Estimating reliability and resolution of probability forecasts through decomposition of the empirical score[*]

Jochen Bröcker

December 8, 2011

Abstract

Proper scoring rules provide a useful means to evaluate probabilistic forecasts. Independent from scoring rules, it has been argued that reliability and resolution are desirable forecast attributes. The mathematical expectation value of the score allows for a decomposition into reliability and resolution related terms, demonstrating a relationship between scoring rules and reliability/resolution. A similar decomposition holds for the empirical (i.e. sample average) score over an archive of forecast–observation pairs. This empirical decomposition though provides a too optimistic estimate of the potential score (i.e. the optimum score which could be obtained through recalibration), showing that a forecast assessment based solely on the empirical resolution and reliability terms will be misleading. The differences between the theoretical and empirical decomposition are investigated, and specific recommendations are given how to obtain better estimators of reliability and resolution in the case of the Brier and Ignorance scoring rule.

1 Probability forecasts, reliability, and resolution

It has long been noted that probabilities, if used and interpreted correctly, provide a consistent means to convey forecast information. Probability forecasts are now used operationally in a large number of different context, ranging from weather predictions to financial and economical forecasts. An ongoing discussion concerns the question of how to define and quantify value of probability forecasts. On the one hand, several virtuous forecast attributes have been identified, most importantly *reliability* and *resolution* (Murphy and Winkler, 1987; Murphy, 1993, 1996). On the other hand, quantitative measures of forecast performance have been proposed, most importantly *proper scoring rules* (Brier,

[*] *Climate Dynamics*, 2011. In the present version, several errors have been corrected.

1950; Brown, 1970; Savage, 1971; Matheson and Winkler, 1976; Gneiting and Raftery, 2007). The connection between scoring rules and the attributes reliability and resolution has been clarified in a series of papers (Murphy, 1973, 1996; Hersbach, 2000; Bröcker, 2009). The main conclusion of that research is that the average score of a forecast can be decomposed into terms which independently quantify the reliability and the resolution of the forecast; both attributes have a positive effect on the average score. These results will be revisited in Section 2.

For the purpose of forecast assessment, it would be of interest to estimate the reliability and resolution terms. The decomposition does not only hold for the true (ensemble) average of the score, but also for the empirical average (over a large sample of forecast–observation pairs; see Sec. 2.3). This suggests to use the reliability and resolution terms from the empirical decomposition as estimators for the corresponding terms in the true decomposition, which has become common practice at least in the meteorological community. The main purpose of this paper is to show that this is not a good idea. There are important differences between the true terms and their empirical counterparts. Roughly speaking, in truth the forecast tends to be more reliable than what the empirical reliability term suggests. At the same time, the true resolution tends to be less than the empirical resolution term. Clearly, any forecast assessment should take this into account. A simplified analysis of this phenomenon is presented in Section 3. Specific recommendations concerning better estimators of the terms are provided for the Brier score and the Ignorance score, two popular scoring rules. A more mathematical treatment will be provided in a forthcoming paper Bröcker (2011).

The remainder of the present section provides a short introduction to probabilistic forecasts, along with a discussion of the two forecast attributes of reliability and resolution. Let Y denote the quantity to be forecast, commonly referred to as the *observation, predictant, verification* or *target.* The observation Y is modelled here as a random variable taking values in a set $\mathbb{K}$, the *state space.* The state space is assumed to be a finite set of mutually exclusive alternatives (e.g. “rain, hail, snow, sunshine”), labelled $1 \dots K$. Values of Y (i.e. elements of $\mathbb{K}$) will be denoted by small lowercase letters like k or l. Any probability distribution over $\mathbb{K}$ is uniquely specified by a *probability vector*, by which we mean a K–dimensional vector p with nonnegative entries so that $\sum_{k=1}^{K} p_k = 1$. Generic probability vectors will be denoted by p and q. (Standard non–bold type will be used for vectors, and I promise that no symbol will be employed for both a vector and a scalar at the same time.) A *probabilistic forecasting scheme* is a scheme whereby information is compiled into probability vectors. More precisely, a probabilistic forecasting scheme is a random variable Γ, the values of which are probability vectors over $\mathbb{K}$. We can also think of Γ as a vector of K random variables $(\Gamma_1 \dots \Gamma_K)$ with the property that

$$\Gamma_k \geq 0 \text{ for all } k, \qquad \sum_k \Gamma_k = 1.$$

Since the probability vectors over $\mathbb{K}$ form a continuum, Γ is, in general, a random variable with continuous range. In order to simplify the presentation though, we assume that the forecasting scheme Γ assumes only a finite number of values $\gamma^{(1)}, \ldots, \gamma^{(D)}$. In situations where this assumption is not fulfilled, the forecasts first have to be mapped onto a new finite set of forecasts $\{\gamma^{(1)} \ldots \gamma^{(D)}\}$, for example by binning (see e.g. Bröcker, 2008a, as well as Section 4.2). We proceed here assuming this step has been applied.

Taking both the observation as well as the forecast as random variables is often referred to as the distributions–oriented approach, see for example Murphy and Winkler (1987); Murphy (1993, 1996). The reason for assuming Γ to be random is that forecasting schemes usually process information available before and at forecast time. For example, if Γ is a weather forecasting scheme with lead time 48h, it will depend on random weather information down to 48h prior to when the observation Y is obtained. The probabilistic association between Γ and the observation Y then determines the predictive power of Γ.

We will now briefly revisit the concepts of *reliability* and *resolution*, which are widely agreed upon as being desirable properties of probability forecasts. On the condition that the forecasting scheme Γ is equal to, say, the probability vector p, the observation Y should be distributed according to p. Reliability means that this holds for all p. We can express reliability as the condition that

$$\mathbb{P}\left[Y = k \,\middle|\, \Gamma = \gamma^{(d)}\right] = \gamma_k^{(d)} \qquad \text{for all } d \text{ and } k \in \mathbb{K}. \tag{1}$$

In view of Equation (1), we define $\pi_k^{(d)} = \mathbb{P}\left[Y = k \,\middle|\, \Gamma = \gamma^{(d)}\right]$ for the conditional probability of the observation being k, given that the forecast is equal to $\gamma^{(d)}$. The reliability condition (1) can then be written simply as $\pi^{(d)} = \gamma^{(d)}$ for all d.

In the literature, other concepts of reliability can be found which are related to but different from the reliability concept presented here. In general, those definitions of reliability are weaker than (i.e. implied by) reliability as defined here. An example is to merely require that on average, the forecast Γ agrees with the climatology of Y (see Sec. 2.2 for a definition). Several concepts of reliability are compared and contrasted in Hamill (2001). Gneiting et al (2007) give a more mathematical treatment; although their formalism is slightly different, the present definition of reliability is equivalent in spirit to what Gneiting et al refer to as *calibration*.

Reliability alone does not make for a useful forecast; for example, even the unconditional probability distribution of Y constitutes a reliable forecast. This forecast though is constant and therefore unable to delineate different values of the observation Y. Therefore, as another desirable forecast attribute, *resolution* has been advocated. Definitions of resolution are often found to be somewhat vague in the literature, although it is generally agreed upon that resolution should depend on π rather than on Γ itself, and that it should be related to the information content of the former, relative to some standard forecast such as the unconditional probability. A more precise definition was given in Bröcker (2009), to be revisited in Subsection 2.2.

