Scoring Probabilistic Forecasts: The Importance of Being Proper

Jochen Bröcker[1,*] and Leonard A. Smith[1,2]

[1]Centre for the Analysis of Time Series
London School of Economics
London, UK

[2]Pembroke College
Oxford University
Oxford, UK

[*]Corresponding Author, cats@lse.ac.uk

August 3, 2006

Abstract

Questions remain regarding how the skill of operational probabilistic forecasts is most usefully evaluated or compared, even though probability forecasts have been a longstanding aim in meteorological forecasting. This paper explains the importance of employing proper scores when selecting beween the various measures of forecast skill. It is demonstrated that only proper scores provide internally consistent evaluations of probability forecasts, justifying the focus on proper scores independently of any attempt to influence the behaviour of a forecaster. Another property of scores, locality, is discussed. Several scores are examined in this light. There is, effectively, only one proper, local score for probability forecasts of a continuous variable. It is also noted that operational needs of weather forecasts suggest that the current concept of a score may be too narrow; a possible generalisation is motivated and discussed in the context of propriety and locality.

1 Introduction

Useful probabilistic forecasts have long been a goal in operational weather forecasting, as has the idea that, by its very nature, the meteorological problem makes a probabilistic solution "desirable if not inevitable" (Petterssen, 1956). Modern telecommunications allow the user of weather model output to construct weather forecasts using simulations from several operational centres. Furthermore, ensembles of simulations under the same model provide flow dependent uncertainty information, superior to the traditional use of historical errors, for translating simulations into probabilistic forecasts (Palmer, 2000; Palmer et al., 2005).

As probability forecasts become more common, the need to select one method from among the plethora of alternatives for constructing and tuning probabilistic forecasts as well as growing interest in how to better quantify the improvement in probabilistic forecasting techniques (Jolliffe and Stephenson, 2003) has stimulated the development or adoption of a number of *scores* (Wilks, 1995; Gneiting and Raftery, 2004; Roulston and Smith, 2002). While the true value of a forecast is its utility to the end user, scores are fundamental to the performance analysis of probabilistic forecasts and, ideally, provide a general measure of forecast quality, independent of any specific end user. We will examine several scores in detail in Section 3, each of which aims to quantify the quality of a probabilistic forecast system given a series of forecast–verification pairs. The main aim of this paper is to demonstrate the requirements on the scores to ensure internally consistent forecast evaluation, rather than how scores could be employed in connection with forecast archives to either evaluate or improve probabilistic forecast systems.

Our main focus is on the importance of using *proper* scores, as outlined in Section 4. After defining this property, it is demonstrated that only proper scores are internally consistent in the sense that a forecast probability distribution is given an optimal expected score when the verification is, in fact, drawn from that probability distribution. Using proper scores may have other positive side effects on the behaviour of forecasters, as argued by Murphy and Winkler (1987). While discussions on motivating the honesty of forecasters is sometimes wide ranging, the importance of using proper scores can be motivated solely on the grounds that mathematically, only proper scores are internally consistent.

Scores used as examples in this paper include the Ignorance, Brier Score, and the Naive Linear Score. Although widely discussed (Wilson et al., 1999), the Naive Linear Score is *not* a proper score. We derive a variant of the Linear Score that is proper. In Section 5 we consider the issues surrounding the notion of *locality*, and note that uncertain observations may drive us to use generalised scores. Concluding remarks are made in Section 6

2 Probabilistic Forecasts

To give a mathematical definition of a probabilistic forecast, let us consider a variable of interest, say the temperature at London Heathrow airport on a specific day. We will use the symbol X to denote the observed value of that variable. The corresponding lower case x denotes any possible value in the range of X. In the case of London Heathrow temperature, x could be any real

number larger than $-273°$C. In this paper we focus on probabilistic forecasts in the form of probability density functions $p(x)$, which express uncertainty over what the possible values of X will be, based on the information in hand. By $p(x)$ we denote the entire function, while the notation $p(X)$ always denotes the value of the function at the particular observation X. Different information may well lead to different probability forecasts for X denoted by $p(x), q(x), r(x), \ldots$.

