# On Using ‘‘Climatology’’ as a Reference Strategy in the Brier and Ranked Probability Skill Scores

SIMON J. MASON

*International Research Institute for Climate Prediction, Columbia University, Palisades, New York*

29 October 2003 and 26 January 2004

ABSTRACT

The Brier and ranked probability skill scores are widely used as skill metrics of probabilistic forecasts of weather and climate. As skill scores, they compare the extent to which a forecast strategy outperforms a (usually simpler) reference forecast strategy. The most widely used reference strategy is that of ‘‘climatology,’’ in which the climatological probability (or probabilities in the case of the ranked probability skill score) of the forecast variable is issued perpetually. The Brier and ranked probability skill scores are often considered harsh standards. It is shown that the scores are harsh because the expected value of these skill scores is less than 0 if nonclimatological forecast probabilities are issued. As a result, negative skill scores can often hide useful information content in the forecasts. An alternative formulation of the skill scores based on a reference strategy in which the outcome is independent of the forecast is equivalent to using randomly assigned probabilities but is not strictly proper. Nevertheless, positive values of the Brier skill score with random guessing as a strategy correspond to positive-sloping reliability curves, which is intuitively appealing because of the implication that the conditional probability of the forecast event increases as the forecast probability increases.

The Brier score is a quadratic measure of error in probabilistic forecasts (Brier 1950). Although the Brier score can be used in multievent situations, it is most commonly used in a dichotomous situation in which an event of interest either occurs or does not occur (Toth et al. 2003). The ranked probability score is a closely related measure that generalizes the Brier score to a multievent situation, but in which the events can be ordered (Epstein 1969; Murphy 1969, 1971). Both scores are measures of the accuracy of the forecast in terms of the probability (or probabilities in the case of the ranked probability score) assigned (Murphy 1993). The scores are widely expressed as skill scores, by which they compare the extent to which a forecast strategy outperforms a (usually simpler) reference forecast strategy. The most widely used reference strategy is that of ‘‘climatology,’’ in which the climatological probability/probabilities of the forecast variable is/are issued perpetually. Skill scores on both measures are widely reported to be low compared to other performance indicators (Wilks 1995), and so these skill scores are often considered harsh standards. While low scores can partly be attributed to sampling errors in the forecast probabilities, most notably when ensemble sizes are small (Kumar et al. 2001), a more fundamental reason for the often negative skill indicated by the scores is detailed in this note. The following discussion refers only to the Brier score, but the conclusions can easily be generalized to the ranked probability skill score.

The average Brier score, BS, for a set of $n$ forecasts is defined as

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2, \tag{1}$$

where $f_i$ is the forecast probability for the $i$th forecast, and $o_i$ is the $i$th outcome, with $o_i = 1$ if the event occurs and $o_i = 0$ otherwise (Brier 1950; Wilks 1995). A commonly used decomposition of the average Brier score is given as

$$\mathrm{BS} = \overline{o}(1 - \overline{o}) + \frac{1}{n}\sum_{k=1}^{m} n_k(f_k - \overline{o}_k)^2 - \frac{1}{n}\sum_{k=1}^{m} n_k(\overline{o}_k - \overline{o})^2 \tag{2}$$

(Murphy 1973). Here $\overline{o}$ represents the climatological probability of the event, and $m$ represents either the number of distinct probabilities with which forecasts are issued or the number of probability bins into which the

*Corresponding author address:* Dr. Simon J. Mason, International Research Institute for Climate Prediction, 61 Route 9W, P.O. Box 1000, Columbia University, Palisades, NY 10964-8000.
E-mail: simon@iri.columbia.edu

