Reliability, Sufficiency, and the Decomposition of Proper Scores*

Jochen Bröcker†

October 23, 2009

Abstract

Scoring rules are an important tool for evaluating the performance of probabilistic forecasting schemes. A scoring rule is called strictly proper if its expectation is optimal if and only if the forecast probability represents the true distribution of the target. In the binary case, strictly proper scoring rules allow for a decomposition into terms related to the resolution and to the reliability of the forecast. This fact is particularly well known for the Brier Score. In this paper, this result is extended to forecasts for finite–valued targets. Both resolution and reliability are shown to have a positive effect on the score. It is demonstrated that resolution and reliability are directly related to forecast attributes which are desirable on grounds independent of the notion of scores. This finding can be considered an epistemological justification of measuring forecast quality by proper scores. A link is provided to the original work of DeGroot and Fienberg (1982), extending their concepts of sufficiency and refinement. The relation to the conjectured sharpness principle of Gneiting et al. (2005a) is elucidated.

1 Introduction

Brown (1970) argues that it seems reasonable to value forecasts (be they probabilistic or other) by a scheme related to the extent to which the forecasts "come true". Scoring rules provide examples for such schemes in the case of probabilistic forecasts. After pioneering work by Good (1952) and Brier (1950), scores were thoroughly investigated in the 1960's and 1970's. The score was effectively thought of as a reward system, inducing (human) experts to provide their judgments or predictions regarding uncertain events in terms of probabilities (Brown, 1970; Savage, 1971). In this respect, scoring rules were devices to elicit probabilities from humans.

The importance of using proper scores (see Section 2 for a definition) was recognised already by Brier (1950) (see also Brown, 1970, for an entertaining discussion and "some horrible examples"). The central argument is that a forecaster's probability assignment should be independent of the particular reward system, which is guaranteed if the reward system constitutes a proper

* *The Quarterly Journal of the Royal Meteorological Society*, Vol.135, No.643, Pg.1512-1519

†Max–Planck–Institut für Physik komplexer Systeme, Nöthnitzer Strasse 34, 01187 Dresden, Germany. email: broecker@pks.mpg.de

score. Savage (1971) (following de Finetti, 1970) points out that this universality property allows for an alternative definition of subjective probability, which is a concept of probability independent of the notion of relative observed frequency.

Over the last decades, computer power has increased enormously. Thus it became feasible to produce probabilistic forecasts for dynamical processes numerically, employing models of ever increasing complexity. Since it is obviously irrelevant whether probabilities are produced by humans or machines, scores provide a tool to evaluate probabilistic numerical forecasting systems, too. In weather forecasting, scores had already been used to evaluate subjective forecasts (issued by expert meteorologists), for example of rain, long before numerical weather forecasts became available (Brier, 1950; Winkler and Murphy, 1968; Epstein, 1969; Murphy and Winkler, 1977). Nowadays, scores are widely applied also in the evaluation of numerically generated probabilistic weather forecasts (Gneiting et al., 2005b; Gneiting and Raftery, 2007; Bröcker et al., 2004; Raftery et al., 2005; Wilks, 2006a; Bröcker and Smith, 2008).

In contrast to the expert–judgment–forecasts considered in earlier works on scores, weather forecasts are often issued over a long period of time under (more or less) stationary conditions, allowing archives of forecast–observation pairs to be collected. Such archives allow to calculate observed frequencies and to compare them with forecast probabilities. For example, if we were to forecast the probability of rain on a large number of occasions, we would like rain to occur on a fraction p of those instances where our forecast was (exactly or around) p. A forecast having this property (up to statistical fluctuations) is called *reliable* (Murphy and Winkler, 1977; Toth et al., 2003; Wilks, 2006b). If a large archive of forecast–observation pairs is available, reliability becomes a sensible property to ask for. As has been widely noted previously though, it is not difficult to produce reliable forecasts if no constraint is put on the information content or resolution of the forecast (the exact meaning of these terms is often left vague, though). In any event, the grand probability (aka climatological frequency) of the target will always be a reliable forecast, and despite the difficulties with the term "information content", many people would presumably agree that this forecast is not very informative.

But how do these virtuous forecast attributes pertain to proper scores? Do proper scores reward reliable forecasts? Does a "better informed" forecaster really achieve a better score? In this paper, these questions are answered in the affirmative (using the appropriate formalisation of "better informed"). In Section 2, after recalling the notion of reliability, it is shown that proper scores allow for a decomposition into terms measuring the resolution and the reliability of the forecast. In particular, reliability turns out to have a direct positive impact on the score. In Section 3, the concept of sufficiency is introduced, generalising similar notions of DeGroot and Fienberg (1982). Sufficiency formalises the idea of "being more informed", and is shown to have a direct positive impact on the resolution term of the score.