A priori, $p(x)$ is only required to be normalised and nonnegative; in symbols

$$\int p(x)\mathrm{d}x = 1. \tag{1}$$

For the discussion in this paper, it is not relevant how the probabilistic forecast came about. It might have been computed using highly sophisticated models or rather simple ones. Of course, to meaningfully evaluate probabilistic forecast systems, access to a forecast archive of forecast-verification pairs is necessary; it is difficult, if not impossible, to usefully evaluate a single probability forecast, and the size of the forecast archive plays a major role in determining the significance of the result, regardless of which score is employed. The aim of this paper is merely to show why only proper scores should be used. This does neither depend on how the forecast is constructed nor on the size of forecast archives.

3 Scores

A *score* attempts to compare X and each of the probabilistic forecasts. Yet X and $p(x)$ are unlike quantities, rendering a point-to-point distance measure such as the square distance inappropriate. Scores provide more general measures of comparison. A score is a function $S(p(x), X)$, where $p(x)$ is a probability density and X is the verification. Note that $S(p(x), X)$ might depend on the whole functional form of $p(x)$ (e.g. by integration). In other words, $S(p(x), X)$ acts on $p(x)$ as an operator. To give the reader an impression of how a score $S(p(x), X)$ would be used to evaluate the quality of a forecast system, assume we had an archive of forecast–verification pairs at our disposal, that is, a large number N of forecasts $\{p_i(x), i = 1 \ldots N\}$ and corresponding verifications $\{X_i, i = 1 \ldots N\}$. The forecast system would then be valued according to its *empirical skill*

$$\langle S \rangle = \frac{1}{N} \sum_{i}^{N} S(p_i(x), X_i).$$

The point of this paper is that not all conceivable scores S should be used for this purpose, but rather only proper ones. As will be explained, this is a property of the function S alone. Throughout this paper, scores are defined like cost functions: small numerical values indicate better forecasts.

Examples for scores include:

The Ignorance Score The Ignorance Score (Good, 1952; Roulston and Smith, 2002) is defined by

$$S(p(x), X) = -\log(p(X)) \tag{2}$$

To our knowledge, it was first mentioned in connection to weather forecasting by Good (1952), who went so far as to suggest that the funding of the UK Met Office should vary with it. Ignorance has been interpreted

in information theoretic terms (Roulston and Smith, 2002) and directly quantifies expected returns in certain betting scenarios commonly used to quantify economic utility.

The Brier Score The Brier Score (Candille and Talagrand, 2004; Jolliffe and Stephenson, 2003) is defined for binary X, i.e. $X = 0$ or 1 only. Intuitively, $P(X = 1)$ "should" be close to 1 if $X = 1$ and close to 0 if $X = 0$. The Brier score quantifies this via:

$$S(p, X) = (X - p)^2, \tag{3}$$

where $p = P(X = 1)$. Note that the use of p here differs slightly from our notational conventions for continuous X.

The Naive Linear Score The Naive Linear Score applies to continuous X and is defined as

$$S(p(x), X) = -p(X). \tag{4}$$

Although often suggested as a possible score, the Naive Linear Score is not proper, as will be demonstrated and discussed in Section 4.

The Proper Linear Score This score applies to continuous X and is defined as

$$S(p(x), X) = \int p^2(z)\mathrm{d}z - 2p(X). \tag{5}$$

It is a strictly proper alternative to the Naive Linear Score of Equation (4). The fact that the additional term $\int p^2(z)\mathrm{d}z$ renders the score strictly proper will be demonstrated in Section 4. Selten (1998) discussed it and contrasted it with the Ignorance.

The Mean Square Error This score can be applied to continuous X with the definition

$$S(p(x), X) = \int (X - z)^2 p(z)\mathrm{d}z \tag{6}$$

This score measures the spread of $p(x)$ around X. If we let m and s be the mean and the standard deviation of $p(x)$ respectively[1], that is

$$m = \int xp(x)\mathrm{d}x$$

and

$$s = \sqrt{\int (x - m)^2 p(x)\mathrm{d}x},$$

this score can be written as

$$S(p(x), X) = (X - m)^2 + s^2. \tag{7}$$

Thus, the Mean Square Error depends on $p(x)$ only through its first and second moment. It does not reflect any other aspect of $p(x)$. The implications of this will be discussed later.

[1]The mean and the standard deviation of $p(x)$ are not to be confused with the sample mean and the sample standard deviation of the observations.

Note that the Proper Linear Score depends on the entire functional form of $p(x)$ (due to the integral in the first term of Equation 5), while both the Ignorance and the Naive Linear Score depend on $p(x)$ only via the single number $p(X)$, the value of $p(x)$ at the verification X. That is, the Ignorance and the Naive Linear Score depend only on the value of the probabilistic forecast at the verification, not on other features of the functional form of $p(x)$. This property is called *locality*, which we return to later in Section 5.