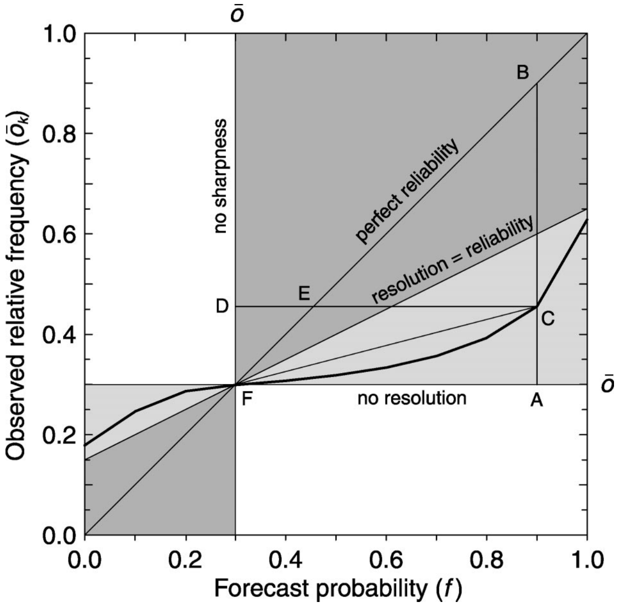

FIG. 1. Attributes diagram showing the areas of skill compared to forecasts of climatology (dark shading) and additional areas of skill compared to random guessing (light shading). The prior probability of the event, $\bar{o}$, is arbitrarily set at 0.3.

forecasts are grouped ($m \leq n$). The forecast probability for bin $k$ is given as $f_k$, and the relative frequency of occurrence of the event when the forecast probability is $f_k$ is denoted $\bar{o}_k$. These three terms can be abbreviated as

$$\mathrm{BS} = \mathrm{UNC} + \mathrm{REL} - \mathrm{RES}, \tag{3}$$

representing uncertainty, reliability, and resolution, respectively. The uncertainty term is independent of the forecasts, being a function only of the inherent uncertainty of the event being forecast, and so it is the reliability and resolution terms that determine the forecast performance.

The components of the Brier score can be illustrated on the attributes diagram (Hsu and Murphy 1986). In Fig. 1 an example is presented in which the thick line represents the empirical curve for an arbitrary set of forecasts for an event with a climatological probability of 0.3. Each point on the curve is defined by the coordinates ($f_k$, $\bar{o}_k$). From Eq. (2), the contributions to the resolution term are represented by the squared vertical distances between the empirical curve and the climatological probability. For a specific point on the curve, the contribution to the resolution term is therefore equivalent to the square of the distance between C and A. These distances are weighted by the number of forecasts represented by the point, $n_k$, and then averaged across all points on the curve. The contributions to the reliability term are represented by the squared vertical distances between the diagonal line of perfect reliability (where $\bar{o}_k = f_k$) and the empirical curve. For the same point, the contribution to the reliability term is therefore equivalent to the square of the distance between C and B. Alternatively, because the diagonal line of perfect reliability is at 45°, the contribution to the reliability term is also equivalent to the squared horizontal distance between the diagonal line of perfect reliability and the empirical curve (between C and E). Thus CB and CE are indicative of two alternative interpretations of the reliability measure: reliability measures the mean squared error between the observed relative frequency and the frequency implied by the forecasts (of which CB is one component), or reliability measures the mean squared error between the actual forecast probability and the probability that should have been assigned (of which CE is one component).

The Brier skill score, BSS, is defined as

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}} \tag{4}$$

(Wilks 1995; Toth et al. 2003). The score represents the level of improvement of the Brier score compared to that of a reference forecast strategy, $\mathrm{BS}_{\mathrm{ref}}$, and is designed to range from 1.0 for a perfect forecast strategy, through 0.0 for one that provides no improvement over the reference strategy, to negative values for strategies that are worse than the reference strategy. (The lower bound is achieved for a set of perfectly bad forecasts, but does not necessarily give a score of $-1.0$, and so negative skill scores need to be interpreted with caution.)