The decomposition of Section 2 is well known for the Brier score (see for example Murphy and Winkler, 1987; Murphy, 1996; Blattenberger and Lad, 1985), a widely used score for forecasting problems with only two categories. The Brier score presumably owes much of its popularity to this decomposition, rendering its interpretation very clear. DeGroot and Fienberg (1982) have derived a similar decomposition for any proper score in the case of binary targets. The relation to the conjectured sharpness principle of Gneiting et al. (2005a) is

elucidated. The appendix contains several more technical points. Appendix A briefly revisits the conditional expectation, a concept that is made extensive use of. Appendix B provides an equivalent characterisation of reliability. In Appendix C, the equivalence between sufficiency according to DeGroot and Fienberg (1982) and as used in this paper is shown. Finally, the derivation of the decomposition (15) is presented in Appendix D.

2 A general decomposition

In this section, a general decomposition of proper scores will be derived. To facilitate the discussion, some convenient notation will be introduced first, supplemented with a brief reminder on proper scores. Let Y denote the quantity to be forecast, commonly referred to as the *observation* or *target*.[1] The observation Y is modelled here as a random variable taking values in a set $\mathcal{I}$. For the sake of simplicity, $\mathcal{I}$ is assumed to be a finite set of alternatives (e.g. "rain/hail/snow/sunshine"), labelled $1 \dots K$. Values of Y (i.e. elements of $\mathcal{I}$) will be denoted by small lowercase letters like k or l.

A *probability assignment* over $\mathcal{I}$ is a K–dimensional vector p with nonnegative entries so that $\sum_{k\in\mathcal{I}} p_k = 1$. The set of all probability assignments over $\mathcal{I}$ is denoted by $\mathcal{P}_\mathcal{I}$. Elements of $\mathcal{P}_\mathcal{I}$ will be denoted by p, q, and r. A *probabilistic forecasting scheme* is a random variable γ with values in $\mathcal{P}_\mathcal{I}$. In other words, the realisations of γ are probability assignments over $\mathcal{I}$. 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 weather information down to 48h prior to when the observation Y is obtained. Designing a forecasting scheme means to model the relationship between this side information and the variable to be forecast (see Murphy and Winkler, 1987; Murphy, 1993, 1996, for a related discussion).[2]

It was already mentioned what reliability means in case that $\mathcal{I}$ contains only two elements (1 and 0, say). In the case of more than two alternatives, this definition of reliability generalises as follows: On the condition that the forecasting scheme is equal to, say, the probability assignment p, the observation Y should be distributed according to p, or in formulae

$$\mathbb{P}\left(Y = k|\gamma = p\right) = p_k \tag{1}$$

for all $k \in \mathcal{I}$. In particular, a reliable forecasting scheme can be written as a conditional probability. As is demonstrated in Appendix B, the reverse is also true: every conditional probability of Y is reliable. In view of Equation (1), we will fix the notation $\pi_k^{(\gamma)} := \mathbb{P}\left(Y = k|\gamma\right)$, $k = 1 \dots K$ for the conditional probability of the observation given the forecasting scheme. Like every conditional probability, $\pi^{(\gamma)}$ is a random quantity. Hence, $\pi^{(\gamma)}$ is a probabilistic forecasting scheme like γ itself. In terms of $\pi^{(\gamma)}$ and γ, the reliability condition (1) can be written simply as $\pi^{(\gamma)} = \gamma$. Since $\pi^{(\gamma)}$ is reliable, it trivially holds that $\pi^{(\gamma)} = \pi^{\left(\pi^{(\gamma)}\right)}$. In any case, $\pi^{(\gamma)}$ is a function of γ, independent of whether γ is reliable or not.

[1] Italics indicate that an expression is to be considered a technical term.

[2] We will not consider forecasting problems which are explicitly dependent on time, as would be necessary for example to take into account seasonal effects.

A *scoring rule* (see for example Matheson and Winkler, 1976; Gneiting and Raftery, 2007) is a function $S(p, k)$ which takes a probability assignment over $\mathcal{I}$ as its first argument and an element of $\mathcal{I}$ as its second argument. For any two probability assignments p and q, the *scoring function* is defined as

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

The interpretation of the scoring function is that if Z is a random variable of distribution q, then $\mathsf{s}(p, q)$ is the mathematical expectation of the score of the assignment p in forecasting Z. It is our convention that a small score indicates a good forecast. 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 $\mathsf{d}(p, q) = 0$ implies $p = q$. The interpretation of $\mathsf{d}(p, q)$ as a divergence is obviously meaningful only if the scoring rule is strictly proper. From now on, scoring rules are assumed to be strictly proper. It is important to note that $\mathsf{d}(p, q)$ is, in general, not a metric, as it is neither symmetric nor does it fulfil the triangle inequality. The quantity

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

is called the *entropy* of p.[3] Table 1 gives a couple of frequently used scoring rules along with the corresponding divergences and entropies.

`Table 1 on top of this or the next page`

For strictly proper scores,

$$\mathsf{e}(p) = \inf_q \mathsf{s}(q, p). \tag{5}$$

Since $\mathsf{s}(q, p)$ is linear in p, Equation (5) demonstrates that for strictly proper scores, the entropy $\mathsf{e}(p)$ is an infimum over linear functions and hence concave (Rockafellar, 1970). For the particular cases listed in Table 1, it should be fairly obvious that the entropy is a measure for the uncertainty inherent in a probability assignment p. For the Brier score and the Ignorance, the entropy is indeed a very common measure of inherent randomness of a distribution. Furthermore, suppose p and q are two probability assignments featuring the same entropy, then intuitively, any mixture of p and q should have a larger inherent uncertainty than any of the individual probability assignments, an intuition which the entropy supports, due to the concavity of $\mathsf{e}(p)$.