2 Probabilistic scoring rules, and their decomposition

In this section, the concept of scoring rules for probability forecasts will be revisited. Scoring rules can be thought of as a way to assign “points” or “rewards” to probability forecasts, providing quantitative indication of success in predicting the observation, in contrast to the more qualitative notions of forecast value discussed in the previous section. Scoring rules measure the success of a single forecast; the overall score of a forecasting scheme is then taken as the average score over individual cases. There are performance measures which cannot be written as an average score over individual instances, one popular example being the Receiver Operating Characteristic (see for example Egan, 1975).

2.1 Definition and properties of scoring rules

A *scoring rule* (see for example Matheson and Winkler, 1976; Gneiting and Raftery, 2007) is a function $S(p, k)$ which takes a probability vector p as its first argument and an element k of $\mathbb{K}$ as its second argument. We will take a small score as indicating a good forecast. One of the most commonly used scoring rules is the Brier score (Brier, 1950), which applies to the case where $k = 0$ or 1 and reads as

$$S(p, k) = (p_1 - k)^2.$$

The Brier score generalises to situations in which $\mathbb{K} = \{1 \dots K\}$ as follows:

$$S(p, k) = \sum_{l=1}^{K} (p_l - \delta_{k,l})^2,$$

where $\delta_{k,l} = 1$ if $k = l$ and 0 otherwise. The Ignorance (or logarithmic scoring rule) is also popular and reads as

$$S(p, k) = -\log(p_k).$$

Many authors have argued that in order to avoid inconsistencies, scoring rules should be *strictly proper* (see also Brown, 1970; Bröcker and Smith, 2007). To define this concept, consider the *scoring function*, which for any two probability vectors p and q is defined as

$$\mathsf{s}(p, q) = \sum_{k \in \mathbb{K}} S(p, k) q_k. \tag{2}$$

The scoring function is to be interpreted as the mathematical expectation value (ensemble average) of the score of a forecasting scheme which issues constant forecasts equal to p for a random variable Y which has in fact distribution q. But if the actual distribution of Y is q, then any probability vector different from q should have a worse average score than q itself. This is the essence of the following definition. A score is called *proper* if the *divergence*

$$\mathsf{d}(p, q) = \mathsf{s}(p, q) - \mathsf{s}(q, q) \tag{3}$$

is nonnegative, and it is called *strictly proper* if it is proper and $\mathsf{d}(p, q) = 0$ implies $p = q$. For strictly proper scoring rules, $\mathsf{d}(p, q)$ can be interpreted as a measure of dissimilarity between p and q. From now on, the term "scoring rule" is supposed to mean "strictly proper scoring rule", unless otherwise stated. The fact that $\mathsf{d}(., ..)$ is positive definite will be exploited extensively in this paper.

For later reference, we define

$$\mathsf{e}(p) = \mathsf{s}(p, p) \tag{4}$$

as the *entropy* associated with the scoring rule S. This nomenclature is motivated by the fact that for the Ignorance, one has $\mathsf{e}(p) = -\sum \log(p_k)p_k$. This quantity is known as the entropy in both statistical physics as well as information theory. For later reference, we note that for the Brier score,

$$\mathsf{e}(p) = 1 - \sum_k p_k^2. \tag{5}$$

2.2 Decomposition of the score's expectation value

A scoring rule provides a means to evaluate a probabilistic forecasting scheme individually for each forecast instant. Therefore, the mathematical expectation value $\mathbb{E}(S(\Gamma, Y))$, henceforth referred to as the *true score*, can be interpreted as an average measure of forecast quality. (Since Γ is random, taking the expectation affects both Γ and Y.) The true score allows for a decomposition into several terms which can be interpreted in terms of resolution and reliability. This result provides a link between resolution and reliability, which are qualitative notions of forecast value, with the true score, which is a quantitative notion of forecast value, thereby justifying the scoring rule methodology. (To get a decomposition of a forecast with a continuum of possible values, the relation (13) below needs to be modified, see Stephenson et al (2008).)

Throughout the paper, we will use the following shorthand

$$\rho_d = \mathbb{P}(\Gamma = \gamma^{(d)})$$

for the probability that Γ assumes the value $\gamma^{(d)}$. Further, we let

$$\bar{\pi}_k = \mathbb{P}(Y = k)$$

be the unconditional distribution of Y, often referred to as the climatology. Note that $\bar{\pi}_k$ is the average of $\pi_k^{(d)}$, in the sense that

$$\bar{\pi}_k = \sum_d \mathbb{P}(Y = k | \Gamma = \gamma^{(d)}) \, \mathbb{P}(\Gamma = \gamma^{(d)}) = \sum_d \pi_k^{(d)} \rho_d.$$

The decomposition of the true score will be presented here as two statements in Equations (6) and (8). For mathematical proofs, the interested reader is deferred to the classical paper by Murphy (1973) for the Brier score (and for binary forecasting problems), to Hersbach (2000) for the Continuous Ranked

Probability Score (not discussed here), and to Bröcker (2009) for the general case.

The first decomposition reads as

$$\mathbb{E}(S(\Gamma, Y)) = \sum_d \mathsf{e}(\pi^{(d)})\rho_d + \sum_d \mathsf{d}(\gamma^{(d)}, \pi^{(d)})\rho_d. \tag{6}$$

The second term on the right hand side in Equation (6) is positive definite, and referred to as the *reliability term.* Recalling that $\gamma^{(d)} = \pi^{(d)}$ indicates a reliable forecast, we see that the reliability term quantifies the average violation of reliability. The first term in Equation (6) is referred to as the *potential score* by Hersbach (2000). This is for the following reason. Consider a forecasting scheme Π defined to be

$$\Pi = \pi^{(d)} \qquad \text{whenever } \Gamma = \gamma^{(d)}. \tag{7}$$

Then the potential score can be shown to be the true score of the forecasting scheme Π, namely

$$\sum_d \mathsf{e}(\pi^{(d)})\rho_d = \mathbb{E}(S(\Pi, Y)).$$

Now Π is, by construction, a reliable forecasting scheme, and Equation (6) says that Π achieves a better true score than Γ, confirming our intuitive understanding that reliability is a virtuous forecast property. The potential score will be subject to the next decomposition, which is

$$\sum_d \mathsf{e}(\pi^{(d)})\rho_d = \mathsf{e}(\bar{\pi}) - \sum_d \mathsf{d}(\bar{\pi}, \pi^{(d)})\rho_d. \tag{8}$$

This relation gives a concise description of the potential score in terms of the fundamental uncertainty $\mathsf{e}(\bar{\pi})$ of Y, less a positive definite term called the *resolution term.*

To interprete the resolution term, we first recall that the expectation value of Π is $\bar{\pi}$. Hence the resolution term describes, roughly speaking, the average deviation of Π from its expectation value $\bar{\pi}$; it can therefore be interpreted as a form of "variance" of Π. The resolution term is indeed given by the usual variance of Π in case of the Brier score. The uncertainty $\mathsf{e}(\bar{\pi})$ can be seen as the true score of the climatology as a forecasting scheme. Hence, the entropy quantifies the ability of the climatology to forecast random samples from itself. As an aside, we note that the decomposition is also valid for improper scores, but then the divergence d is no longer positive definite, and improving the forecast in terms of reliability or resolution does not necessarily yield a better score.

