

## Scoring Probabilistic Forecasts: The Importance of Being Proper

JOCHEN BRÖCKER

*Centre for the Analysis of Time Series, London School of Economics, London, United Kingdom*

LEONARD A. SMITH

*Centre for the Analysis of Time Series, London School of Economics, London, and Pembroke College, Oxford University, Oxford, United Kingdom*

(Manuscript received 4 November 2005, in final form 23 May 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 long-standing aim in meteorological forecasting. This paper explains the importance of employing proper scores when selecting between 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 independent of any attempt to influence the behavior of a forecaster. Another property of scores (i.e., 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 generalization 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” (Petersen 1956). Modern telecommunications allow the user of weather model output to construct weather forecasts using simulations from several operational centers. 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 future 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 behavior of forecasters, as ar-

---

*Corresponding author address:* Jochen Bröcker, Centre for the Analysis of Time Series, London School of Economics, Houghton St., London WC2A 2AE, United Kingdom.  
E-mail: cats@lse.ac.uk

gued 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 generalized 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 lowercase  $x$  denotes any possible value in the range of  $X$ . In the case of London Heathrow temperature,  $x$  could be any real number larger than  $-273^\circ\text{C}$ . In this paper we focus on probabilistic forecasts in the form of probability density functions (PDFs)  $p(x)$ , 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  $p(x)$ ,  $q(x)$ ,  $r(x)$ , etc.

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

$$\int p(x) dx = 1. \quad (1)$$

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 an 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 neither depends 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  $\{p_i(x), i = 1, \dots, N\}$  and corresponding verifications  $\{X_i, i = 1, \dots, N\}$ . 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.

### a. The ignorance score

The ignorance score (Good 1952; Roulston and Smith 2002) is defined by

$$S[p(x), X] = -\log[p(X)]. \quad (2)$$

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 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.

### b. The Brier score

The Brier score (Candille and Talagrand 2005; 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

$$S(p, X) = (X - p)^2, \quad (3)$$

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

### c. The naive linear score

The naive linear score applies to continuous  $X$  and is defined as

$$S[p(x), X] = -p(X). \quad (4)$$

Although often suggested as a possible score, the naive linear score is not proper, as will be demonstrated and discussed in section 4.

### d. The proper linear score

This score applies to continuous  $X$  and is defined as

$$S[p(x), X] = \int p^2(z) dz - 2p(X). \quad (5)$$

It is a strictly proper alternative to the naive linear score in Eq. (4). The fact that the additional term  $\int p^2(z) dz$  renders the score strictly proper will be demonstrated in section 4. Selten (1998) discussed it and contrasted it with the ignorance score.

### e. The mean square error

This score can be applied to continuous  $X$  with the following definition:

$$S[p(x), X] = \int (X - z)^2 p(z) dz. \quad (6)$$

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,<sup>1</sup> that is,

$$m = \int xp(x) dx$$

and

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

this score can be written as

$$S[p(x), X] = (X - m)^2 + s^2. \quad (7)$$

Thus, the mean square error depends on  $p(x)$  only through its first and second moment. It does not reflect

<sup>1</sup> 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.

any other aspect of  $p(x)$ . The implications of this will be discussed later.

Note that the proper linear score depends on the entire functional form of  $p(x)$  [due to the integral in the first term of Eq. (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  $q(x)$

$$\int S[p(x), z]q(z) dz \geq \int S[q(x), z]q(z) dz. \quad (8)$$

In other 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 Eq. (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 [viz.  $q(x)$ ]. This can be written as

$$\text{forecasted skill of } q(x): \int S[q(x), z]q(z) dz.$$

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

$$\text{forecasted skill of } p(x): \int S[p(x), z]q(z) dz.$$

This is the integral on the left-hand side of Eq. (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 necessary 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 are considered.

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

$$\int S[p(x), z]q(z) dz = \int S[q(x), z]q(z) dz + \left\{ \int S[p(x), z]q(z) dz - \int S[q(x), z]q(z) dz \right\}. \tag{9}$$

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 [Eq. (9)] is well known as the *reliability–sharpness* decomposition (Wilks 1995). To show this, we write the Brier score as

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

because  $E_q(X) = q$ , 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, Eq. (10) becomes the same decomposition as Eq. (9). This shows also that the Brier score is strictly proper, because the parenthesized term in Eq. (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

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

with equality if and only if  $p(x) = q(x)$  (i.e., Kullback–Leibler inequality; Kullback and Leibler 1951). The proper linear score is indeed also strictly proper, given that

$$\int [q(z) - p(z)]^2 dz \geq 0, \tag{12}$$

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

$$\begin{aligned} \int [q(z)^2 + p(z)^2 - 2p(z)q(z)] dz &= \int q(z)^2 dz \\ &+ \int S[p(x), z]q(z) dz, \end{aligned} \tag{13}$$