The aim now is to derive a decomposition of the expectation $\mathbb{E}\left[S(\gamma, Y)\right]$ of the score achieved by the forecasting scheme γ. Since γ is random, the expectation affects both γ and Y. In this paper, extensive use will be made of the conditional expectation. A few words about this concept, along with the most important properties, can be found in Appendix A. An elementary property of the mathematical expectation gives

$$\mathbb{E}\left[S(\gamma, Y)\right] = \mathbb{E}\left[\mathbb{E}\left[S(\gamma, Y)|\gamma\right]\right]. \tag{6}$$

[3] Gneiting and Raftery (2007) refer to $-\mathsf{e}(p)$ as either the generalised entropy function or the information measure, but since entropy is commonly interpreted as a *lack* of information, the entropy is defined here as $\mathsf{e}(p)$.

Name	scoring rule $S(p,k)$	divergence $\mathsf{d}(q,p)$	entropy $\mathsf{e}(p)$
Brier[a]	$\lvert y-p\rvert^2$	$\lvert p-q\rvert^2$	$p(1-p)$
Ignorance[b]	$-\log p_k$	$\sum_l -\log\left(\frac{p_l}{q_l}\right) q_l$	$\sum_l -\log(p_l)p_l$
CRPS[c]	$\int (F(z)-H(k-z))^2 \mathrm{d}z$	$\int (F(z)-G(z))^2 \mathrm{d}z$	$\int F(z)(1-F(z))\mathrm{d}z$
PSS[d]	$-\frac{p_k^{\alpha-1}}{\lVert p\rVert_\alpha^{\alpha-1}}$	$\lVert q\rVert_\alpha - \frac{\langle q, p^{\alpha-1}\rangle}{\lVert p\rVert_\alpha^{\alpha-1}}$	$-\lVert p\rVert_\alpha$
PLS[e]	$\sum_l p_l^2 - 2p_k$	$\sum (p_l-q_l)^2$	$-\sum p_l^2$

[a]For binary cases (i.e. $\mathcal{I}=\{0,1\}$).

[b]Propriety follows from Jensen's inequality.

[c]Continuous Ranked Probability Score – Here F and G are the cumulative distribution functions corresponding to p and q, respectively.

[d]Pseudo-spherical Scores – Here $\alpha>1$, while $\lVert p\rVert_\alpha = \left[\sum_l p_l^\alpha\right]^{1/\alpha}$. Propriety follows from Hölder's Inequality.

[e]Proper Linear Score, also referred to as the quadratic score. For binary cases (i.e. $\mathcal{I}=\{0,1\}$), this score is equivalent to the Brier score

Table 1: Scoring rule, divergence, and entropy for several common scores. All sums extend over $\mathcal{I}$. See Epstein (1969); Murphy (1971) for a discussion of the Ranked Probability Score. Matheson and Winkler (1976); Gneiting and Raftery (2007) discuss scoring rules for continuous variables.

To calculate the conditional expectation $\mathbb{E}\left[S(\gamma,Y)|\gamma\right]$, the probability of Y given γ is needed, but this is just $\pi^{(\gamma)}$, whence

$$\mathbb{E}\left[S(\gamma,Y)|\gamma\right] = \mathsf{s}(\gamma,\pi^{(\gamma)}). \tag{7}$$

Substituting with Equation (7) in (6) results in

$$\mathbb{E}\left[S(\gamma,Y)\right] = \mathbb{E}\,\mathsf{s}(\gamma,\pi^{(\gamma)}). \tag{8}$$

From Equations (3) and (4) we get

$$\mathsf{s}(\gamma,\pi^{(\gamma)}) = \mathsf{e}(\pi^{(\gamma)}) + \mathsf{d}(\gamma,\pi^{(\gamma)}). \tag{9}$$

Taking the expectation on both sides of Equation (9) and substituting for the right hand side in (8), we obtain

$$\mathbb{E}\left[S(\gamma,Y)\right] = \mathbb{E}\,\mathsf{e}(\pi^{(\gamma)}) + \mathbb{E}\,\mathsf{d}(\gamma,\pi^{(\gamma)}). \tag{10}$$

The first term in Equation (10), the expectation of the entropy of $\pi^{(\gamma)}$, can be decomposed further. Consider the (nonrandom) forecasting scheme obtained by taking the expectation of $\pi^{(\gamma)}$,

$$\bar{\pi} := \mathbb{E}\,\pi^{(\gamma)} \tag{11}$$

It is easily seen that $\bar{\pi}$ is just the unconditional probability of Y, which in meteorology is often referred to as the *climatology* of Y. Since $\mathsf{s}(\bar{\pi},\pi^{(\gamma)})$ is linear in $\pi^{(\gamma)}$ and $\bar{\pi}$ is not random, it follows immediately from Equation (11) that

$$\mathbb{E}\,\mathsf{s}(\bar{\pi},\pi^{(\gamma)}) = \mathsf{s}(\bar{\pi},\bar{\pi}) = \mathsf{e}(\bar{\pi}). \tag{12}$$