2.3 Decomposition of the empirical score

We will now discuss a decomposition similar to the previous one, albeit not for the true score, but rather for the empirical (sample) average of the score over an *archive* T of forecast–observation pairs. More specifically, $T = \{(\Gamma(n), Y(n)), n =$

$1 \dots N\}$, where $(\Gamma(n), Y(n))$ are independent and identically distributed realisations of the forecast–observation pair (Γ, Y). The vector components of $\Gamma(n)$ are still written as subscripts, that is $\Gamma(n) = (\Gamma_1(n) \dots \Gamma_K(n))$, and we have

$$\Gamma_k(n) \geq 0 \text{ for all } k, n, \qquad \sum_k \Gamma_k(n) = 1 \text{ for all } n.$$

Given an archive T, the true score can be estimated by the *empirical score*

$$\hat{S} = \frac{1}{N} \sum_{n=1}^{N} S\left(\Gamma(n), Y(n)\right).$$

A decomposition similar to the one we have seen in Section 2.2 holds for the empirical score, to be discussed now. Let N_{kd} be the number of instances in the data set where $\Gamma(n) = \gamma^{(d)}$ and at the same time $Y(n) = k$. Further, let $N_{\bullet d} = \sum_k N_{kd}$ be the number of instances in the data set where $\Gamma(n) = \gamma^{(d)}$, that is, we sum over the rows of the contingency table N_{kd} and replace k with a bullet $\bullet$. Obviously, $\sum_d N_{\bullet d} = N$, the total number of instances in the data set. Now, define for $d = 1 \dots D$ and $k = 1 \dots K$ the *relative observed frequencies*

$$o_k^{(d)} = \frac{N_{kd}}{N_{\bullet d}}. \tag{9}$$

For every d, the vector $o^{(d)} = (o_1^{(d)}, \dots, o_K^{(d)})$ is a probability vector. Moreover, if we let the number of instances go to infinity, then due to the law of large numbers,

$$o_k^{(d)} \to \pi_k^{(d)}, \tag{10}$$

so that $o_k^{(d)}$ can be interpreted as an estimate of $\pi_k^{(d)}$. Furthermore, we set

$$\bar{o}_k = \frac{N_{k\bullet}}{N}$$

with $N_{k\bullet} = \sum_d N_{kd}$. Note that $\bar{o}_k$ is the empirical average of $o_k^{(d)}$ over d. On the other hand, $\bar{o} = (\bar{o}_1, \dots, \bar{o}_K)$ is an approximation of the climatological probability distribution of Y, namely

$$\bar{o}_k \to \bar{\pi}_k \qquad \text{for } N \to \infty. \tag{11}$$

Finally, we set

$$r_d = \frac{N_{\bullet d}}{N}.$$

If we let the number of instances go to infinity, then

$$r_d \to \rho_d. \tag{12}$$

In terms of these objects, the decompositions

$$\frac{1}{N} \sum_n S(\Gamma(n), Y(n)) = \sum_d \mathsf{e}(o^{(d)})\, r_d + \sum_d \mathsf{d}(\gamma^{(d)}, o^{(d)})\, r_d \tag{13}$$

and

$$\sum_d \mathsf{e}(o^{(d)})\, r_d = \mathsf{e}(\bar{o}) - \sum_d \mathsf{d}(\bar{o}, o^{(d)})\, r_d \tag{14}$$

hold. The decompositions (13) and (14) might be considered as empirical versions (i.e. summation over samples) of the decompositions (6) and (8). The first and second terms in the decomposition (13) are referred to as the *empirical potential score* and the *empirical reliability*, respectively, while the terms in the decomposition (14) are called *empirical uncertainty* and *empirical resolution*, respectively.

Empirical (i.e. sample) averages are commonly used to estimate ensemble averages. In particular, the empirical score can be employed as an estimator of the true score. Indeed, due to the law of large numbers, the fluctuations in the empirical score are expected to become small, and furthermore we have the identity

$$\mathbb{E}\left(\frac{1}{N}\sum_n S(\Gamma(n), Y(n))\right) = \mathbb{E}\left(S(\Gamma, Y)\right), \tag{15}$$

At first glance, it might appear reasonable to use the terms in the empirical decomposition to estimate the corresponding terms in the true decomposition. Even more so, because we can formally arrive at Equations (6) and (8) if we replace the quantities r, o and $\bar{o}$ with their respective limit values $\rho, \pi, \bar{\pi}$ in the relations (13) and (14). We will see though that this is not a good idea, as there are systematic deviations between the empirical and their corresponding true terms in the decomposition. These deviations will be calculated explicitely in Sections 3.1 and 3.2 for the Ignorance score and the Brier score, respectively; in the remainder of the present section, we will discuss heuristic arguments as to why the empirical decomposition terms give a misleading picture of the true forecast value.

In view of the decomposition (13, 14), we might think about defining a new, "calibrated" probabilistic forecasting scheme Φ as follows: whenever $\Gamma(n) = \gamma^{(d)}$ for some d, we set $\Phi(n) = o^{(d)}$. Note that Φ is an approximation of the reliable forecast Π discussed above by Equation (7). The empirical score of this forecast is given by the empirical potential score

$$\frac{1}{N}\sum_n S(\Phi(n), Y(n)) = \sum_d \mathsf{e}(o^{(d)})\, r_d. \tag{16}$$

The decomposition (13) shows that the empirical score of Φ is always better than that of Γ (the empirical reliability term of Φ would be zero). This seems to suggest a foolproof way of improving forecast skill. We should be suspicious though about the fact that this strategy would *always* improve the score, even if the original forecast were in fact reliable! How can this be? Roughly speaking, this comes about because we "recalibrate" not only the real deviations from reliability, but also those arising merely through sampling variations. This has a systematic and detrimental effect. Although Φ approximates Π, the two are not identical and therefore Φ is in fact not fully reliable. Hence, the true reliability term of Φ is larger than zero, and therefore the true score of Φ is worse than its empirical score. Another way to see this is that we are evaluating the recalibrated forecast Φ "in sample", meaning that the same data is used to both build and then evaluate Φ. The forecast Φ "knows" the data already and appears to be better than it actually would be it were evaluated on instances of the data set that were not used to generate Φ.

Clearly, these conclusions should bear on the interpretation of the empirical decomposition as well as on any possible application. For example, we might contemplate to use the resolution term to determine an appropriate binning for a forecasting scheme with a continuous range. The resolution term though becomes the better the more bins we use, since the information in the forecast becomes ever more detailed. Indeed, if we went to the extreme and set the bins so that each bin contains exactly one forecast ($\Gamma(n)$, say), then the corresponding observed frequency would be built upon a single observation $Y(n)$. Thus, the recalibrated forecast Φ would seem to be perfect. In truth, Φ is only good in forecasting observations which have been used already to construct Φ, but utterly useless in forecasting *new* observations.

3 Comparison between empirical and true terms in the decomposition

In view of the practical significance of the score decompositions, it would be of value to have at least some approximation of the difference between the true and the empirical decomposition terms. Such approximation will now be provided; we focus on the two most common scoring rules, namely the Ignorance and the Brier score.