4 Proper Scores

At first glance, the various scores presented above possess no distinctive features qualifying them as particularly useful in valuing probabilistic forecasts. As will be shown in this section though, some of these scores are *proper*, while others are not. We will first define this property and subsequently explain why improper scores lead to conclusions inconsistent with common sense, thus motivating the importance of being proper.

Mathematically, a score is proper if for any two probability densities $p(x)$ and $q(x)$

$$\int S(p(x), z)q(z)\mathrm{d}z \geq \int S(q(x), z)q(z)\mathrm{d}z. \tag{8}$$

In words: the minimum of the left hand side over all possible choices of $p(x)$ is obtained if $p(x) = q(x)$ for all x. A score is strictly proper if this happens *only* if $p(x) = q(x)$ for all x.

The central argument for employing only proper scores becomes apparent when the meaning of the two integrals in Equation 8 is explained. In short, a proper score will always prefer a probabilistic forecast if it is, in fact, more accurate. Suppose $q(x)$ is our bespoke forecast. If we knew the verification X, the skill of the forecast $q(x)$ would be $S(q(x), X)$. Although we do not know X at the moment, we still can compute the skill we *expect* to obtain by averaging the quantity $S(q(x), X)$ over all possible values of X using the forecast we possess (namely $q(x)$). This can be written as

$$\text{Forecasted Skill of } q(x): \int S(q(x), z)q(z)\mathrm{d}z.$$

This is the integral on the right–hand–side of Equation 8. Given an additional forecast $p(x)$, we can again employ $q(x)$ to evaluate the expected skill of $p(x)$, which is

$$\text{Forecasted Skill of } p(x): \int S(p(x), z)q(z)\mathrm{d}z$$

This is the integral on the left–hand–side of Equation 8. Note that $q(x)$ was used to predict the skill of $p(x)$. Propriety implies that the latter integral is always larger than the former, or in other words that we expect $p(x)$ to be *less* skillful than $q(x)$ when the expectation is evaluated using $q(x)$. Otherwise, the score we are using leads to a contradiction: It would rank $p(x)$ above $q(x)$ even if X was actually drawn from $q(x)$. This is a property of the score alone, not of the particular distributions $p(x)$ or $q(x)$. Under a proper score, we would likewise expect $q(x)$ to be less skillful than $p(x)$ if the expected skill was calculated using $p(x)$ instead of $q(x)$. Propriety is a property of the score, it is neither neccessary

to assume that X is drawn from any kind of "true" distribution nor that any kind of data is to hand. The question of whether the employed score is proper or not can be answered before any data is considered.

Alternatively, consider any two forecasts $p(x)$ and $q(x)$. Trivially we can write

$$\begin{aligned} \int S(p(x), z)q(z)\mathrm{d}z = & \int S(q(x), z)q(z)\mathrm{d}z \\ + & \quad [\int S(p(x), z)q(z)\mathrm{d}z - \int S(q(x), z)q(z)\mathrm{d}z]. \end{aligned} \tag{9}$$

If S is proper, the term in square brackets is positive definite. Strict Propriety means that the term in square brackets is strictly positive definite. Thus, if X was drawn from the the distribution $q(x)$, the skill of any forecast, if measured according to a (strictly) proper score, could be decomposed into the skill of $q(x)$ plus a (strictly) positive definite term. Again, this holds for any two $p(x), q(x)$.

For the Brier score this decomposition (9) is well known as the *Reliability–Sharpness* decomposition (Wilks, 1995). To show this, write the Brier score as

$$\begin{aligned} E_q(X-q)^2 &= E_q(X-p+p-q)^2 \\ &= E_q(X-p)^2 + (p-q)^2 + 2E_q(X-p)(p-q) \\ &= E_q(X-p)^2 - (p-q)^2, \end{aligned} \tag{10}$$

since $E_q(X) = q$, where E_q indicates expectation with respect to q. The first term on the right hand side is the Brier score of p. Adding $(p-q)^2$ on both sides, Equation (10) becomes the same decomposition as Equation (9). This shows also that the Brier score is strictly proper, since the parenthesised term in Equation (9) is $(p-q)^2$ and thus indeed strictly positive definite.

