.

We can apply this function to the probabilities issued by each of the five forecast strategies.

```
R> o <- dat20$Obs
R> bs_mat <- rbind(Clim = bs_decomp(o = o, p = dat20$Clim),
+                  CondClim = bs_decomp(o = o, p = dat20$CondClim),
+                  Ens = bs_decomp(o = o, p = dat20$Ens),
+                  PP = bs_decomp(o = o, p = dat20$PP),
+                  CondPP = bs_decomp(o = o, p = dat20$CondPP))
R> print(bs_mat*1e4)

          UNC  RES   REL  TOT
Clim     2297    0  11.4 2308
CondClim 2297  886  18.6 1429
Ens      2297 1092 142.7 1348
PP       2297 1066  52.9 1284
CondPP   2297 1195  86.9 1189
```

Note that these decompositions have been calculated using the isotonic regression-based terms (the default in `bs_decomp()`), which is in contrast to Allen *et al.* (2023), where the bias corrected estimators were employed.

We can similarly calculate the conditional decomposition terms using the `bs_decomp_cond()` function

```
bs_decomp_cond(o, p, states, bins = NULL, method = "isotonic")
```

This differs from `bs_decomp()` only in that it includes an additional argument `states`, which contains a vector of the states on which to condition the decomposition.

Applying this to the five forecast strategies considered here gives

```
R> state <- dat20$state
R> bs_mat_c <- rbind(Clim = bs_decomp_cond(o, dat20$Clim, state),
+                    CondClim = bs_decomp_cond(o, dat20$CondClim, state),
+                    Ens = bs_decomp_cond(o, dat20$Ens, state),
+                    PP = bs_decomp_cond(o, dat20$PP, state),
+                    CondPP = bs_decomp_cond(o, dat20$CondPP, state))
R> print(bs_mat_c*1e4)

         UNC_A RES_A  RES_F|A  RES_A|F REL_F|A  TOT
Clim      1411   886 3.81e-14 8.86e+02   897.6 2308
CondClim  1411   886 3.81e-14 4.34e-14    18.6 1429
Ens       1411   886 3.80e+02 1.75e+02   317.6 1348
PP        1411   886 3.67e+02 1.87e+02   240.1 1284
CondPP    1411   886 3.73e+02 6.48e+01   151.6 1189
```

These terms mirror the results presented in Figure 4 of Allen *et al.* (2023).

This information can be visualised more succinctly using the `plot_decomp()` function.

```
R> plot_decomp(terms_un = bs_mat["Ens", ]*1e4,
+               terms_cnd = bs_mat_c["Ens", ]*1e4, title = "COSMO-E")
```

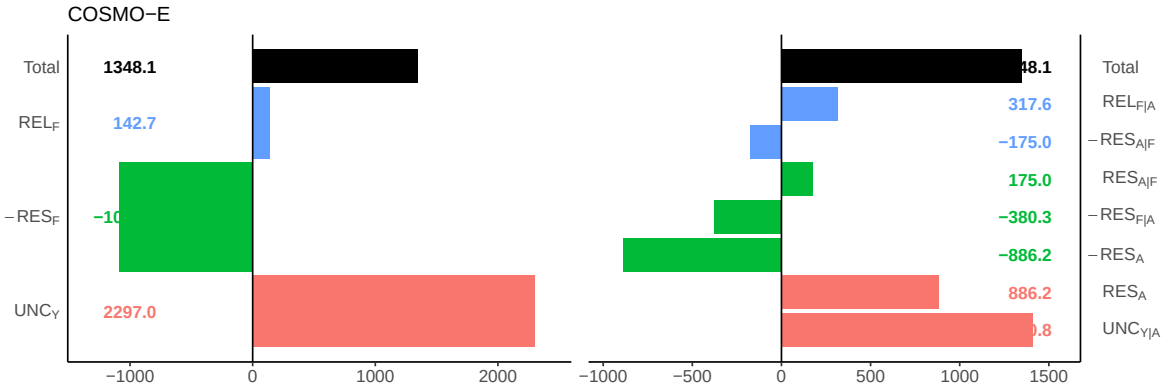

Figure 1: Bar plot containing the unconditional and conditional decomposition terms for the COSMO-E forecasts.

```
plot_decomp(terms_un = NULL, terms_cnd = NULL,
            title = "", waterfall = FALSE, dec_places = 1)
```