3.1 The Ignorance score

The central result of this section are the following relations between the empirical uncertainty, resolution, and reliability with their respective true counterparts:

$$\mathbb{E}\left(\mathsf{e}(\bar{o})\right) = \mathsf{e}(\bar{\pi}) - \frac{K-1}{2N} \tag{17}$$

$$\mathbb{E}\left(\sum_{d=1}^{D} \mathsf{d}(\bar{o}, o^{(d)}) r_d\right) = \sum_{d=1}^{D} \mathsf{d}(\bar{\pi}, \pi^{(d)})\rho_d + \frac{(K-1)(D-1)}{2N} \tag{18}$$

$$\mathbb{E}\left(\sum_{d=1}^{D} \mathsf{d}(\gamma^{(d)}, o^{(d)}) r_d\right) = \sum_{d=1}^{D} \mathsf{d}(\gamma^{(d)}, \pi^{(d)})\rho_d + \frac{(K-1)D}{2N} \tag{19}$$

The proof has been deferred to Appendix A. These relations allow for interesting conclusions. The empirical resolution term overestimates the true resolution of the forecast, that is, it suggests more resolution than can in fact be obtained

(Eq. 18). Further, the empirical reliability term overestimates the true reliability term, so that in truth, the original forecast Γ is *more* reliable than the empirical reliability term suggests (Eq. 19). These effects diminish with increasing number of samples N, but grow with the number of categories K and also with the number of different forecast values D. This means that binning a continuous forecast Γ among too many bins will increase the empirical reliability and resolution terms, therefore spuriously inflating the potential score. The Ignorance is special in that the correction terms depend only on the dimensions of the forecasting problem.

Practical recommendations for the Ignorance

The relations (17–19) should be interpreted as

$$\mathsf{e}(\bar{\pi}) \cong \mathsf{e}(\bar{o}) + \frac{K-1}{2N}, \tag{20}$$

$$\sum_{d=1}^{D} \mathsf{d}(\bar{\pi}, \pi^{(d)})\rho_d \cong \sum_{d=1}^{D} \mathsf{d}(\bar{o}, o^{(d)})r_d - \frac{(K-1)(D-1)}{2N}, \tag{21}$$

$$\sum_{d=1}^{D} \mathsf{d}(\gamma^{(d)}, \pi^{(d)})\rho_d \cong \sum_{d=1}^{D} \mathsf{d}(\gamma^{(d)}, o^{(d)})r_d - \frac{(K-1)D}{2N}. \tag{22}$$

That is, the right hand sides provide estimates of the left hand sides. These estimates are easily computed in practice. They consist of the empirical uncertainty, resolution, and reliability terms, augmented by certain corrections which do not even depend on the actual forecast and observation data but just on the dimensions of the problem. It needs to be kept in mind though that these are still only estimators. Further, in their derivation, we have assumed that ρ, π, and $\bar{\pi}$ are not too different from their estimates r, o, and $\bar{o}$. A more in–depth analysis shows that if this is true, then the correction terms have χ^2 distributions, and then the relations (17–19) are exact. It is often recommended that for the χ^2 distribution to apply, no entry in the contingency table N_{kd} should be less than five.

3.2 The Brier score

We will now derive similar results for the Brier score. Again, the proof can be found in the Appendix. We claim that

$$\mathbb{E}\,\mathsf{e}(\bar{o}) = \mathsf{e}(\bar{\pi}) - \frac{1}{N}\mathsf{e}(\bar{\pi}), \tag{23}$$

$$\mathbb{E}(\sum_d \mathsf{d}(\bar{o}, o^{(d)})r_d) = \sum_d \mathsf{d}(\bar{\pi}, \pi^{(d)})\rho_d + \frac{1}{N}\left(\sum_d \nu_d \mathsf{e}(\pi^{(d)}) - \mathsf{e}(\bar{\pi})\right), \tag{24}$$

$$\mathbb{E}(\sum_d \mathsf{d}(\gamma^{(d)}, o^{(d)})r_d) = \sum_d \mathsf{d}(\gamma^{(d)}, \pi^{(d)})\rho_d + \frac{1}{N}\sum_d \nu_d \mathsf{e}(\pi^{(d)}), \tag{25}$$

with $\nu_d = 1 - (1 - \rho_d)^N$, a quantity which is effectively unity unless both N and ρ_d are very small. In the original version of this paper, it was claimed that that the term in brackets on the right hand side of Equation (24) is never negative, which would imply that the empirical resolution overestimates the true resolution. For the Brier score though, this statement is wrong in general. The proof assumed that the entropy e is convex, while in fact it is concave. However, it is worth noting that the term in question is positive for many choices of $\pi^{(d)}, \rho_d, d = 1 \ldots D$. To make this more precise, note that for the Brier score, $\mathsf{e}(p) = 1 - \sum_k p_k^2$. Therefore, the relation

$$\sum_d \mathsf{e}(\pi^{(d)})\nu_d \geq \mathsf{e}(\bar{\pi}) \tag{26}$$

evidently holds under the condition that the $\pi^{(d)}$ live in a sphere of radius $\sqrt{1 - \frac{\mathsf{e}(\bar{\pi})}{\sum_d \nu_d}}$, centered at the origin. A simple calculation shows that $\mathsf{e}(\bar{\pi}) \leq \frac{K-1}{K}$, while the ν_d are very nearly equal to one, unless N is very small, as discussed. We can conclude that the empirical resolution overestimates the true resolution if the $\pi^{(d)}$ are restricted to a sphere of radius no less than $\sqrt{1 - \frac{K-1}{KD}}$.

The correction terms in Equations (23) and (25) are indeed non–negative because the entropy of the Brier score is never negative. We can conclude that for the Brier score, the empirical resolution and reliability terms *typically* overestimate the true resolution and reliability terms. Again, in truth the forecast is more reliable and less resolved than the empirical decomposition suggests.

Practical recommendations for the Brier score

Due to the dependence of the correction terms in Equations (23–25) on the unknown quantities $\pi^{(d)}, \bar{\pi}$, and ν_d, they seem to be of less practical value than the corresponding equations for the Ignorance, which only depend on K, N,and D. As a possible remedy, we can replace these unknown quantities by their sample estimators, thereby obtaining at least rough guidance as to the magnitude of the deviations.

$$\mathsf{e}(\bar{\pi}) \cong \mathsf{e}(\bar{o}) + \frac{1}{N}\mathsf{e}(\bar{o}), \tag{27}$$

$$\sum_d \mathsf{d}(\bar{\pi}, \pi^{(d)})\rho_d \cong \sum_d \mathsf{d}(\bar{o}, o^{(d)})r_d - \frac{1}{N}\left(\sum_d \mathsf{e}(o^{(d)}) - \mathsf{e}(\bar{o})\right), \tag{28}$$

$$\sum_d \mathsf{d}(\gamma^{(d)}, \pi^{(d)})\rho_d \cong \sum_d \mathsf{d}(\gamma^{(d)}, o^{(d)})r_d - \frac{1}{N}\sum_d \mathsf{e}(o^{(d)}), \tag{29}$$

The right hand sides provide estimates of the left hand sides, and again are easily computed in practice. A few words of caution might seem in order when using Equations (27–29), as these are again only estimators. In their derivation, we have assumed that ρ, π, and $\bar{\pi}$ are not too different from their estimates r, o, and $\bar{o}$. Taking the discussion for the Ignorance as a guidance, no entry in the contingency table N_{kd} should be less than five.