We next demonstrate briefly whether or not the further scores mentioned in the last section are (strictly) proper. The Ignorance is strictly proper, as can be derived from the fact that

$$\int -\log\left(\frac{p(z)}{q(z)}\right) q(z)\mathrm{d}z \geq 0, \tag{11}$$

with equality if and only if $p(x) = q(x)$ (Kullback-Leibler Inequality (Kullback and Leibler, 1951)). The Proper Linear Score is indeed also strictly proper, given the fact that

$$\int (q(z) - p(z))^2 \,\mathrm{d}z \geq 0, \tag{12}$$

with equality if and only if $p(x) = q(x)$ for all x. The left–hand–side of Equation 12 can be written as

$$\int [q(z)^2 + p(z)^2 - 2p(z)q(z)] \,\mathrm{d}z = \int q(z)^2\mathrm{d}z + \int S(p(x), z)q(z)\mathrm{d}z, \tag{13}$$

which is the proper linear score plus the square integral over $q(x)$ which, being a constant, does not enter the minimisation over $p(x)$. Therefore, the score is minimal if and only if $p(x) = q(x)$.

The Naïve Linear Score however is *improper*: even if X were drawn from $q(x)$, the Naïve Linear Score would not judge $q(x)$ the best. Probability density functions $p(x)$ different from $q(x)$ would rank higher than $q(x)$. In short, there are $p(x)$ which would be judged to have greater skill. In fact, for any given $q(x)$ it is always possible to find a $p(x)$ so that

$$\int -p(z)\, q(z)\mathrm{d}z \leq \int -q(z)\, q(z)\mathrm{d}z.$$

Actually, the Naïve Linear Score favours a $p(x)$ featuring a very small spread and which is centred at a point $\bar{x}$ for which $q(\bar{x})$ is very large. To see this, consider first the case where $q(x)$ is not constant. Then there is a $\bar{x}$ so that

$$-q(\bar{x}) < \int -q(z)q(z)\ \mathrm{d}z = \int S(q(x), z)q(z)\ \mathrm{d}z.$$

This point $\bar{x}$ is a point where $q(x)$ is larger than average. If we take an arbitrary kernel function $g(x)$ that has a continuous derivative, is symmetric, and normalised and define $p_\sigma(x) = \frac{1}{\sigma} g(\frac{x-\bar{x}}{\sigma})$, i.e. center the kernel at $\bar{x}$ with spread σ, it follows that

$$\int S(p_\sigma(x), z)q(z)\mathrm{d}z = \int -p_\sigma(z)q(z)\mathrm{d}z \to -q(\bar{x}).$$

In other words, the Naïve Linear Score rewards assigning excess probability to high–probability x, which requires assigning too low probability to low–probability x. If however $q(x)$ is constant, then

$$\int -p(z)q(z)\mathrm{d}z = -q$$

for any $p(x)$. So in this case the score does not discriminate between forecasts at all.

The non–propriety of the Naïve Linear Score would also emerge as a consequence of a far more general result due to Bernardo (1979) (see also Page 8), namely that for continuous variables all smooth, proper and local scores are affine functions of the Ignorance. The notion of locality, briefly mentioned at the end of Section 3, will be returned to in Section 5. Proper scores in general have been characterised by Gneiting and Raftery (2004).

The Mean Square Error is improper as well. This can be seen as follows. Let m_p and s_p be the mean and the standard deviation of $p(x)$. Likewise let m_q and s_q be the mean and the standard deviation of $q(x)$. Using the representation Equation 7 we have

$$\begin{aligned} \int S(p(x), z)q(z)\mathrm{d}z &= \int (z - m_p)^2 q(z)\mathrm{d}z + s_p^2 \\ &= (m_q - m_p)^2 + s_q^2 + s_p^2. \end{aligned}$$

But this quantity is *not necessarily* larger than

$$\int S(q(x), z)q(z)\mathrm{d}z = 2s_q^2,$$

as it would have to be for the Mean Square Error to be proper. In fact, as for the Naïve Linear Score, a density $p(x)$ centred around the mean m_q and having small standard deviation s_p^2 would achieve a better score than $q(x)$ itself.