(In fact, the relation $\mathbb{E}\,\mathsf{s}(\bar{\pi}, \gamma) = \mathsf{e}(\bar{\pi})$ is true for any reliable forecasting scheme γ.) Adding and subtracting $\mathbb{E}\,\mathsf{s}(\bar{\pi}, \pi^{(\gamma)})$ on the right hand side of Equation (10) and using Equation (12) we arrive at

$$\mathbb{E}\,[S(\gamma, Y)] = \mathsf{e}(\bar{\pi}) - \mathbb{E}\,\mathsf{d}(\bar{\pi}, \pi^{(\gamma)}) + \mathbb{E}\,\mathsf{d}(\gamma, \pi^{(\gamma)}). \tag{13}$$

Equation (13) constitutes the desired decomposition of the expectation of $S(\gamma, Y)$. This decomposition is completely analogous to and a generalisation of the well known decomposition of the Brier score. The three terms in Equation (13) will be (from left to right) referred to as the uncertainty of Y, the resolution term[4], and the reliability term. Note that for the Brier score, Equation (13) indeed yields the known decomposition.

The remainder of this section will provide an intuitive interpretation of each term in Equation (13). Firstly, the uncertainty of Y is the entropy of the climatology, which can be seen as the expectation value of the score of the climatology as a forecasting scheme. In other words, it quantifies the ability of the climatology to forecast random draws from itself.

Secondly, note that the resolution term $\mathbb{E}\,\mathsf{d}(\bar{\pi}, \pi^{(\gamma)})$ is always positive definite, due to the strict propriety of the score. The resolution term contributes negatively to the score. Since the resolution term describes, roughly speaking, the deviation of $\pi^{(\gamma)}$ from its expectation value $\bar{\pi}$ (see Equation 11), it can be interpreted as a form of "variance" of $\pi^{(\gamma)}$. The resolution term is indeed given by the standard variance of $\pi^{(\gamma)}$ in case of the Brier score.

It might seem counterintuitive that the larger this "variance", the better the score. Consider an event which happens with 50% chance. Then $p = 0.5$ is a reliable forecast which has zero resolution with respect to any score. Consider another forecast which says either $p = 0.1$ or $p = 0.9$, and which is also reliable. (It follows that the outcomes $p = 0.1$ and $p = 0.9$ must occur with frequency 0.5) This forecast is clearly more useful than the former; if it says "0.1", we know for sure that the event is unlikely, while "0.9" is a reliable indicator of a likely event. The fact that the second forecast is more informative is actually the reason for its larger variability.

Finally, the reliability term (which is again positive definite) describes the deviation of γ from $\pi^{(\gamma)}$. Recalling that $\gamma = \pi^{(\gamma)}$ indicates a reliable forecast, the interpretation of the reliability term as the average violation of reliability becomes obvious.

3 A decomposition of the resolution term

The decomposition (13) demonstrates how the score changes if the forecast scheme γ changes in such a way that $\pi^{(\gamma)}$ remains constant. In this case, any deviation of γ from $\pi^{(\gamma)}$ has adverse effects on the score. But in general, changing γ means that $\pi^{(\gamma)}$ changes, too. Thus, changes in γ usually entail changes in both the reliability and the resolution term of the decomposition (13). The changes in the resolution term are investigated in this section. It will turn out that, roughly speaking, the resolution term quantifies the information content of the forecast scheme.

The concept of forecast sufficiency, introduced by DeGroot and Fienberg (1982), formalises the notion of being "more or less informed" and allows for

[4] Also called sharpness term

the partial ordering of forecasting schemes. As will be seen in this section, γ_1 will have at least the same resolution as γ_2 if γ_1 is sufficient for γ_2. Thus, the expectation value of the score reproduces the same ordering as sufficiency. This result establishes a connection between a quantitative notion of information as provided by the score, and a qualitative notion of information content as provided by sufficiency. This is analogous to the relation between the reliability term of the decomposition (13) and the qualitative reliability condition (1).

A forecasting scheme γ_1 is called *sufficient* for a forecasting scheme γ_2 if

$$\pi^{(2)} = \mathbb{E}\left[\pi^{(1)}|\gamma_2\right], \tag{14}$$

where the abbreviations $\pi^{(1)} := \pi^{(\gamma_1)} = \mathbb{P}(Y|\gamma_1)$ and analogously for $\pi^{(2)}$ were used.[5] In Appendix C, it is shown that the present notion of sufficiency is equivalent to the corresponding definition of DeGroot and Fienberg (1982).