The most widely used reference strategy for calculating the Brier skill score is that of ‘‘climatology,’’ in which the climatological probability of the forecast variable is issued perpetually. Climatology is an appealing reference strategy because it is intended to provide an indication of whether the forecasts are better than having no forecast information at all, apart from knowledge of the historical likelihood of the event (it is usually assumed that the prior probability of the event is stationary). In addition, by combining Eqs. (2) and (4), the Brier skill score simplifies conveniently because both the resolution and reliability terms are 0 for forecasts of climatological probabilities: the resolution term disappears because all forecasts are for the same probability ($m = 1$ and $n_k = n$, and so $\bar{o}_k = \bar{o}$), while the reliability term disappears because the forecast probability for all forecasts equals the observed relative frequency ($f_k = \bar{o}_k$, assuming that $\bar{o}_k$ is stationary). The Brier skill score relative to forecasts of climatological probabilities therefore reduces to

$$\mathrm{BSS}_{\mathrm{clim}} = \frac{\mathrm{RES} - \mathrm{REL}}{\mathrm{UNC}} \tag{5}$$

(Wilks 1995; Toth et al. 2003). Hsu and Murphy (1986) demonstrate that, according to Eq. (5), skill is indicated whenever RES > REL, and so the empirical curve on the attributes diagram needs to be steeper than the reliability = resolution line on Fig. 1. By this reckoning, areas where skill is indicated are shaded dark.

It is informative to calculate the expected value of the Brier skill score. From Eq. (4) and from the fact that the Brier score for climatological forecasts reduces to the uncertainty term, the expected Brier skill score is

$$E(\mathrm{BSS}_{\mathrm{clim}}) = 1 - \frac{E(\mathrm{BS})}{\mathrm{UNC}}. \tag{6}$$

The expected value of the Brier score [the numerator in the right-hand side of Eq. (6)] can be determined from the expected relative frequencies of contributions to the score when the forecast event verifies, compared to when it does not verify. These relative frequencies are defined by the climatological probability of the event, and are $\overline{o}$ and $1 - \overline{o}$, respectively. When the forecast event verifies, the Brier score for each forecast is $(f_i - 1)^2$, otherwise the score is $f_i^2$. Hence,

$$\begin{aligned} E(\mathrm{BSS}_{\mathrm{clim}}) &= 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}[\overline{o}(f_i - 1)^2 + (1 - \overline{o})f_i^2]}{\overline{o}(1 - \overline{o})} \\ &= -\frac{\frac{1}{n}\sum_{i=1}^{n}(f_i - \overline{o})^2}{\overline{o}(1 - \overline{o})}, \end{aligned} \tag{7}$$

which states that the expected Brier skill score is a function of the forecast probabilities. This dependence of the expected score on the forecasts is atypical of a number of other commonly used skill scores such as the hit skill score (Wilks 1995), linear error in probability space (LEPS; Ward and Folland 1991; Potts et al. 1996), and the Gerrity score (Gerrity 1992; Livezey 2003). It has some important implications that can best be indicated by expressing the forecast probability in terms of its departure, $d_i$, from the climatological probability ($f_i = \overline{o} + d_i$). Equation (7) then simplifies to

$$E(\mathrm{BSS}_{\mathrm{clim}}) = \frac{-\frac{1}{n}\sum_{i=1}^{n} d_i^2}{\overline{o}(1 - \overline{o})}. \tag{8a}$$

Because the numerator of Eq. (8) defines the variance of a set of forecasts about the climatological probability of the event, forecast systems with greater variance in the forecast probabilities will have a lower (more negative) expected Brier skill score than those with smaller variance. For any single forecast Eq. (8a) simplifies to

$$E(\mathrm{BSS}_{\mathrm{clim}}) = \frac{-d_i^2}{\overline{o}(1 - \overline{o})}, \tag{8b}$$

which, since $d_i^2 \geq 0$, implies that the expected Brier skill score, with climatology as the reference forecast strategy, is less than 0 for any forecast that differs from the climatological probability. There are two important implications of Eq. (8): the expected Brier skill score can be optimized by issuing climatological forecast probabilities, and the forecast may contain some potentially usable information even when $\mathrm{BSS}_{\mathrm{clim}}$ is less than 0.