Sometimes only the mean of $p(x)$ is eventually used as a forecast. The error in the mean can be taken as a score, effectively setting

$$S(p(x), X) = (X - m_p)^2.$$

This score is proper, but not strictly proper, as follows from the fact that $\int (z - m_p)^2 q(z)\mathrm{d}z$ is minimal if $m_p = m_q$, in particular if $p(x) = q(x)$, yet *every other*

pdf $p(x)$ having the same first moment m_p will achieve the same score, no matter how distorted the distribution is! Even a forecast that, for example, assigned zero probability wherever $q(x)$ is non-zero but had the same mean would achieve the same score.

To conclude, we note that proper scores and only proper scores are internally consistent in that the score $S(q(x), X)$ assigns an optimal expected value to $q(x)$ if and only if X is distributed according to $q(x)$. Note that philosophical arguments over the existence of a "true" probability distribution play no role in the entire discussion of this paper. It is tempting to think of the skill of a forecast $p(x)$ as its distance to a true (in any sense) conditional probability describing the relation between our information and the unknown variable X. Since we do not assume the existence of such a "true" probability distribution, much less having access to it, we are unable to consider *distance measures* between probability distributions, gainfully explored in other circumstances by Kleeman (2002). A proper score merely ensures consistency.

5 Locality and Non-locality

A score is *local* if the probabilistic forecast is evaluated only at the actual verification. As an example, contrast the (nonlocal) Proper Linear Score, which involves the functional operation of integrating over $p(x)^2$, with the (local) Ignorance Score, which is simply the logarithm of the probability density function taken at the verification. In other words, a score is local if and only if it can (with a slight abuse of notation) be written as

$$S(p(x), X) = S(p(X), X).$$

Thus, for local scores, S does not act on the whole function $p(x)$ any more but is just a usual function of the two real numbers $p(X)$ and X. Therefore, it makes sense to define *smooth* local scores as local scores for which the function S has continuous partial derivatives with respect to these two arguments.[2]

At first sight, it might seem unreasonable that features of the forecast other than the value it assigned to the verification should matter. Yet it is possible that domain knowledge suggests any appropriate forecast should have, for example, some smoothness properties; one may want to restrict the possible variations in the probability forecast *a priori*, without having looked at the data[3]. This can also be useful when scores are employed for training models translating numerical model simulations into probability density forecasts. A ubiquitous problem here is to limit model complexity, which can be addressed by enforcing certain measures of smoothness upon the probability density functions (regularisation). Since a finite sample of verifications is *never* sufficient to either confirm or deny the presence of such properties, smoothness has to be enforced either by restricting the class of density functions considered to smooth

[2]Note that a similar definition for nonlocal scores would require a substantially more advanced concept of smoothness, since in general nonlocal scores involve function*al* operations.

[3]The definition of locality as given here must not be confused with issues related to scoring forecasts for spatial fields. There the question arises whether fields should have some smoothness properties over space, rather than over different verifications.

functions a priori, or by augmenting the score with a term that penalises non–smooth densities, essentially rendering the score nonlocal[4] [5].

A separate reason for using nonlocal measures in a particular problem would arise if the Ignorance score is not considered suitable. For example, the Ignorance score is infinity if the forecast assigns vanishing probability to an event that obtains. If we wish to usefully evaluate forecasts which insist on assigning zero probability to events that occur, we would have to resort to other scores. Inasmuch as Ignorance is the only smooth, proper and local score for continuous variables (Bernardo, 1979)[6], this implies switching to a nonlocal score.

Nonlocal evaluation measures also arise naturally when the value of the verification is uncertain, although the whole concept of scores needs a slight alteration in this situation. Suppose we have a probabilistic forecast $p(x)$ for X, but we in fact observe Z, which is X corrupted with additive observation noise. Assuming that the density of the noise is known, the conditional density $\kappa(z|x)$ of Z given X can be computed. Any forecast $p(x)$ for X gives rise to a forecast $\bar{p}(z)$ for Z by means of

$$\bar{p}(z) = \int \kappa(z|x)p(x)\mathrm{d}x.$$

Applying a score S to $\bar{p}(z)$ and Z, we can define the *generalised score* $\bar{S}$ for $p(x)$ by setting

$$\bar{S}(p(x), Z) = S(\bar{p}(z), Z).$$

Here the right–hand–side defines the left–hand–side. We then define a generalised score to be proper if for any $q(x)$ we have

$$\int \bar{S}(p(x), z)\bar{q}(z)\mathrm{d}z \geq \int \bar{S}(q(x), z)\bar{q}(z)\mathrm{d}z, \tag{14}$$

where, as for $p(x)$,

$$\bar{q}(z) = \int \kappa(z|x)q(x)\mathrm{d}x.$$

If S is proper, $\bar{S}$ is proper as well. If S is strictly proper though, $\bar{S}$ is *not neccessarily* strictly proper, since if $\bar{q}(z) = \bar{p}(z)$, this does not neccessarily mean equality of $p(x)$ and $q(x)$. Although $\bar{S}$ is not a score in the original definition of Section 3, it is clearly a nonlocal quantity.