Before continuing with score decompositions, the rather technical condition (14) is given a somewhat informal interpretation. Suppose the forecaster who is running forecasting scheme γ_1, albeit having no access to the current value of γ_2, collected a large archive of past values of γ_2 and hence is able to fit a good approximation to $\mathbb{P}(\gamma_2|\gamma_1)$. With this information, the forecaster tries to mimic forecasting scheme γ_2 as follows. The forecaster's mimicry version of γ_2 (which is denoted by γ_2^*) is just a random draw of $\mathbb{P}(\gamma_2|\gamma_1)$ (conditioned on the forecaster's own forecast γ_1). Since the expectation value of the score depends only on the joint distribution of the forecast scheme and Y, the mimicry forecast γ_2^* will achieve the same score as the real γ_2 (in expectation) if the compound distribution of (γ_2, Y) and (γ_2^*, Y) are the same. It is straight forward to work out that the latter condition is equivalent to (14). In brief, if γ_1 is sufficient for γ_2, then by appropriate randomisation of γ_1, a forecast γ_2^* is obtained which has the same statistical properties as γ_2. Note also that in particular γ_1 is sufficient for γ_2 if γ_2 can be written as a function of γ_1.

In Appendix D, it is shown that if γ_1 is sufficient for γ_2, it holds that

$$\mathbb{E}\,\mathsf{d}(\bar{\pi}, \pi^{(2)}) = \mathbb{E}\,\mathsf{d}(\bar{\pi}, \pi^{(1)}) - \mathbb{E}\,\mathsf{d}(\pi^{(2)}, \pi^{(1)}). \tag{15}$$

Keeping in mind that $\mathbb{E}\,\mathsf{d}(\bar{\pi}, \pi^{(1)})$ and $\mathbb{E}\,\mathsf{d}(\bar{\pi}, \pi^{(2)})$ are the resolution terms of γ_1 and γ_2, respectively, and that $\mathsf{d}(\ldots)$ is never negative, Equation (15) demonstrates that the resolution of γ_2 will be at most that of γ_1.

To summarise, Equations (13) and (15) together allow for the following conclusions as to the approach of scoring forecasting schemes using strictly proper scores:

- On average (that is, in terms of the expectation value), the forecasting scheme $\pi^{(\gamma)}$ achieves the best possible score among all forecasts for which γ is sufficient. If the score is strictly proper, $\pi^{(\gamma)}$ is uniquely defined through this optimum property, in the sense that any forecast for which γ is sufficient is either equal to $\pi^{(\gamma)}$ or it will have a worse score. This can be considered as a proof of the conjectured sharpness principle of Gneiting et al. (2005a), reinterpreted in our framework.

[5] If both γ_1 and γ_2 are reliable, then condition (14) modifies to $\gamma^2 = \mathbb{E}\left[\gamma^1|\gamma^2\right]$. In this situation, γ^1 is said to be *at least as refined as* γ^2.

- Per se, it is impossible to say how the score will rank unreliable forecast schemes, even if one is sufficient for the other. The lack of reliability of one forecast scheme might be out-balanced by the lack of resolution of the other.
- It is also not clear how the score will rank forecast schemes (reliable or unreliable) as long as none of the two forecast schemes is sufficient for the other. It seems plausible that the actual ranking of such forecasts will depend on the particular scoring rule employed.

4 An Example

In this section, an example is discussed, in order to illustrate the concepts of this paper. Assume a weather forecasting centre issues ensemble forecasts on a daily basis. To be specific, we look at forecasts for the two metre temperature at Heligoland weather station (WMO 10015), measured at noon. The observations are distributed among three categories, referred to as "warm", "moderate", and "cold", defined as follows. From the actual observations, a reference temperature is computed as a fourth order trigonometric polynomial of time. This reference temperature is subtracted from both forecasts and observations. An observation is classified as "warm", "moderate", or "cold", respectively, if the observation (anomaly) is above $1.21°$, between $1.21°$ and $-1.03°$, or below $-1.03°$, respectively. These categories were chosen so as to have a climatological probability of $1/3$.

Forecaster Alice devises a forecast scheme γ_A by assigning to each category the relative number of ensemble members which fall into that category. Assuming an ensemble of 15 members, there are 136 different possible values of γ_A in total, that is, 136 triplets of probabilities (p_1, p_2, p_3) assigned to the three possible events. In fact, it turns out that 4 of them never get issued.

Forecaster Bob devises a forecast scheme γ_B which works very much like Alice's, only that Bob is given only 5 ensemble members. They are a random subselection of the ensemble members that Alice is using. Otherwise, Bob follows Alice's procedure, and consequently his forecast comprises 21 potential values.

Forecaster Charleen uses a simplified form γ_C of Bob's forecasting scheme. She assings probability one to the category that is most likely according to Bob's forecast and zero to the other two. Hence, Charleen's forecast scheme can assume only 3 possible values.

Firstly, let us consider the reliability of these forecasts. In the present context, reliability means the following. Fix a triplet (p_1, p_2, p_3) among the possible values of, say, Alice's forecast scheme, and consider only days where her forecast equals (p_1, p_2, p_3). Then the observed frequencies of "warm", "moderate", and "cold" over this restricted set of days should, in the long run, converge to p_1, p_2, and p_3, respectively. Note that in reality, a forecasting scheme can only proven to be reliable in the sense of a statistical test, not in the sense of a mathematical equality. In any event, the long time observed frequencies, seen as a function of (p_1, p_2, p_3), yield $\pi^{(\gamma_A)}$, abbreviated as $\pi^{(A)}$. For Bob (resp. Charleen), $\pi^{(B)}$ (resp. $\pi^{(C)}$) are defined similarly.