4 Example

4.1 Artificial data

First, an example using artificial data will be discussed. More specifically, we will simply set the probabilities $\pi_k^{(d)}$ and ρ_d for all $k = 1 \dots K$ and $d = 1 \dots D$. This will be done so that we can vary D in order to generate various experiments. Next, we set $\gamma^{(d)} = Q(\pi^{(d)})$, where for Q we choose some function mapping probability vectors on probability vectors, in order to introduce slight deviations from reliability. We then generate data $\{(\Gamma(n), Y(n)), n = 1 \dots N\}$ by applying for every n independently the following protocol:

1. Draw a number δ from $1 \dots D$ so that the probability of $\delta = d$ is equal to ρ_d. Set $\Gamma(n) = \gamma^{(\delta)}$
2. Draw $Y(n)$ from $1 \dots K$ so that the probability of $Y(n) = k$ is equal to $\pi_k^{(d)}$.

We still have to define $\pi_k^{(d)}$ and ρ_d. To keep matters simple, we will fix $K = 3$. The set of all three dimensional probability vectors is called the standard 2–simplex. The standard 2–simplex can be subdivided into M^2 smaller simplices of equal size, with vertices given by

$$(\frac{k}{M}, \frac{l}{M}, \frac{m}{M}), \qquad k, l, m = 0 \dots M, \qquad k + l + m = M.$$

We denote these simplices by Δ_d, with d running from 1 to $M^2 = D$. We then choose $\pi^{(d)}$ as the centre point of the simplex Δ_d, for each d. These points have coordinates

$$(\frac{k + 1/3}{M}, \frac{l + 1/3}{M}, \frac{m + 1/3}{M}), \qquad k, l, m = 0 \dots M - 1, \qquad k + l + m = M - 1$$

and

$$(\frac{k + 2/3}{M}, \frac{l + 2/3}{M}, \frac{m + 2/3}{M}), \qquad k, l, m = 0 \dots M - 2, \qquad k + l + m = M - 2$$

Finally, we set $\rho_d = 1/D$ for all d. We can think of $\pi^{(d)}$ as a coarse grained version of some random probability vector with uniform distribution over the standard 2–simplex. Our eventual forecasts are given by distorting π as follows:

$$\gamma_k^{(d)} = c_d \cdot (\pi_k^{(d)})^{1 + k/2}$$

where c_d is a normalisation constant. Thereby, we introduce some mild deviation from reliability.

We generated data sets T comprising $N = 365$ forecast–observation pairs, imitating a year's worth of data. We are now in a position to compute all terms in both the empirical and the true decomposition. In order to validate

Equations (17–19) though, we need the expectation value of the empirical uncertainty, resolution, and reliability. In order to approximate these, we generated 100 statistically identical copies of the data set T. For each copy, the empirical uncertainty, resolution, and reliability terms where computed; the results were averaged, and also the standard deviations were recorded.

Figure 1 shows the results for the Ignorance score. The entire analysis was carried out for several values of D, shown on the x–axes. (In fact, M runs from 1 to 6 in these plots, and $D = M^2$.) All plots show the empirical term minus the true term, divided by the true score; here "term" is to be read as "score" in plot (a), "uncertainty" in plot (b), "reliability" in plot (c), and "resolution" in plot (d). The average of these values is indicated by circles, with 2σ confidence bars attached. The dashed lines show what our theory predicts for these values. Overall, it appears that theory and experiments are in fairly good agreement.

More specifically, plot (a) demonstrates that, up to statistical fluctuations, the expectation value of the empirical score is given by the true score, thereby verifying Equation (15). Plot (b) shows the difference between the empirical and true uncertainty term, relative to the true score, marked with balls. Firstly, the empirical uncertainty underestimates the true uncertainty, consistent with our predictions. Equation (17) asserts that the difference should be equal to $-(K-1)/N$; this is indicated with the dashed line. The empirical and true uncertainty term do not depend on D, but the true score does, whence the dashed line is not constant. Evidently, the experiments show that Equation (17) is valid. Plot (c) shows the difference between the empirical and true reliability term, relative to the true score, marked with balls. Again, consistent with our qualitative statements, the empirical reliability overestimates the true reliability (by up to 10% of the true score in this case). Equation (18) asserts that the difference should be equal to $(K-1)(D-1)/N$; this is indicated with the dashed line. Theory and experiment turn out to be in qualitative agreement, with small but apparently systematic deviations, discussed below. Plot (d) finally shows the difference between the empirical and true resolution term, relative to the true score, marked with balls. As predicted by theory, the empirical resolution term overestimates the true resolution term (again by up to 10% of the true score). Equation (19) asserts that the difference should be equal to $(K-1)D/N$; this is indicated with the dashed line. Again, theory and experiment agree qualitatively. It seems though that the empirical resolution as well as the reliability are larger than what would be consistent with our theory, especially for large values of D. It is to be noted that this does not disprove our qualitative statement that the empirical resolution overestimates the true resolution; it is only that our quantitative theory still underestimates this effect. The likely reason for this difference is the discussed approximation on which this estimate is based. It stops to be valid if $N_{k,d}$ is very small for some k, d. The expectation value of $N_{k,d}$ can be as small as $1/D^2$, which is on the order of $1/N$ for $D \cong 20$. Thus, for D larger than this value, it is almost certain that $N_{k,d} = 0$ for some k, d, violating the condition that $N_{k,d}$ be at least five.

Results for the Brier score are shown in Figure 2. The main findings are very similar to the case of the Ignorance score. We will therefore discuss a few important points, only. The agreement between theory and experiment is even better than in the case of the Ignorance, which is to be expected, as no approximations are involved. As mentioned in the recommendations for the Brier score, in order to apply the corrections, the true values for $\bar{\pi}_k, \pi_k^{(d)}$, and ρ_d have to be replaced with their respective approximations $\bar{o}, o^{(d)}$ and r_d. The dotted lines have been obtained in this way. It appears to be a good approximation in the present situation, but further investigation is needed to confirm the general suitability of this approach.

4.2 Weather data

We now present an example using actual weather forecasting data. In terms of measurements, we use two metre temperature data from the weather station at Heligoland in the German Bight, taken daily at 12:00 UTC. In terms of forecasts, dynamical weather forecasts from the European Centre for Medium Range Weather Forecasts (ECMWF) for two metre temperature are used. The ECMWF maintains an ensemble prediction system with 50 members. The ensemble members comprise runs of a global weather model, each generated with slightly perturbed initial conditions (there is also an unperturbed run, the control, which is not used here). Forecasts were available from 1 January, 2001, until 31 December, 2005, from the then operational ECMWF prediction system, featuring lead times from one to ten days and a spatial resolution (for the ensembles) of about 80 kilometres. (For more information, the reader is referred to Persson and Grazzini, 2005). Only the forecast information relevant for Heligoland is used. We will focus on lead times of 7 days, only, for which we have 1812 forecast–observation pairs. Next, the data is modified in the following way:

1. A climate normal is computed by fitting a fourth order trigonometric polynomial to the temperature data.

2. We define $Y(n)$ to be one if the temperature on day n falls below the climate normal, while $Y(n) = 2$ if the temperature exceeds the normal. Thereby, we have a binary forecasting problem, that is, $\mathbb{K} = \{1, 2\}$;

3. We set $\Gamma_1(n) = \frac{m+\frac{1}{2}}{M+1}$, where m is the number of ensemble members below the climate normal, and $M = 50$ is the total number of ensemble members. Clearly $\Gamma_2(n) = 1 - \Gamma_1(n) = \frac{M-m+\frac{1}{2}}{M+1}$.