6 Conclusions

Insightful evaluation and inter-comparison of probability forecasts requires a careful choice of score to quantify the agreement between historical forecast-verification pairs. We focus on a few scores for the case that each forecast consists of a probability density function and each verification consists of a real

[4]Scores including the derivative at verification points are still nonlocal according to the common definitions, although they could be attested a certain "pseudo–locality"

[5]Requirements of smoothness or parsimony might be desired for reasons not directly connected with skill, and therefore might not be considered as part of the score. We thank Devin Kilminster for stressing this point.

[6]The exact statement of this result is that every local, smooth and proper score for continuous variables is an affine function of the Ignorance

number. This list of scores is not exhaustive. Furthermore, probabilistic forecasts for discrete events allow for further measures of skill not mentioned here. Our main point is that only proper scores are internally consistent. Another property of scores, locality, is discussed. Several scores are examined in this light. By Bernardo's theorem, Ignorance is effectively the only proper local score for continuous variables. Locality also appears to be a desirable property of a score, yet the case for local scores is less compelling than for proper scores. It would be interesting to identify and investigate when nonlocal scores for continuous variables would be highly valued.

When using scores to evaluate probabilistic forecasting systems it is critical to consider the performance of the system over a duration sufficiently long to obtain robust results. Ultimate evaluation of operational probabilistic forecast systems may require including the fact that the verifying observation is itself uncertain, and thus a move to generalised scores. A possible generalisation was motivated and discussed in the context of propriety and locality. Proper scores allow an internally consistent evaluation, making their use an important feature in the valuation and further improvement of these forecasts and the models behind them.

Acknowledgements

This work was supported by the DIME EPSRC/DTI Faraday Partnership under grant GR/R92363/01; ENSEMBLES and the National Oceanographic and Atmospheric Administration (NOAA) under grant 1-RAT-S592-04001. Furthermore, the authors gratefully acknowledge fruitful discussions with Liam Clarke, Devin Kilminster, Antje Weisheimer, and Kevin Judd.

References

J. M. Bernardo. Expected information as expected utility. *Annals of Statistics*, 7(7):686–690, 1979.

G. Candille and O. Talagrand. Evaluation of probabilistic prediction systems for a scalar variable. *Quaterly Jornal of the Royal Meteorological Society*, pages 1–23, 2004.

Tilmann Gneiting and Adrian Raftery. Strictly proper scoring rules, prediction, and estimation. Technical Report 436, Department of Statistics, University of Washington, 2004.

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

Ian T. Jolliffe and David B. Stephenson. *Forecast Verification*. Wiley, 2003.

Richard Kleeman. Measuring dynamical prediction utility using relative entropy. *Journal for the Athmospheric Sciences*, 59:2057–2072, 2002.

S. Kullback and R. A. Leibler. *On information and sufficiency*. Number 22(1) in Annals of Mathematical Statistics. McGraw Hill, 1 edition, March 1951.

A. H. Murphy and R. L. Winkler. A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338, Jul 1987.

T.N. Palmer. Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics*, 63(2):71–116, February 2000.

T.N. Palmer, G.J. Shutts, R. Hagedorn, F. Doblas-Reyes, T. Jung, and Leutbecher M. Representing model uncertainty in weather and climate prediction. *Annual Review of Earth and Planetary Sciences*, 33:163–193, 2005.

S. Petterssen. *Weather Analysis and Forecasting*. McGraw Hill, London, second edition, 1956.

M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(130):1653–1660, 2002.

Reinhard Selten. Axiomatic characterisation of the quadratic scoring rule. *Experimental Economics*, 1:43–62, 1998.

Daniel S. Wilks. *Statistical Methods in the Athmospheric Sciences*, volume 59 of *International Geophysics Series*. Academic Press, first edition, 1995.

L.J. Wilson, W.R. Burrows, and A. Lanzinger. A strategy for verification of weather element forecasts from an ensemble prediction system. *Monthly Weather Review*, 127(6):956–970, 1999.