This function takes a vector `terms_un` containing the unconditional decomposition terms, and a vector `terms_cnd` containing the conditional decomposition terms, and plots the results either using a bar plot (the default, as in Figures 2-4 of Allen *et al.* (2023)) or a waterfall plot (`waterfall = TRUE`). A custom title to the plot can be chosen using `title`, while the `dec_places` argument allows the user to specify how many decimal places should be displayed.

If one of `terms_un` and `terms_cnd` is not specified, then the function returns a plot containing only the decomposition terms that have been given. Otherwise, the two decompositions are displayed alongside each other. Of course, an error is returned if neither argument is specified.

As an example, consider the COSMO-E ensemble forecasts at a threshold of 20 degrees. Figure 1 displays a bar plot containing both the unconditional and conditional decomposition terms for the COSMO-E forecasts, while Figure 2 displays the same information in a waterfall plot. Of course, it is straightforward to obtain analogous plots corresponding to the other forecasting methods.

The results presented thus far have been specific for forecasts issued for whether the daily maximum temperature will exceed 20 degrees celcius. It is straightforward to calculate the decomposition terms corresponding to other thresholds using the same approach. Figure 3 displays the unconditional and conditional decomposition terms as a function of the daily maximum temperature threshold, as in Figure 6 of Allen *et al.* (2023). Again, we illustrate this here using the COSMO-E forecasts.

Finally, as discussed in Allen *et al.* (2023), an optimal re-calibration scheme should remove the miscalibration in the COSMO-E output, without sacrificing the information contained in these forecasts. The relative improvement gained by such a re-calibration scheme is therefore equal to the reliability component, RES_F, of the COSMO-E forecasts, divided by the total score. To assess the efficiency of the post-processing method, we can compare this theoretical improvement to the actual improvement gained by post-processing. Figure 4 shows

```
R> plot_decomp(terms_un = bs_mat["Ens", ]*1e4,
+               terms_cnd = bs_mat_c["Ens", ]*1e4,
+               title = "COSMO-E", waterfall = T)
```

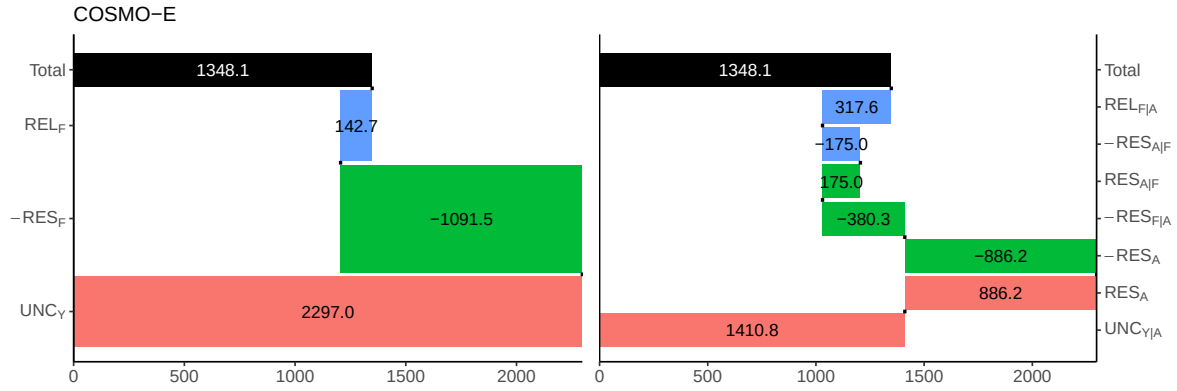

Figure 2: Waterfall plot containing the unconditional and conditional decomposition terms for the COSMO-E forecasts.

