Decompositions of Proper Scores

Jochen Bröcker

Technical Report
Max–Planck–Institut für Physik komplexer Systeme
Nöthnitzer Strasse 34
01187 Dresden
Germany

October 10, 2007

Abstract

Scoring rules are an important tool for evaluating the performance of probabilistic forecasts. A popular example is the Brier score, which allows for a decomposition into terms related to the sharpness (or information content) and to the reliability of the forecast. This feature renders the Brier score a very intuitive measure of forecast quality. In this paper, it is demonstrated that all strictly proper scoring rules allow for a similar decomposition into reliability and sharpness related terms. This finding underpins the importance of proper scores and yields further credence to the practice of measuring forecast quality by proper scores. Furthermore, the effect of averaging multiple probabilistic forecasts on the score is discussed. It is well known that the Brier score of a mixture of several forecasts is never worse that the average score of the individual forecasts. This property hinges on the convexity of the Brier score, a property not universal among proper scores. Arguably, this phenomenon portends epistemological questions which require clarification.

1 Introduction

Brown (1970) argues that it seems reasonable to value forecasts (be they probabilistic or other) by a scheme related to the extend 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); 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 judgements 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 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 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.

Owing to the enormous increase in computer power over the last decades, it became computationally feasible to numerically produce probabilistic forecasts for dynamical processes, 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., 2004; Gneiting and Raftery, 2007; Bröcker et al., 2004; Raftery et al., 2005; Wilks, 2006a).

In contrast to the expert–judgement–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 for archives of forecast–observation pairs to be collected. This fact allows to reconsider the interpretation of probabilities as long time observed frequencies. 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 we have available a large archive of forecast–observation pairs, reliability becomes a sensible property to ask for.

But how does reliability pertain to proper scores? Do proper scores reward reliable forecasts? In this paper, this question is answered in the affirmative. As will be shown in Section 2, proper scores allow for a decomposition into terms measuring the information content and the reliability of the forecast. This decomposition is well known for the Brier score (see for example Murphy and Winkler, 1987; Murphy, 1996), 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. We will see that any proper score can be thus decomposed, with the terms in the decomposition bearing the same interpretation as the corresponding terms in the decomposition of the Brier score. In particular, it is shown that both forecast resolution and reliability have a positive effect on the score. This result emphasises the importance of proper scores, and furthermore, since resolution and reliability are arguably desirable properties, provides an intuitive explanation why forecasts achieving a good score should indeed be considered good forecasts.

A further interesting aspect of the Brier score is its convexity with respect to the forecast probability. An important consequence is that a weighted average over several probabilistic forecasts results in a score which is better than the weighted average over the individual scores of the constituent forecasts. This feature is *not* universal to proper scores, and a counterexample (using the spherical score) is presented. But since combining probability forecasts by a weighted average is widely practised, this rises interesting epistemological questions as to whether averaging forecasts is an inherently good idea (implying that scores ought to be convex), or just a habit owing to the fact that many popular scores happen to be convex.

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 variable which is to be forecast, commonly referred to as the observation. The observation Y takes values in a space E, which can range from just a set of few alternatives (e.g. "rain/sunshine") to state spaces of very complex dynamical systems. The observation Y is a random variable, which we model by assuming it to be a function on a probability space $(\Omega, \mathbb{P})$. Values of Y (i.e. elements of E) will be denoted by small lowercase letters like x, y, or z. A probability forecast is a function $\gamma(y, \omega)$ with $y \in E$ and $\omega \in \Omega$, which, for ω held fixed, is a probability density function of y. In other words, γ is a mapping which associates with every ω a probability density function on E and can hence be considered a *random* probability density on E. The definition of γ needs a slight (and obvious) alteration if E contains discrete elements. In this case, $\gamma(y, \omega)$ denotes the probability (rather than the density) associated with y. Generic probability densities on E will be denoted by $p(\cdot)$, $q(\cdot)$, and $r(\cdot)$. As is often the custom in stochastics, the argument ω will be omitted as long it is obvious that the quantity in question is a random variable. The reason for assuming γ to be random is that forecasts usually depend on some side information that will become available at forecast time. For example, if γ is a weather forecast with lead time 48h, it will depend on weather information down to 48h prior to when the observation Y obtains. To give a further example, if γ is a forecast for wind speed over the next couple of seconds, it might for example depend on wind speed measurements up to several minutes into the past. The task of designing a forecast system is effectively to model the relationship between this side information and what is to be forecasted (see Murphy and Winkler, 1987; Murphy, 1993, 1996, for a related discussion).[1]