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

The naive linear score however is *improper*: even if  $X$  were drawn from  $q(x)$ , the naive linear score would not judge  $q(x)$  the best. There are PDFs different from  $q(x)$ , which would rank higher than  $q(x)$ . In fact, for any given  $q(x)$  it is always possible to find a  $p(x)$  so that

$$\int -p(z)q(z) dz \leq \int -q(z)q(z) dz.$$

Actually, the naive linear score favors a  $p(x)$  featuring a very small spread and which is centered at a certain 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

$$-q(\bar{x}) < \int -q(z)q(z) dz = \int S[q(x), z]q(z) dz.$$

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 normalized and define

$$p_\sigma(x) = \frac{1}{\sigma} g\left(\frac{x - \bar{x}}{\sigma}\right),$$

that is, center the kernel at  $\bar{x}$  with spread  $\sigma$ , it follows that

$$\int S[p_\sigma(x), z]q(z) dz = \int -p_\sigma(z)q(z) dz \rightarrow -q(\bar{x}).$$

In other words, the naive 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

$$\int -p(z)q(z) dz = -q$$

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

The nonpropriety of the naive linear score would also emerge as a consequence of a far more general result due to Bernardo (1979; see also section 5), 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 characterized 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 in Eq. (7) we have

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

But this quantity is *not necessarily* larger than

$$\int S[q(x), z]q(z) dz = 2s_q^2,$$

as it would have to be for the mean square error to be proper. In fact, as for the naive linear score, a density  $p(x)$  centered 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  $\int (z - m_p)^2 q(z) dz$  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 nonzero 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  $X$ . Because 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 nonlocality

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 PDF 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.<sup>2</sup>

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.<sup>3</sup> This can also be useful when scores are employed for training models that translate numerical simulations into probability density forecasts. A ubiq-

<sup>2</sup> Note that a similar definition for nonlocal scores would require a substantially more advanced concept of smoothness, because, in general, nonlocal scores involve functional operations.

<sup>3</sup> 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.

uitous problem here is to limit model complexity, which can be addressed by enforcing certain measures of smoothness upon the PDFs (regularization). Because 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 considered class of density functions to smooth functions a priori, or by augmenting the score with a term that penalizes nonsmooth densities, essentially rendering the score nonlocal.<sup>4,5</sup>

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 that 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),<sup>6</sup> 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

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

Applying a score  $S$  to  $\bar{p}(z)$  and  $Z$ , we can define the *generalized 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 generalized score to be proper if for any  $q(x)$  we have

$$\int \bar{S}[p(x), z]\bar{q}(z) dz \geq \int \bar{S}[q(x), z]\bar{q}(z) dz, \quad (14)$$

<sup>4</sup> Scores including the derivative at verification points are still nonlocal according to the standard definition of locality, although they could be attested a certain “pseudolocality.”

<sup>5</sup> 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 D. Kilminster for stressing this point.

<sup>6</sup> The exact statement of this result is that every local, smooth, and proper score for continuous variables is an affine function of the ignorance.

where, as for  $p(x)$ ,

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

If  $S$  is proper,  $\bar{S}$  is proper as well. If  $S$  is strictly proper though,  $\bar{S}$  is *not necessarily* strictly proper, because if  $\bar{q}(z) = \bar{p}(z)$ , this does not necessarily 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 intercomparison of probability forecasts requires a careful choice of score to quantify the agreement of historical forecast–verification pairs. We focus on a few scores for the case that each forecast consists of a PDF and each verification consists of a real 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. 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 generalized scores. A possible generalization 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.

*Acknowledgments.* This work was supported by the DIME EPSRC/DTI Faraday Partnership under Grant GR/R92363/01; and 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

Bernardo, J. M., 1979: Expected information as expected utility. *Ann. Stat.*, **7**, 686–690.

- Candille, G., and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.*, **131**, 2131–2150.
- Gneiting, T., and A. Raftery, 2004: Strictly proper scoring rules, prediction, and estimation. Tech. Rep. 436, Department of Statistics, University of Washington.
- Good, I. J., 1952: Rational decisions. *J. Roy. Stat. Soc.*, **XIV**, 107–114.
- Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, 256 pp.
- Kleeman, R., 2002: Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.*, **59**, 2057–2072.
- Kullback, S., and R. A. Leibler, 1951: On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86.
- Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338.
- Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.*, **63** (2), 71–116.
- , G. J. Shutts, R. Hagedorn, F. Doblas-Reyes, T. Jung, and M. Leutbecher, 2005: Representing model uncertainty in weather and climate prediction. *Annu. Rev. Earth Planet. Sci.*, **33**, 163–193.
- Petterssen, S., 1956: *Weather Analysis and Forecasting*. 2d ed. McGraw Hill, 505 pp.
- Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660.
- Selten, R., 1998: Axiomatic characterisation of the quadratic scoring rule. *Exp. Econ.*, **1**, 43–62.
- Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. International Geophysics Series, Vol. 59, Academic Press, 464 pp.
- Wilson, L. J., W. R. Burrows, and A. Lanzinger, 1999: A strategy for verification of weather element forecasts from an ensemble prediction system. *Mon. Wea. Rev.*, **127**, 956–970.