The fact that the expected Brier skill score can be optimized by issuing climatological forecast probabilities does not imply that the skill score is improper. Proper scoring rules are those that are optimized when the forecast probability corresponds to the forecaster's true belief in the probability of an event occurring, and strictly proper scoring rules are those for which no other strategy provides an equally optimal score (Murphy and Epstein 1967). It can be demonstrated that $\mathrm{BSS}_{\mathrm{clim}}$ is a strictly proper score by considering the expected Brier skill score for a case, $i$, for which the forecaster's true belief in the probability of the event occurring is $p_i$, and then differentiating with respect to the issued forecast probability, $f_i$:

$$\frac{\partial E(\mathrm{BSS}_{\mathrm{clim}})}{\partial f_i} = \frac{2(p_i - f_i)}{\overline{o}(1 - \overline{o})}. \tag{9}$$

The skill score can be optimized by equating Eq. (9) to 0, which occurs only when $p_i = f_i$, indicating that the score is strictly proper. By optimizing the individual contributions to the skill score, as implied by Eq. (9), the skill score for a set of forecasts, with varying $p_i$ and $f_i$, is also optimized.

That the Brier skill score with climatology as the reference strategy is a strictly proper scoring rule [Eq. (9)] is not inconsistent with the fact that the expected value of the score can be optimized by repeatedly issuing the climatological probability [Eq. (8)]. Equation (8) applies only in the absence of any reason for expecting the forecast event to be more or less likely than usual. In this instance the forecaster should issue the climatological probability as the forecast in preference to any other strategy (such as assigning all probability to one specific outcome, or randomly assigning probabilities). In contrast, a number of other skill scores have an expected score of 0 for all naïve forecast strategies, and the forecaster is effectively free to choose any of these strategies. These scores have the property of equitability (Gandin and Murphy 1992; Mason 2003), which $\mathrm{BSS}_{\mathrm{clim}}$ lacks. However, in the specific context of the Brier skill score, the lack of equitability may be a desirable feature: a nonclimatological forecast should imply that the forecaster believes that the probability of the event is different from normal, and where that implied belief is unfounded the forecaster is penalized by the Brier score. Specifically, it can be shown that perpetual forecasts of nonclimatological values give a Brier skill score that is equal to the negative of the squared departure from the climatological probability divided by the uncertainty [Eq. (8)].

Although climatology is appropriately the best forecast strategy when the forecaster has no meaningful posterior information, the fact that Eq. (8) is negative for nonclimatological forecasts ($d_i \neq 0$) means that the forecast may contain some potentially usable information even when $\mathrm{BSS}_{\mathrm{clim}}$ is less than 0. It could be argued that if the relative frequency of the outcome is conditional upon the forecast, even if reliability and resolution are far from perfect, then the forecast does contain usable information. In fact, as long as there is some resolution then the forecasts are potentially usable since all that is required is a recalibration to make the forecasts reliable (Murphy 1966), although the additional constraint that resolution is monotonically (and positively) related to forecast probability would be desirable. For a forecast strategy that issues nonclimatological forecast probabilities it therefore makes sense to compare the forecasts to a strategy in which the observed relative frequency is independent of the forecast probability. Then either the outcomes can be seen as random occurrences or the forecasts can be viewed as being randomly shuffled, while in both cases the marginal distribution of the forecast probabilities remains unchanged. In this case, although there are still $m > 1$ probability bins, $\overline{o}_k = \overline{o}$ for all $k$, and so, as with forecasts of climatological probabilities, the resolution term is 0. However, the reliability term now becomes nonzero:

$$\mathrm{REL}_{\mathrm{ran}} = \frac{1}{n}\sum_{k=1}^{n} n_k(f_k - \overline{o})^2. \qquad (10)$$

Equation (10) defines the mean squared departure of the forecast probabilities from the climatological probability, which could be considered a measure of the sharpness of the forecasts[1] (Wilks 1995). So the reliability term for random forecasts becomes equal to the sharpness term (denoted SHP), and the Brier skill score with random forecasts as the reference strategy, $\mathrm{BSS}_{\mathrm{ran}}$, then becomes

$$\mathrm{BSS}_{\mathrm{ran}} = \frac{\mathrm{SHP} + \mathrm{RES} - \mathrm{REL}}{\mathrm{SHP} + \mathrm{UNC}}. \qquad (11)$$

This revised skill score involves simple adjustments to Eq. (5) for forecast sharpness and is equitable (as shown below), unlike Eq. (5).