It was already mentioned what reliability means in case that E contains only two elements (1 and 0, say). Suppose we have a long series of independent realisations of Y and corresponding γ. For all $p \in [0, 1]$, the fraction of instances where $Y = 1$ among all instances where $\gamma(1) \cong p$ should be $\cong p$ in the long run. This amounts to saying that

$$\mathbb{P}(Y = 1 | \gamma = p) = p. \tag{1}$$

In the case of general state spaces E, this definition of reliability generalises as follows: On the condition that the forecast is equal to, say, the density $p(\cdot)$, the observation Y should be distributed according to $p(\cdot)$, or in formulae

$$\mathbb{P}(Y \in A | \gamma = p(\cdot)) = \int_A p(y) \mathrm{d}y \tag{2}$$

for every set $A \subset E$. Our reasoning as to why reliability (as defined through Equ. 2) is a desirable property was based on heuristic arguments. But Equation (2) could also be derived from the postulate that γ be some conditional probability. This is demonstrated in Appendix A. The conditional probability $\mathbb{P}(Y \in A|\gamma)$ (on the left hand side of Equation (2)) denotes the conditional probability of the observation given the forecast. It is assumed throughout the paper that this conditional probability possesses a density, which will be denoted as $\pi(\cdot)$. Like every conditional probability density, $\pi(\cdot)$ is a random quantity. Hence, strictly speaking, $\pi(\cdot)$ is a mapping which associates with every ω a probability density function on E, very much like γ. In terms of π and γ, Equation (2) (i.e. reliability of γ) can be written simply as

$$\pi = \gamma. \tag{3}$$

It is important to note that *in any case*, π is γ–measurable, that is, effectively a functional of γ (see e.g. Breiman, 1973). In particular, in calculations where γ is assumed to be fixed, π is also fixed.

We now turn our attention to *scoring rules* (see Matheson and Winkler, 1976; Gneiting and Raftery, 2007, whom we follow closely in terms of notation). A scoring rule is a function $S(p, y)$ which takes a probability density over E as its first argument and an element of E (i.e. an observation) as its second argument. For any two densities $p(\cdot)$ and $q(\cdot)$, the *scoring function* is defined as

$$\mathsf{s}(p, q) = \int S(p, y) q(y) \mathrm{d}y. \tag{4}$$

[1]We are not considering forecasting problems which are explicitly dependent on time, for example to take into account seasonal effects.

Name	scoring rule $S(p,y)$	divergence $\mathsf{d}(q,p)$	entropy $\mathsf{e}(p)$
Brier	$\vert y-p\vert^2$	$\vert p-q\vert^2$	$p(1-p)$
Ignorance[a]	$-\log(p(y))$	$\int -\log\left(\frac{p}{q}(z)\right) q(z)\mathrm{d}z$	$\int -\log(p(z))p(z)\mathrm{d}z$
CRPS[b]	$\int (F(z)-H(y-z))^2\mathrm{d}z$	$\int (F(z)-G(z))^2\mathrm{d}z$	$\int F(z)(1-F(z))\mathrm{d}z$
PSS[c]	$-p(y)^{\alpha-1}/\Vert p\Vert_\alpha^{\alpha-1}$	$\Vert q\Vert_\alpha - \int q(z)p(z)^{\alpha-1}\mathrm{d}z/\Vert p\Vert_\alpha^{\alpha-1}$	$-\Vert p\Vert_\alpha$
PLS[d]	$\int p^2(z)\mathrm{d}z - 2p(y)$	$\int (p(z)-q(z))^2\mathrm{d}z$	$-\int p^2(z)\mathrm{d}z$

[a]Propriety follows from Jensen's inequality.
[b]Continuous Ranked Probability Score – Here F and G are the cumulative distribution functions corresponding to p and q, respectively.
[c]Pseudospherical Scores – Here $\alpha > 1$, while $\Vert p\Vert_\alpha$ denotes the L_α–norm. Propriety follows from Hölder's Inequality.
[d]Proper Linear Score, also referred to as the quadratic score.

Table 1: Scoring rule, divergence, and entropy for several common scores. All integrals extend over E. 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.