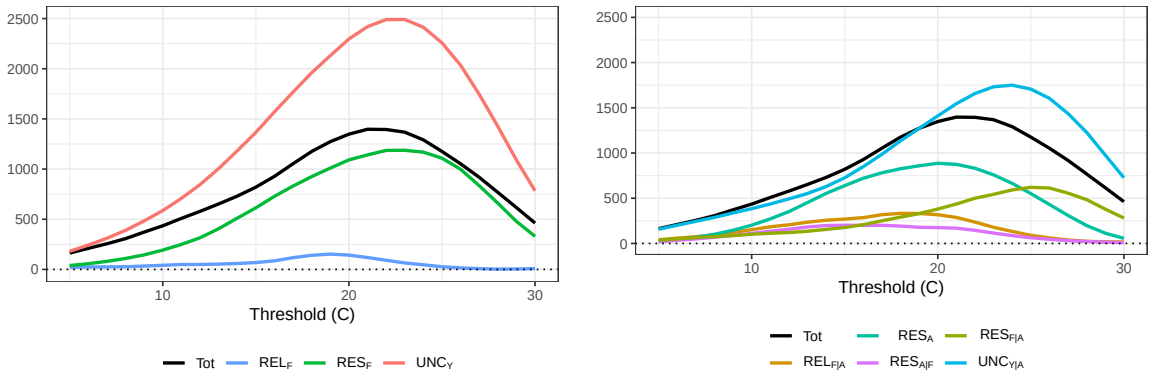

Figure 3: Terms of the unconditional decomposition (left) and the conditional decomposition (right) for the COSMO-E ensemble forecasts, as a function of the temperature threshold.

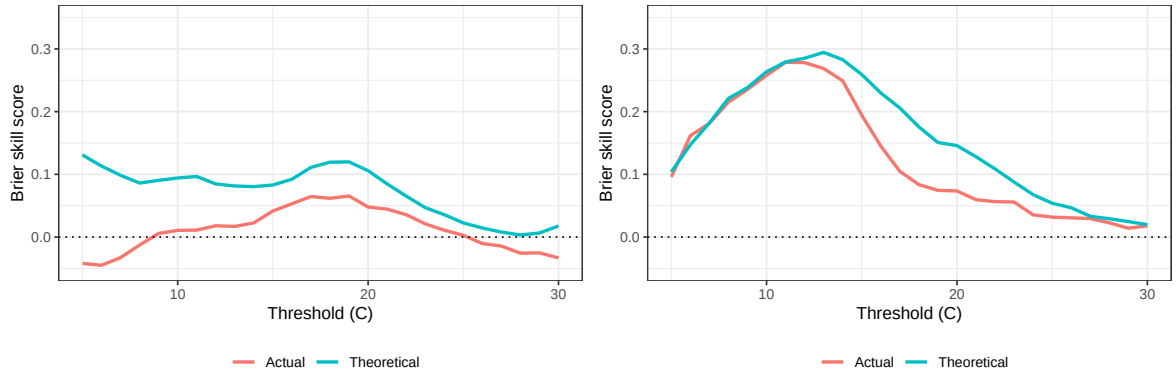

Figure 4: The theoretical relative improvement in the Brier score gained by statistically post-processing the COSMO-E probability forecasts, and the actual relative improvement gained in practice by the logistic regression model, shown as a function of the temperature threshold (left). And the theoretical and materialised relative improvement in the Brier score gained by incorporating altitude information into the statistical post-processing model, as a function of the temperature threshold (right).

the relative improvement (which corresponds to a Brier skill score) obtained by the logistic regression-based post-processing model, as a function of the threshold. The improvements gained by post-processing are typically below the theoretical improvements associated with an optimal re-calibration scheme, suggesting the chosen post-processing model is not the most effective approach, particularly for smaller thresholds.