Contrary to popular belief, Alice's and Bob's counting approach leads to a

Name	Av. Score	Resolution	Reliability
Alice:	-0.418 ± 0.011	0.1594 ± 0.0032	0.0744 ± 0.0028
Bob:	-0.351 ± 0.014	0.0956 ± 0.0024	0.0784 ± 0.0019
Charleen:	0.616 ± 0.019	0.0456 ± 0.0001	0.9949 ± 0.0023

Table 2: The score (i.e. its expectation value), resolution, and reliability, estimated for three forecasting schemes. The proper linear score was used. The data set comprised 1810 forecast–observation pairs for two–metre temperature anomalies at WMO 10015.

reliable forecast only in the limit of infinitely many ensemble members, even if the original ensemble is reliable. Since Charleen's ensemble is also clearly not reliable, all three forecast schemes can be expected to have a nonzero reliability term. To estimate the reliability and resolution terms, a leave–one–out approach was employed (Bröcker, 2008), in order to avoid overly biased results. The entire data set comprised 1810 instances.

`table 2 on top of this or the next page`

Results using the proper linear score are shown in Table 2. By choice of the categories, the climatology for each category is $1/3$, hence the uncertainty has a value of $\mathsf{e}(\bar{\pi}) = -1/3$. With this in mind, the results in Table 2 are seen to obey the decomposition (13), within the confidence limits. From the construction of the forecasts, it is intuitively clear that Alice posesses the largest amount of information, and reassuringly, Table 2 demonstrates this. Alice's forecast features the largest resolution, followed by Bob and finally Charleen. This is consistent with the theoretical results in this paper, since Alice's scheme is sufficient for Bob's, while Bob's scheme is sufficient for Charleen's. The latter is true because Charleen's forecasting scheme is a function of Bob's, which implies sufficiency (see the discussion of sufficiency in Section 3). To see that Alice's forecasting scheme is sufficient for Bob's, note that Bob's forecast can be written as a function of Alice's and a random variable r, which models the random selection of 5 ensemble members out of 15 without replacement. To mimick Bob's forecast, Alice simply draws 5 times without replacement from the categories "warm", "moderate", and "cold", according to the probabilities assigned by herself, and subsequently counts the occurences of each event. Thereby, Alice creates a forecasting scheme which, although not *equal* to Bob's, has exactly the same statistical properties. More formally, we can verify condition (14). By definition $\pi_k^{(B)} = \mathbb{E}\left[\delta_{Y,k}|\gamma_B\right]$ and similarly for $\pi^{(A)}$, where $\delta_{Y,k} = 1$ if $Y = k$ and 0 otherwise. But

$$\begin{aligned}\mathbb{E}\left[\delta_{Y,k}|\gamma_B\right] &= \mathbb{E}\left[\mathbb{E}\left[\delta_{Y,k}|\gamma_A, r\right]|\gamma_B\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\delta_{Y,k}|\gamma_A\right]|\gamma_B\right] \\ &= \mathbb{E}\left[\pi_k^{(A)}|\gamma_B\right]\end{aligned}$$

which is condition (14). Here it was used that γ_B is a function of γ_A and r, and that Y is independent of r.

Using $\pi^{(A)}$, $\pi^{(B)}$, and $\pi^{(C)}$ from the numerical computations, we can verify Equation (15). For Alice and Bob, $\mathbb{E}\,\mathsf{d}(\pi^{(A)}, \pi^{(B)}) = 0.0703 \pm 0.0023$; for Bob and Charleen, $\mathbb{E}\,\mathsf{d}(\pi^{(B)}, \pi^{(C)}) = 0.05 \pm 0.0012$. These numbers are, within confidence limits, equal to the respective difference in resolution in Table 2, consistent with

Equation (15).

5 Conclusion

The score of a probabilistic forecast was shown to decompose into terms related to the uncertainty in the observation, the resolution of the forecast, and its reliability, generalising corresponding results for the Brier score. The only property required of the score is that it be strictly proper. By using a widely accepted characterisation of reliability, and furthermore by generalising the concepts of sufficiency and refinement due to DeGroot and Fienberg (1982), it was argued that both the resolution and the reliability term in the decomposition quantify forecast attributes for which the case can been made independently (i.e. not referring to scoring rules). These results provide an epistemological justification of measuring forecast quality by proper scores. Furthermore, the relation to the conjectured sharpness principle of Gneiting et al. (2005a) was mentioned.

Acknowledgements

The author gratefully acknowledges fruitful discussions with Kevin Judd, University of Western Australia, as well as the members of the Time Series Analysis group and the Max–Planck–Institute for the Physics of Complex Systems, in particular Gianluigi Del Magno and Holger Kantz.

Appendix

A Conditional expectation

Some readers might be unfamiliar with the notion of conditional expectations as used in this paper. This appendix is supposed to provide some clarification. For a mathematical treatment of the conditional expectation, the reader is referred to textbooks of stochastics and probability theory, for example Breiman (1973). Let X and Y be random variables. The conditional probability density $p(x|y)$ of X given $Y = y$ describes the distribution of values of X which are consistent with the condition $Y = y$. The conditional expectation of X given Y is defined as

$$\mathbb{E}\left[X|Y\right] = \int xp(x|Y)\mathrm{d}x.$$

The conditional expectation of X given Y is a function of the value of Y; one also defines

$$\mathbb{E}\left[X|Y = y\right] = \int xp(x|y)\mathrm{d}x$$

which is a function of the nonrandom parameter y. The conditional expectation is completely determined by the joint distribution of X and Y.