The interpretation of the scoring function is that if Z is a random variable of distribution q, then $\mathsf{s}(p,q)$ is the score we expect the density p to achieve in forecasting Z. It is our convention that a small score indicates a good forecast. A score is called *proper* if the *divergence*

$$\mathsf{d}(p,q) = \mathsf{s}(p,q) - \mathsf{s}(q,q) \tag{5}$$

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. 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

$$\mathsf{e}(p) = \mathsf{s}(p,p) \tag{6}$$

is called the *entropy* of p.[2] For strictly proper scores,

$$\mathsf{e}(p) = \inf_q \mathsf{s}(q,p). \tag{7}$$

Since $\mathsf{s}(q,p)$ is linear in p, Equation (7) demonstrates that for strictly proper scores, the entropy $\mathsf{e}(p)$ is an infimum over linear functions and hence concave (Rockafellar, 1970). From now on, the scoring rule S will be assumed to be strictly proper! Table 1 gives a couple of frequently used scoring rules along with the corresponding divergences and entropies.

`Table 1 on top of this page`

Going back to the forecasting problem, the *expected score* of the forecast γ is given by $\mathbb{E}(S(\gamma,Y))$. Since γ is random, it is also affected by the expectation, unlike as in the discussion following Equation (4). However, we can write

$$\mathbb{E}(S(\gamma,Y)) = \mathbb{E}(\mathbb{E}(S(\gamma,Y)|\gamma)) \tag{8}$$

[2]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, we define $\mathsf{e}(p)$ to be the entropy.

To calculate the conditional expectation $\mathbb{E}(S(\gamma, Y)|\gamma)$, we need the density of Y given γ. But this is just π, whence

$$\mathbb{E}(S(\gamma, Y)|\gamma) = \mathsf{s}(\gamma, \pi). \tag{9}$$

Substituting with Equation (9) in Equation (8) results in

$$\mathbb{E}(S(\gamma, Y)) = \mathbb{E}(\mathsf{s}(\gamma, \pi)). \tag{10}$$

We will now turn to the mentioned decomposition of the expected score $\mathbb{E}(S(\gamma, Y))$ (Equ. 10). It is easily seen that

$$\mathsf{s}(\gamma, \pi) = \mathsf{e}(\pi) + \mathsf{d}(\gamma, \pi). \tag{11}$$

Taking the expectation on both sides of Equation (11) and substituting for the right hand side in Equation (10), we obtain

$$\mathbb{E}(S(\gamma, Y)) = \mathbb{E}\mathsf{e}(\pi) + \mathbb{E}\mathsf{d}(\gamma, \pi). \tag{12}$$

The first term in Equation (12), the average entropy of π, can be decomposed further. Consider the density obtained by taking the average over π,

$$c(y) := \mathbb{E}\pi(y, \omega) \tag{13}$$

(the average of course extends over ω). It is easily seen that $c(y)$ is just the unconditional density of Y, which in meteorology is often referred to as the climatology of Y. Since $\mathsf{s}(c, \pi)$ is linear in π and c is not random, it follows immediately from Equation (13) that

$$\mathbb{E}\mathsf{s}(c, \pi) = \mathsf{s}(c, c) = \mathsf{e}(c). \tag{14}$$

Adding and subtracting $\mathbb{E}\mathsf{s}(c, \pi)$ on the right hand side of Equation (12) and using Equation (14) we arrive at

$$\mathbb{E}\mathsf{s}(\gamma, y) = \mathsf{e}(c) - \mathbb{E}\mathsf{d}(c, \pi) + \mathbb{E}\mathsf{d}(\gamma, \pi). \tag{15}$$

Equation (15) constitutes the desired decomposition of the expected score of the probabilistic forecast γ. This decomposition is, as we will argue, completely analogous to and a generalisation of the well known decomposition of the Brier score. The three terms in Equation (15) will be (from left to right) referred to as the uncertainty of Y, the resolution term, and the reliability term. As a starting point for the discussion of the decomposition Equation (15), the reader might want to convince himself (with the help of Table 1) that for the Brier score, Equation (15) indeed yields the known decomposition. As to the general decomposition, let us first consider the entropy. For the particular cases listed in Table 1, the interpretation of the entropy as a measure for the uncertainty inherent in the density c should pose no problem. For the Brier score and the Ignorance, the entropy is indeed a very common measure of inherent randomness of a distribution. In general, the entropy can be interpreted as the expected score of the unconditional density (climatology) as a forecast, or in other words the ability of the density to forecast random draws from itself. Furthermore, suppose p and q are two densities featuring the same entropy, then any mixture of p and q should have a larger inherent uncertainty and hence a larger entropy than the individual densities, which is the case due to the concavity of $\mathsf{e}(p)$. The resolution term $\mathbb{E}\mathsf{d}(c, \pi)$ contributes negatively to the score (i.e. the larger the resolution term, the smaller and hence the better the score). Note that due to the strict propriety of the score, the resolution is always positive definite. Since the resolution term describes the average deviation of π from its average c (see Equation 13), it can be interpreted as a form of variance of π (which reduces to the standard variance of π in case of the Brier score). Finally, the reliability term (which is again positive definite) describes the average deviation of γ from π. Recalling that $\gamma = \pi$ indicates a reliable forecast, the interpretation of the reliability term as the average violation of reliability becomes obvious.