Using this archive of forecast–observation pairs, we are going to investigate the Brier score for the forecasting scheme Γ. More specifically, we will calculate the empirical score as well as the empirical uncertainty, resolution, and reliability term on the left hand side of Equations (23–25). Further, we will compute the correction terms on the right hand side of these Equations (i.e. the terms carrying the prefactor $1/N$), using the approximation discussed before, namely we replace the true probabilities $\bar{\pi}$ and $\pi^{(d)}$ with their respective approximations

$\bar{o}$ and $o^{(d)}$. The experiments using artificial data provided tentative confirmation that this approximation is justified.

We are now in a position to address the following two questions. Firstly, does our theory yield appreciable corrections to the empirical uncertainty, resolution, and reliability term in this example? Our second question is related to the estimators provided for the true uncertainty, resolution, and reliability terms through Equations (27–29). In particular, the right hand sides should (at least roughly) be independent of N. This is the second issue we are going to check. Note that by construction, Γ can assume at most 51 values. Hence there is, in this example, some sort of "natural" value for D. As in the example using artificial data though, we would like to carry out the analysis for several values of D, as we expect a strong dependence of the corrections on D. For this reason, we will distribute the forecasts $\{\Gamma_n, n = 1 \dots N\}$ among D different bins of roughly equal population and use the in–bin average of Γ_n as the new set of coarse–grained forecasts. The investigated values for D were $3, 5, 8, 12, 17$, and 23. It is not suggested here that these would be appropriate values of D for a serious performance assessment of the discussed forecasts. Although the results of this paper are, in principle, valid for any D (apart from the reservations mentioned at the end of Secs. 3.1, 3.2), too large values of D will clearly have detrimental effect on the results. For example, the relative observed frequencies $r, o, \bar{o}$ will feature a large variance if D is too large, and so will the empirical uncertainty, resolution, and reliability terms, diminishing their information content. The question of how to choose D (more generally, the bins) appropriately probably merits further investigation, although some suggestions have been made, see for example Bröcker (2008b).

Figure 3 demonstrates that our theory suggests considerable corrections to the empirical uncertainty, resolution, and reliability term, and hence the first question can be answered in the affirmative. Plot (a) shows the correction to the reliability term, relative to the empirical reliability term itself, as a function of D. Plot (b) shows the correction to the resolution, relative to the empirical resolution term itself, again as a function of D. We see that the corrections to the empirical reliability term can amount to up to 35%. All quantities were computed using $N = 1095$ instances which were drawn randomly and with replacement from the entire data set of 1812 samples. This resampling experiment was repeated 100 times, allowing to estimate mean and standard deviations for all quantities. The width of the error bars in Figure 3 represents two standard deviations. Although we have five years worth of data available in total, the resampled data sets comprised only three years as this was deemed to be more realistic for many applications in weather and climate.

Figure 4 shows the empirical score (plot a) as well as the estimates (as suggested by Eqs. 27–29), of the true uncertainty, resolution, and reliability terms (plot b–d, respectively). All quantities are plot as a function of D. In order to check the dependence of these estimates on the sample size N, we computed the estimates for different values of N, namely $N = a \cdot 365$ with $a = 2 \dots 5$ years. Results for two years are presented with error bars (two standard deviations obtained through resampling the entire data set). The other lines correspond to $a = 3$ years (solid), $a = 4$ years (dashed), and $a = 5$ years (dotted). Clearly, the results are not independent of N, but the differences are on the same order of magnitude as the sampling variations represented by the error bars.

5 Concluding remarks

Scoring rules provide a useful means to evaluate probabilistic forecasts (as long as only strictly proper scoring rules are employed). The mathematical expectation value of the score allows for a decomposition into terms which quantify the reliability and the resolution of the forecast. (Reliability and resolution are desirable forecast attributes for which the case can be made independently of scoring rules.) A similar result holds for the empirical (or sample average) score over an archive of forecast–observation pairs, decomposing the empirical score into empirical resolution and reliability terms. It has been demonstrated in this paper that the empirical reliability and resolution terms do not agree well with the true reliability and resolution. The empirical reliability is too large, suggesting a too large departure from reliability. As a consequence, the empirical decomposition provides a too optimistic estimate of the potential score (i.e. of the optimum score which could be obtained through recalibration). Hence, a forecast assessment based solely on the empirical resolution and reliability terms will be misleading. Specific recommendations have been given as to how better estimators of reliability and resolution can be obtained in the case of the Brier and Ignorance Score. Our theoretical investigations have been tested and confirmed in a numerical experiment using artificial data. Furthermore, the practical feasibility of the given recommendations was demonstrated in a numerical example using actual weather data.

There are several points which call for further investigation. In the statistical analysis of forecast verification methods, it is mostly assumed that the individual samples, or forecast–observation pairs, are serially independent, or at least very nearly so. It is clear however that weather data, both observations as well as forecasts, display strong temporal correlations in general. Even though we are not using the observations directly but rather the anomalies, assuming them to be serially independent is clearly an idealistic assumption. The question how serial dependencies enter in the statistical analysis of reliability and resolution of probabilistic forecast remains a subject of future investigation. The obvious difficulty here is that data can be serially independent in only one way but serially dependent in many different ways. A very preliminary guess is that in the presence of temporal correlations, the difference between empirical and true decomposition terms (provided that the empirical score still makes sense at all) does not scale with $\frac{1}{N}$, as it is now, but rather inversely proportional to the effective number of independent instances, which is less than N.

As a second point, the presented results call for extension to situations in which the observation Y occupies a continuous range (such as the real numbers). There is a fairly advanced theory for evaluating such forecasts, including scoring rules, and even a decomposition of the true score into uncertainty, reliability, and resolution terms (Matheson and Winkler, 1976; Hersbach, 2000; Gneiting and Raftery, 2007; Gneiting et al, 2007). The problem is that there is as yet no equivalence to the empirical decomposition, or in other words, estimators for the empirical decomposition terms have yet to be proposed. The only option available so far is to project the observations onto a finite set of exclusive alternatives and apply the methodology discussed in this paper.

Acknowledgements

Stimulating discussions with the members of the Time Series Analysis group at the Max Planck Institute for the Physics of Complex Systems, in particular Holger Kantz and Stefan Siegert, are gratefully acknowledged. In addition, a number of thoughtful comments by two anonymous referees helped to improve this work.

A Proof of Equations (17–19)

To demonstrate the relations (17–19), we use that

$$\mathbb{E}\,\mathsf{e}(\bar{o}) = \mathbb{E}\,\mathsf{e}(\bar{\pi}) - \mathbb{E}\,\mathsf{d}(\bar{\pi}, \bar{o}) \tag{30}$$

$$\mathbb{E}(\mathsf{e}(o^{(d)})|r_d) = \mathsf{e}(\pi^{(d)}) - \mathbb{E}(\mathsf{d}(\pi^{(d)}, o^{(d)})|r_d). \tag{31}$$

For a proof of these facts, the reader is referred to Bröcker (2011). Now using the expression for d in case of the Ignorance score, we obtain

$$\mathsf{d}(\bar{\pi}, \bar{o}) = \sum_{k=1}^{K} -\log\left(\frac{\bar{\pi}_k}{\bar{o}_k}\right)\bar{o}_k. \tag{32}$$

Writing $\frac{\bar{\pi}}{\bar{o}} = 1 + x$ and assuming that x is small compared to 1, the expression (32) can be expanded to second order in x. With this approximation, we obtain by taking the expectation

$$\mathbb{E}\,\mathsf{d}(\bar{\pi}, \bar{o}) = \frac{K-1}{2N},$$

which, with the help of Equation (30), gives Equation (17).