Let $f(x, y)$ be some function. Often, the notation $\mathbb{E}_X[f(X, Y)]$ is found, which is supposed to mean the expectation of X with Y being held constant. This notion becomes problematic if X and Y are dependent, since then keeping Y constant has side effects on X. Then, $\mathbb{E}_X[f(X, Y)]$ can either mean to calculate $\mathbb{E}[f(X, y)]$ as a function of the nonrandom parameter y and then setting

$y = Y$, or it can mean $\mathbb{E}[f(X,Y)|Y]$. In general, these notions coincide only if X and Y are independent. The conditional expectation can also be thought of as calculating $\mathbb{E}[f(X,y)]$ as a function of the nonrandom parameter y, but *taking into account* the side effects on the distribution of X. The most important rules for manipulations involving the conditional expectation are the following. Let X, Y, and Z be random variables. Then

1. $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$.
2. $\mathbb{E}[Y|X]$ is linear in Y.
3. If Y is a function of X, then $\mathbb{E}[YZ|X] = Y\,\mathbb{E}[Z|X]$.
4. If Y is a function of X, then $\mathbb{E}[Z|Y] = \mathbb{E}[\mathbb{E}[Z|X]|Y]$ *(law of iterated expectations)*.
5. $\mathbb{P}(Y = k|X) = \mathbb{E}[\delta_{Y,k}|Y]$, where $\delta_{Y,k} = 1$ if $Y = k$ and 0 otherwise.

B An alternative definition of reliability

In this section, it will be shown that any conditional probability is a reliable forecasting scheme. The reader is assumed to be familiar with the basic notions of probability theory (see e.g. Breiman, 1973, chapter 4). Let γ be a probabilistic forecasting scheme which can be written as a conditional probability, that is

$$\mathbb{P}(Y = k|\mathcal{F}) = \gamma_k \tag{16}$$

for all $k \in E$ and some random variable $\mathcal{F}$, modelling the information that γ is built upon[6]. On both sides of Equation (16), we take the mathematical expectation conditioned on γ. The right hand side gives back γ_k. To compute the left hand side, note that because of Equation (16), γ is a function of $\mathcal{F}$. Hence

$$\begin{aligned}\mathbb{E}[\mathbb{P}(Y = k|\mathcal{F})|\gamma] &= \mathbb{E}[\mathbb{E}[\delta_{Y,k}|\mathcal{F}]|\gamma] \\ &= \mathbb{E}[\delta_{Y,k}|\gamma] \\ &= \mathbb{P}(Y = k|\gamma),\end{aligned} \tag{17}$$

where $\delta_{Y,k} = 1$ if $Y = k$ and 0 otherwise. This demonstrates that $\mathbb{P}(Y = k|\gamma) = \gamma_k$, which is the condition for reliability.

C Sufficiency and refinement of DeGroot and Fienberg

Let γ_1, γ_2 and $\pi^{(1)}, \pi^{(2)}$ as in Section 3. With these definitions, γ_1 is sufficient for γ_2 if $\pi^{(2)} = \mathbb{E}\left[\pi^{(1)}|\gamma_2\right]$. In this appendix, it is shown that this is equivalent to the sufficiency condition given by DeGroot and Fienberg (1982), Equation (4.3). To state the latter condition, we assume that the conditional probability of γ_1 given Y and the conditional probability of γ_2 given Y, respectively, have densities

[6]Readers familiar with the concept of σ–algebras will have realised that an arbitrary σ–algebra can be substituted for $\mathcal{F}$

$g_1(p|Y)$ and $g_2(p|Y)$, respectively. Furthermore, the conditional probability of γ_2 given γ_1 is assumed to have a density $h(\gamma_2|\gamma_1)$. With these conventions, γ_1 is sufficient for γ_2 in the sense of of DeGroot and Fienberg (1982), if

$$g_2(\gamma_2|Y) = \int_{\mathcal{P}_\mathcal{I}} h(\gamma_2|\gamma_1)\, g_1(\gamma_1|Y)\, \mathrm{d}\gamma_1. \tag{18}$$

Multiplying both sides by $\bar{\pi}$ and dividing by the density of γ_2 we obtain

$$\pi^{(2)}(\gamma_2) = \int_{\mathcal{P}_\mathcal{I}} \pi^{(1)}(\gamma_1)\, f(\gamma_1|\gamma_2)\, \mathrm{d}\gamma_1, \tag{19}$$

with $f(\gamma_1|\gamma_2)$ being the conditional probability of γ_1 given γ_2. Here we need to write explicitely that $\pi^{(1)}$ and $\pi^{(2)}$ depend on γ_1 and γ_2, respectively. But the right hand side of Equation (19) is just $\mathbb{E}\left[\pi^{(1)}|\gamma_2\right]$.