Presumably the most important aspect of the decomposition Equation (15) is that it demonstrates the positive effect of both forecast resolution and reliability on the score. As resolution and reliability are arguably desirable properties of a probabilistic forecasts in their own right, Equation (15) in fact provides an intuitive explanation why forecasts that achieve a good score should indeed be considered good forecasts. Furthermore, as this result depends solely on the propriety of the score, it further emphasises the importance of proper scores.

3 The Effects of Mixing Different Forecasts

In this section, the effect of averaging probabilistic forecasts on the resulting score is discussed. It is not uncommon to have several probabilistic forecasting systems available for one and the same forecasting problem, for example weather forecasts from different forecasting centres or simply opinions from a number of disagreeing experts. A simple and widely practised way to combine these forecasts is by convex mixture (Winkler, 1989; Clemen, 1989; Doblas-Reyes et al., 2005; Hall and Mitchell, 2007). For the discussion in this paper, it is unimportant how the weights are obtained, but possible ways could be by maximising the score over an archive of forecast–observation pairs or simply gut feeling. Convex mixing of heterogeneous probability forecasts is known to have a positive effect on the Brier score. More specifically, any convex combination of forecasts achieves a score not worse than the same convex combination applied to the scores of the individual forecasts. Since the Brier score is effectively the mean square error, this behaviour is well known in statistical learning and regression too, where averaging over heterogeneous models is known as ensembling (Krogh and Sollich, 1997).

The property responsible for this phenomenon is of course the convexity of the Brier score. Suppose that $S(p, Y)$ is a scoring rule which is convex as a function of p for every Y. Furthermore, let $\gamma_\nu, \nu = 1 \ldots N$ be a collection of probability forecasts over E. Finally, let $\tau_\nu, \nu = 1 \ldots N$ be convex mixing coefficients, that is $\tau_\nu \geq 0$ and $\sum_\nu \tau_\nu = 1$. Then

$$\sum_\nu \tau_\nu \mathbb{E}\left[S(\gamma_\nu, Y)\right] \geq \mathbb{E}\left[S(\sum_\nu \tau_\nu \gamma_\nu, Y)\right]. \tag{16}$$

The left hand side are the expected scores of the individual γ_ν, averaged over ν, while the right hand side is the expected score of the forecast combination $\sum_\nu \tau_\nu \gamma_\nu$.

In general, scores do not need to be convex, although many popular ones are. The spherical scores provide examples of non-convex scores, as the following example demonstrates. Suppose that $Y = 0$ or 1 only, with $\mathbb{P}(Y = 1) = c$. We consider a constant (i.e. nonrandom) “forecast” $\gamma(1) \in [0, 1]$, that is, we are essentially looking for a good approximation to the climatology c. The expected spherical score (of order $\alpha = 2$) of γ is

$$\mathbb{E}S(\gamma, Y) = -\frac{\gamma(1) \cdot c + \gamma(0) \cdot (1 - c)}{\sqrt{\gamma(1)^2 + \gamma(0)^2}}. \tag{17}$$

Expanding this to second order for small $\gamma(1)$ (and using $\gamma(0) = 1 - \gamma(1)$) gives

$$\mathbb{E}S(\gamma, Y) \cong \frac{\gamma(1)^2}{2}(1 - 3c) - \gamma(1) \cdot c + (c - 1). \tag{18}$$

Hence if $c > 1/3$, the expected score is concave for small $\gamma(1)$, and consequently Equation (16) can get reversed for small $\gamma_\nu(1)$. For the spherical score this effect is small, but it becomes very pronounced for the pseudospherical scores at higher numbers of α.

Although this phenomenon is surprising, we did not find any compelling argument why scores should be convex. This point is worth investigating further, as it obviously bears on the common practice of averaging forecasts. There might be reasons why using a mix of heterogeneous models is an inherently good idea. This would imply that only convex scores ought to be used (or scores which feature a weaker form of convexity or monotonicity). But mixing forecast might as well have become common practice simply due to the fact that many popular scores happen to be convex. In this case, mixing forecasts needs justification on a case by case basis, for example by establishing that the particular task at hand naturally suggests a convex score.