To prove Equation (18), we use again the expression for d to get

$$\sum_d \mathsf{d}(\pi^{(d)}, o^{(d)})r_d = \sum_{d=1}^{D}\sum_{k=1}^{K} -\log\left(\frac{\pi_k^{(d)}}{o_k^{(d)}}\right) o_k^{(d)} r_d$$

We now use the same trick and write $\frac{\pi_k^{(d)}}{o_k^{(d)}} = 1 + x$. Assuming that x is small compared to 1, expanded to second order in x, and taking the expectation, we obtain

$$\mathbb{E}(\sum_d \mathsf{e}(o^{(d)})r_d) = \mathbb{E}(\mathsf{e}(\pi)) - \frac{D(K-1)}{2N}$$

with the help of Equation (31). We now subtract this equation from Equation (17) and use Equations (8) and (14) to establish Equation (18). Finally, Equation (19) follows because (17)−(18)+(19) must give Equation (15).

B Proof of Equations (23–25)

To prove these relations, we use properties of the multinomial distribution to conclude that

$$\mathbb{E}(\bar{o}_k^2) = \mathbb{E}\left(\frac{N_k}{N}\right)^2 = \frac{\bar{\pi}_k(1-\bar{\pi}_k)}{N} + \bar{\pi}_k^2.$$

With the expression for the entropy of the Brier score, this gives Equation (23) Along very similar lines, we obtain

$$\mathbb{E}(\mathsf{e}(o^{(d)})|r_d) = \frac{N_{\bullet d}-1}{N_{\bullet d}}\mathsf{e}(\pi^{(d)});$$

Strictly speaking, both sides are undefined whenever $N_{\bullet d} = 0$ or equivalently $r_d = 0$, which happens with non–vanishing probability $(1-\rho_d)^N$. This problem disappears now, as we multiply with r_d and agree that both sides are zero whenever $r_d = 0$. We obtain

$$\mathbb{E}(\mathsf{e}(o^{(d)})|r_d)r_d = \mathsf{e}(\pi^{(d)})r_d + \frac{I_d}{N}\mathsf{e}(\pi^{(d)}),$$

where $I_d = 1$ if $r_d > 0$ and 0 otherwise. We now take the expectation on both sides and obtain

$$\mathbb{E}(\sum_d \mathsf{e}(o^{(d)})r_d) = \mathbb{E}\,\mathsf{e}(\pi) - \frac{1}{N}\sum_d \nu_d \mathsf{e}(\pi^{(d)}), \tag{33}$$

where $\nu_d = \mathbb{P}(r_d > 0) = 1-(1-\rho_d)^N \cong 1-e^{-N\rho_d}$. As with the Ignorance score, we now subtract Equation (33) from Equation (23) and use Equations (8) and (14) to get Equation (24) for the resolution term. To finish the proof, Equation (25) follows because (23)−(24)+(25) must give Equation (15).

References

Brier GW (1950) Verification of forecasts expressed in terms of probabilities. Monthly Weather Review 78(1):1–3

Bröcker J (2008a) On reliability analysis of multi-categorical forecasts. Nonlinear Processes in Geophysics 15(4):661–673, URL http://www.nonlin-processes-geophys.net/15/661/2008/

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

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

Bröcker J (2011) Estimating reliability and resolution of probability forecasts using proper scoring rules (in preparation)

Bröcker J, Smith LA (2007) Scoring probabilistic forecasts: The importance of being proper. Weather and Forecasting 22(2):382–388

Brown TA (1970) Probabilistic forecasts and reproducing scoring systems. Tech. Rep. RM–6299–ARPA, RAND Corporation, Santa Monica, CA

Egan JP (1975) Signal detection theory and ROC analysis, 1st edn. Academic Press series in cognition and perception, Academic Press

Gneiting T, Raftery A (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102:359–378

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69(2):243–268, DOI 10.1111/j.1467-9868.2007.00587.x, URL http://dx.doi.org/10.1111/j.1467-9868.2007.00587.x

Hamill TM (2001) Interpretation of rank histograms for verifying ensemble forecasts. Monthly Weather Review 129(3):550–560

Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting 15(5):559–570

Matheson JE, Winkler RL (1976) Scoring rules for continuous probability distributions. Management Science 22(10):1087–1096

Murphy AH (1973) A new vector partition of the probability score. Journal of Applied Meteorology 12(4):595–600, DOI 10.1175/1520-0450(1973)

Murphy AH (1993) What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather and Forecasting 8(2):281–293

Murphy AH (1996) General decompositions of MSE-based skill scores: Measures of some basic aspects of forecast quality. Monthly Weather Review 124(10):2353–2369

Murphy AH, Winkler RL (1987) A general framework for forecast verification. Monthly Weather Review 115:1330–1338

Persson A, Grazzini F (2005) User guide to ECMWF forecast products. Tech. rep., European Centre for Medium Range Weather Forecasts, URL http://www.ecmwf.int/products/forecasts/guide/user_guide.pdf

Savage LJ (1971) Elicitation of personal probabilities and expectation. Journal of the American Statistical Association 66(336):783–801

Stephenson DB, Coelho CAS, Jolliffe IT (2008) Two extra components in the brier score decomposition. Weather and Forecasting 23(4):752–757, DOI 10.1175/2007WAF2006116.1, URL http://journals.ametsoc.org/doi/abs/10.1175/2007WAF2006116.1, http://journals.ametsoc.org/doi/pdf/10.1175/2007WAF2006116.1