Similarly, the ratio of $\text{RES}_{A|F}$ to the total score provides us with a theoretical measure of the improvement gained by incorporating the state information into the post-processing model. This can also be plotted against the actual improvements, as quantified by the Brier skill score of the conditionally post-processed forecasts with the standard post-processed forecasts as reference. If the actual improvements coincide with the theoretical improvements, then it suggests the state information is efficiently captured within the conditional post-processing model; conversely, a large difference between the theoretical and actual improvements indicates that this information is not optimally included in the model. Figure 4 illustrates that this information is well-captured for smaller thresholds, despite the model's simplicity, but not for intermediate thresholds.

4. Discussion

This vignette describes the conditional decompositions of proper scores proposed by Allen *et al.* (2023), and reproduces the results therein. This package is still in development, and there are several possible extensions that could be made. From an implementation standpoint, the `plot_decomp()` function is not robust to the plot window and scale of the variables, and can therefore lead to some values of the decomposition being unreadable.

There are also several theoretical extensions that could be included within the package. Firstly, the package currently only contains functionality to apply the conditional decomposition to the Brier score. However, as discussed in Section 2, estimators of the conditional

decomposition terms can readily be obtained for any proper scoring rule for binary outcomes, via Equation 20. Alternative scoring rules, e.g. the logarithmic score, could therefore easily be added.

Moreover, this general framework also applies to forecasts made for outcomes defined on more complex spaces. For example, for real-valued outcomes, the framework can be applied using the continuous ranked probability score (CRPS). However, in this case, there is currently no canonical approach with which to generate re-calibrated forecasts q. Work is ongoing to achieve this using isotonic distributional regression (Henzi *et al.* 2021), which additionally leads to non-negativity results for the resolution and reliability components.

Finally, the conditional decomposition can also be employed with more than one auxiliary variable, as discussed in Allen *et al.* (2023). It is generally difficult to reliably estimate the terms of this decomposition using the classical and bias corrected estimators — especially when the sample size is small, or the number of possible states is large — though the isotonic regression-based approach can handle this relatively easily. While the package can currently only handle single, discrete choices of the auxiliary variable, it should be straightforward to extend the package to deal with this.

References

Allen S, Ferro CAT, Kwasniok F (2023). "A conditional decomposition of proper scores: quantifying the sources of information in a forecast." *Quarterly Journal of the Royal Meteorological Society*.

Bröcker J (2009). "Reliability, sufficiency, and the decomposition of proper scores." *Quarterly Journal of the Royal Meteorological Society*, **135**, 1512–1519.

Bröcker J (2012). "Estimating reliability and resolution of probability forecasts through decomposition of the empirical score." *Climate Dynamics*, **39**, 655–667.

Dimitriadis T, Gneiting T, Jordan AI (2021). "Stable reliability diagrams for probabilistic classifiers." *Proceedings of the National Academy of Sciences*, **118**.

Ferro CAT, Fricker TE (2012). "A bias-corrected decomposition of the Brier score." *Quarterly Journal of the Royal Meteorological Society*, **138**, 1954–1960.

Henzi A, Ziegel JF, Gneiting T (2021). "Isotonic distributional regression." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **83**, 963–993.

Murphy AH (1973). "A new vector partition of the probability score." *Journal of Applied Meteorology*, **12**, 595–600.

Siegert S (2017). "Simplifying and generalising Murphy's Brier score decomposition." *Quarterly Journal of the Royal Meteorological Society*, **143**, 1178–1183.

Affiliation:

Sam Allen
University of Bern

Institute of Mathematical Statistics and Actuarial Science
Alpeneggstrasse 22
3012 Bern, Switzerland
E-Mail: sam.allen@unibe.ch
and
Oeschger Centre for Climate Change Research