4 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. This result is the natural generalisation of the well known decomposition of the Brier score (i.e. the quadratic error). The only property required of the score is that it be strictly proper, whence this finding emphasises the importance of using proper scores. Furthermore, the effect of averaging several probabilistic forecasts on the score was investigated. Averaging has positive effects if the score is convex. However, proper scores are not necessarily convex. Therefore it is argued that this popular and often recommended practice need be scrutinised again.

5 Acknowledgements

The author gratefully acknowledges fruitful discussions with the members of the Centre for the Analysis of Time Series, London School of Economics, in particular Liam Clarke, Milena Cuellar and Leonard A. Smith, as well as members of the Time Series Analysis group and the Max–Planck–Institute for the Physics of Complex Systems.

A An Alternative Definition of Reliability

In this section, a different definition of reliability will be provided. More specifically, we will derive Equation (2) from a mathematical postulate. For a more streamlined formulation, the probabilistic forecast is assumed to have the form of a random probability measure Γ on E, rather than a density. The appropriate form of the reliability criterion Equation (2) reads as

$$\mathbb{P}(Y \in A|\Gamma) = \Gamma(A). \tag{19}$$

The postulate is simply that the probabilistic forecast Γ can be written as some conditional probability of Y given "something". What is this "something"? As was discussed in Section 2, Γ usually depends on some side information. For the purpose of the present discussion, it suffices to assume that the side information can be subsummised in a random variable I. The postulate mentioned above can hence be written as

$$\mathbb{P}(Y \in A|I) = \Gamma(A) \tag{20}$$

For the discussion in this (and many similar papers), the exact nature of the side information I is unimportant. In this respect, it is convenient to know that a probabilistic forecast is reliable if and only if it can be represented as a conditional probability. The "only if"–part is obvious: If the forecast is reliable, then according to Equation (19) it can be written as a conditional probability. To see the "if"–part, note that, because of Equation (20), Γ is measurable with respect to I, which effectively means that Γ can be written as a function of the side information I. For this reason, an elementary property of the conditional expectation gives

$$\mathbb{E}(\mathbb{P}(Y \in A|I)|\Gamma) = \mathbb{P}(Y \in A|\Gamma) \tag{21}$$

Using again Equation (20), this time to substitute for $\mathbb{P}(Y \in A|I)$ in Equation (21), we obtain

$$\mathbb{E}(\Gamma(A)|\Gamma) = \mathbb{P}(Y \in A|\Gamma). \tag{22}$$

But the left hand side is obviously just $\Gamma(A)$, which demonstrates that Γ is reliable in the sense of Equation (19).

References

Leo Breiman. *Probability.* Addison-Wesley-Publishing, 1973.

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

Jochen Bröcker, Liam Clarke, Devin Kilminster, and Leonard A. Smith. Scoring probabilistic forecasts. In *First THORPEX International Science Symposium*, Montréal, 2004.

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

Robert T. Clemen. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5, 1989.

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

F.J. Doblas-Reyes, R. Hagedorn, and T. Palmer. The rationale behind the success of multi-model ensembles in seasonal forecasting. part ii: Calibration and combination. *Tellus*, 57(3), 2005.

Edward S. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8, 1969.

T. Gneiting, A. Raftery, A. H. Westveld III, and T. Goldmann. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 2004.

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

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

Stephen G. Hall and James Mitchell. Combining density forecasts. *International Journal of Forecasting*, 23(1), 2007.

Anders Krogh and Peter Sollich. Statistical mechanics of ensemble learning. *Physical Review E*, 55(1), 1997.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10), 1976.

Allan H. Murphy. A note on the ranked probability score. *Journal of Applied Meteorology*, 10, 1971.

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

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

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

Allan H. Murphy and Robert L. Winkler. A general framework for forecast verification. *Monthly Weather Review*, 115:1330–1338, 1987.

Adrian E. Raftery, Tilman Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 2005.

T. Rockafellar. *Convex Analysis*. Princeton University Press. Princeton, 1970.

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

Zoltan Toth, Olivier Talagrand, Guillem Candille, and Yuejian Zhu. Probability and ensemble forecasts. In Ian T. Jolliffe and David B. Stephenson, editors, *Forecast Verification*, chapter 7. John Wiley & Sons, Ltd., 2003.

Daniel S. Wilks. Comparison of ensemble–MOS methods in the Lorenz'96 setting. *Meteorological Applications*, 13, 2006a.

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

Robert L. Winkler. Combining forecasts: A philosophical basis and some current issues. *International Journal of Forecasting*, 5, 1989.

Robert L. Winkler and Allan H. Murphy. "Good" probability assessors. *Journal of Applied Meteorology*, 7, 1968.