List of figures

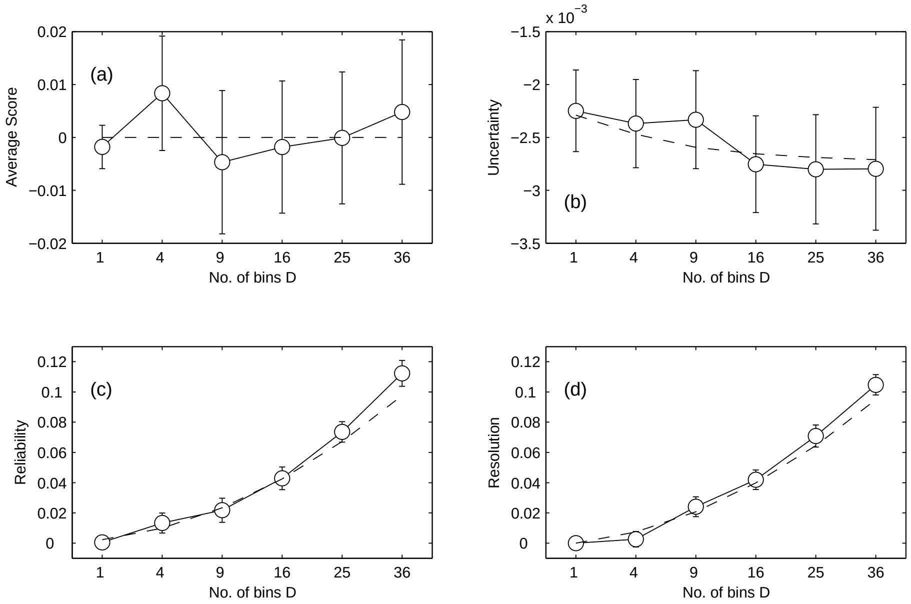

Figure 1: Results of both theory (dashed lines) as well as experiment (balls) are shown for the Ignorance score. The different plots refer to the average score (a), the uncertainty term (b), the reliability term (c), and the resolution term (d). For all quantities, the difference between the empirical and the true terms are shown, divided by the true score. All quantities are plot as a function of D, the number of bins. Bars indicate $\pm 2\sigma$ confidence intervals. Our theoretical investigations are confirmed. The empirical and true score agree on average; Empirical uncertainty underestimates true uncertainty; Empirical reliability and resolution terms overestimate true reliability and resolution terms.

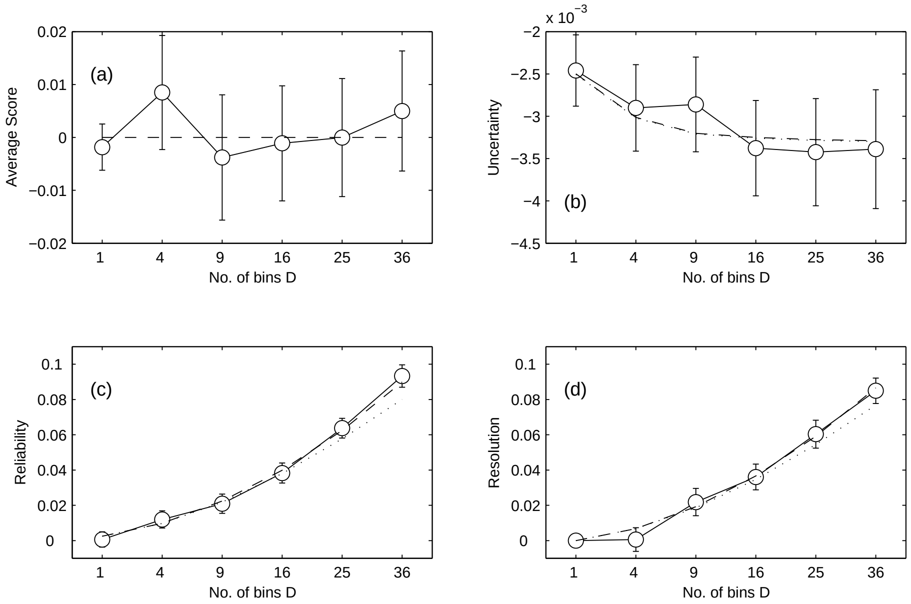

Figure 2: Results of both theory (dashed lines) as well as experiment (balls) are shown for the Brier score. The different plots refer to the average score (a), the uncertainty term (b), the reliability term (c), and the resolution term (d). For all quantities, the difference between the empirical and the true terms are shown, divided by the true score. All quantities are plot as a function of D, the number of bins. Bars indicate $\pm 2\sigma$ confidence intervals. Our theoretical investigations are again confirmed, as in the Ignorance case. In addition, estimates of the theoretical differences are shown with dotted lines. These estimates roughly agree with the exact values.

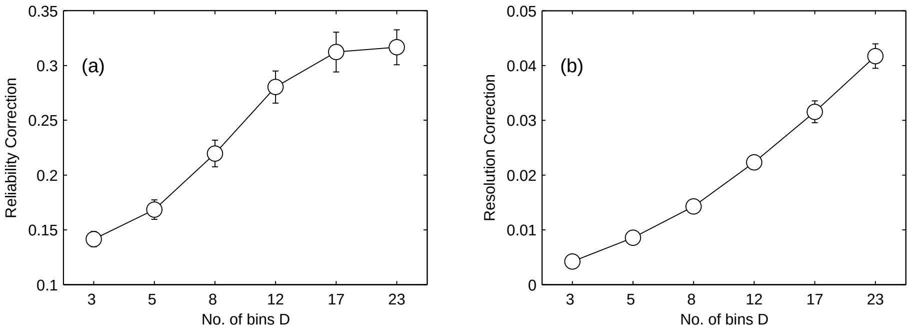

Figure 3: Plot (a): correction to the reliability term, relative to the empirical reliability term itself, as a function of D. Plot (b): correction to the resolution, relative to the empirical resolution itself, as a function of D, the number of bins. All quantities were computed using $N = 1095$ instances which were drawn randomly and without replacement from the entire data set of 1812 samples. This resampling experiment was repeated 100 times, with error bars representing two standard deviations. The resampled data sets comprised only three years.

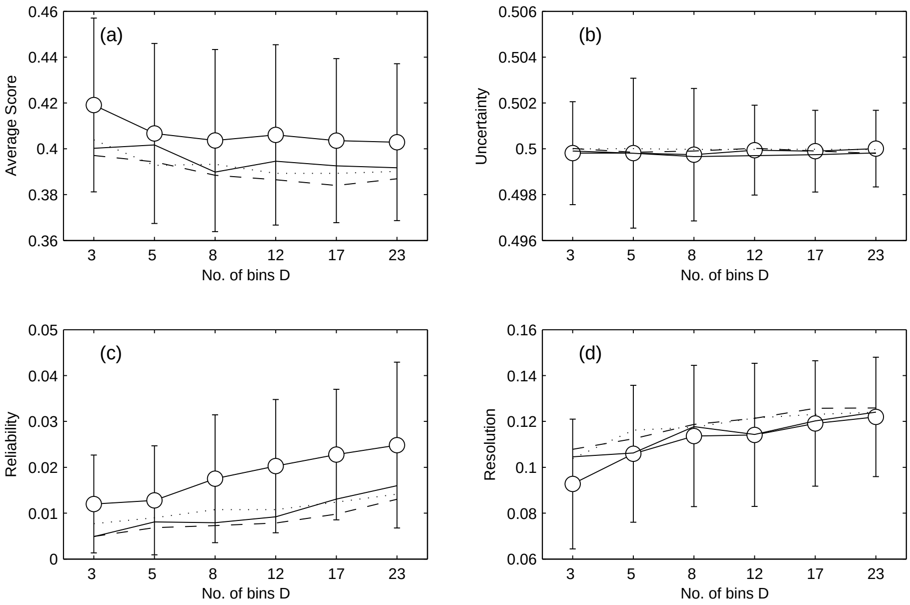

Figure 4: Plot (a) shows the empirical score; the estimates of the true uncertainty, resolution, and reliability terms (as suggested by Eqs. 27–29) are shown in plot (b–d), respectively. All quantities are plot as a function of D, the number of bins. Estimates for $N = a \cdot 365$ with $a = 2 \dots 5$ years were computed. Results for two years are presented with error bars (two standard deviations obtained through resampling the entire data set). The other lines correspond to $a = 3$ years (solid), $a = 4$ years (dashed), and $a = 5$ years (dotted). Clearly, the results are not independent of N, but the differences are on the same order of magnitude as the sampling variations represented by the error bars.