D Derivation of Equation 15

Still, γ_1, γ_2 and $\pi^{(1)}, \pi^{(2)}$ are as in Section 3, with $\pi^{(1)}$ being sufficient for $\pi^{(2)}$. By just applying definitions, we get

$$\begin{aligned} \mathsf{d}(\bar{\pi}, \pi^{(2)}) &= \mathsf{s}(\bar{\pi}, \pi^{(2)}) - \mathsf{s}(\pi^{(2)}, \pi^{(2)}) \\ &= \mathsf{s}(\bar{\pi}, \pi^{(2)}) - \mathsf{s}(\pi^{(1)}, \pi^{(1)}) \\ &\quad - (\mathsf{s}(\pi^{(2)}, \pi^{(2)}) - \mathsf{s}(\pi^{(1)}, \pi^{(1)})). \end{aligned} \tag{20}$$

The mathematical expectation of the first term can be written as

$$\begin{aligned} \mathbb{E}\,\mathsf{s}(\bar{\pi}, \pi^{(2)}) &= \mathbb{E}\left[\mathbb{E}\left[S(\bar{\pi}, Y)|\gamma_2\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[S(\bar{\pi}, Y)|\gamma_1\right]\right] \\ &= \mathbb{E}\,\mathsf{s}(\bar{\pi}, \pi^{(1)}), \end{aligned} \tag{21}$$

using elementary properties of the conditional expectation and the fact that $\bar{\pi}$ is not random. Next, the expectation of the third term is considered:

$$\begin{aligned} \mathbb{E}\,\mathsf{s}(\pi^{(2)}, \pi^{(2)}) &= \mathbb{E}\left[\mathsf{s}(\pi^{(2)}, \mathbb{E}\left[\pi^{(1)}|\gamma_2\right])\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathsf{s}(\pi^{(2)}, \pi^{(1)})|\gamma_2\right]\right] \\ &= \mathbb{E}\,\mathsf{s}(\pi^{(2)}, \pi^{(1)}). \end{aligned} \tag{22}$$

The first equality is due to sufficiency; the second is valid because $\pi^{(2)}$ is a function of γ_2, so it can be taken under any expectation conditioned on γ_2; and the third equality uses elementary properties of the conditional expectation. Taking the expectation over Equation (20) and using Equations (21, 22), we obtain Equation (15).

References

Blattenberger G, Lad F. 1985. Separating the Brier score into calibration and refinement components: A graphical exposition. *The American Statistician* **39**(1): 26–32.

Breiman L. 1973. *Probability*. Addison-Wesley-Publishing.

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

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

Bröcker J, Clarke L, Kilminster D, Smith LA. 2004. Scoring probabilistic forecasts. In *First THORPEX International Science Symposium*. Montréal.

Bröcker J, Smith LA. 2008. From ensemble forecasts to predictive distribution functions. *Tellus A* **60**: 663–678.

Brown TA. 1970. Probabilistic forecasts and reproducing scoring systems. Technical Report RM–6299–ARPA, RAND Corporation.

DeGroot MW, Fienberg SE. 1982. Assessing probability assessors: calibration and refinement. *Statistical Decision Theory and Related Topics* **1**(3): 291–314.

Epstein ES. 1969. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology* **8**: 985–987.

de Finetti B. 1970. Logical foundations and measurement of subjective probability. *Acta Psychologica* **34**.

Gneiting T, Balabdaoui F, Raftery AE. 2005a. Probabilistic forecasts, calibration, and sharpness. Technical report, Department of Statistics, University of Washington.

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

Gneiting T, Raftery A, Westveld III AH, Goldmann T. 2005b. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review* **133**: 1098–1118.

Good IJ. 1952. Rational decisions. *Journal of the Royal Statistical Society* **XIV(1)**: 107–114.

Matheson JE, Winkler RL. 1976. Scoring rules for continuous probability distributions. *Management Science* **22**(10).

Murphy AH. 1971. A note on the ranked probability score. *Journal of Applied Meteorology* **10**: 155.

Murphy AH. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting* **8**.

Murphy AH. 1996. General decompositions of MSE-based skill scores: Measures of some basic aspects of forecast quality. *Monthly Weather Review* **124**.

Murphy AH, Winkler RL. 1977. Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics* **26**(1): 41–47.

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

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* **133**(5): 1155–1174.

Rockafellar RT. 1970. *Convex Analysis*. Princeton University Press. Princeton.

Savage LJ. 1971. Elicitation of personal probabilities and expectation. *Journal of the American Statistical Association* **66**(336).

Toth Z, Talagrand O, Candille G, Zhu Y. 2003. Probability and ensemble forecasts. In Jolliffe IT and Stephenson DB, editors, *Forecast Verification*, chapter 7, 137–163. John Wiley & Sons, Ltd., Chichester.

Wilks DS. 2006a. Comparison of ensemble–MOS methods in the Lorenz'96 setting. *Meteorological Applications* **13**(3): 243–256.

Wilks DS. 2006b. *Statistical Methods in the Athmospheric Sciences*, volume 59 of *International Geophysics Series*. Academic Press, second edition.

Winkler RL, Murphy AH. 1968. "Good" probability assessors. *Journal of Applied Meteorology* **7**.