Since $\mathrm{SHP} + \mathrm{UNC} > 0$, skill is indicated whenever $\mathrm{SHP} + \mathrm{RES} > \mathrm{REL}$. The area of skill, as defined by Eq. (11) can be indicated on the attributes diagram, but first the distance represented by the sharpness term needs to be identified. From Eq. (10), the sharpness term is represented by the squared horizontal distance between the empirical curve and the climatological probability. For an arbitrary point on the curve, the contribution to the sharpness term is therefore equivalent to the square of the distance between C and D. If the distance $(\mathrm{CD})^2$, and hence $(\mathrm{AF})^2$, is equal to the sharpness term, and $(\mathrm{AC})^2$ to the resolution, then the distance $\mathrm{CF}^2$ equals SHP + RES. With reliability represented by the distance $(\mathrm{CE})^2$, skill is indicated wherever $\mathrm{CF} > \mathrm{CE}$, which is true for all points below the no-resolution line to the left of the climatological probability and above the no-resolution line to the right (light and dark shaded areas of Fig. 1). In effect, therefore, skill is indicated relative to random guessing whenever the slope of the reliability curve is positive. That a positively sloping reliability curve indicates positive skill is intuitively appealing since it indicates that the probability of the event occurring does increase (by however small an amount) as the forecast probability increases.

The equitability of Eq. (11) can be demonstrated in the same manner as in Eq. (7), that is, by calculating the expected value of the score

$$E(\mathrm{BSS}_{\mathrm{ran}}) = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{[\overline{o}(f_i - 1)^2 + (1 - \overline{o})f_i^2]}{(f_i - \overline{o})^2 + \overline{o}(1 - \overline{o})} = 0. \qquad (12)$$

Equation (12) indicates that the expected value of the Brier skill score with random probabilities as the reference strategy is 0, regardless of the forecast. Hence, any naïve forecasting strategy will give a 0 score, and, more pertinently, any forecast with at least some useful information will give a positive skill score. However, it remains to be asked whether $\mathrm{BSS}_{\mathrm{ran}}$ is a strictly proper scoring rule. As for Eq. (9), let $p_i$ represent the forecaster's true belief in the probability of the event occurring. From Eqs. (4) and (11), the expected Brier skill score is

$$\begin{aligned} E(\mathrm{BSS}_{\mathrm{ran}}) &= 1 - \frac{E(\mathrm{BS})}{\mathrm{SHP} + \mathrm{UNC}} \\ &= 1 - \frac{p_i(f_i - 1)^2 + (1 - p_i)f_i^2}{(f_i - \overline{o})^2 + \overline{o}(1 - \overline{o})} \\ &= 1 - \frac{f_i^2 - 2f_ip_i + p_i}{f_i^2 - 2f_i\overline{o} + \overline{o}}. \end{aligned} \qquad (13)$$

Equation (13) is maximized when its first derivative is set to 0:

$$\frac{\partial E(\mathrm{BSS}_{\mathrm{ran}})}{\partial f_i} = \frac{2f_i(1 - f_i)(p_i - \overline{o})}{(f_i^2 - 2f_i\overline{o} + \overline{o})^2} = 0$$
$$\Rightarrow f_i = \{0, 1\}, \qquad (14)$$

which indicates that the skill score can be optimized by hedging probabilities toward deterministic forecasts of the event (i.e., $f_i = 0$ or $f_i = 1$). Thus the gain in equitability is at the loss of propriety, which is a serious drawback and effectively precludes its use in most contexts (Jolliffe and Stephenson 2003). In combination with Eq. (5), however, Eq. (11) may have some value in that it does provide positive scores when the forecasts have some information content. Although information

[1] Sharpness is difficult to define adequately [see discussion by Potts (2003)], but for unbiased forecasts $\overline{f} = \overline{o}$, and so Eq. (10) becomes

$$\mathrm{REL}_{\mathrm{ran}} = \frac{1}{n}\sum_{k=1}^{m} n_k(f_k - \overline{f})^2 = \frac{1}{n}\sum_{i=1}^{n}(f_i - \overline{f})^2,$$

which is the variance of the forecast probabilities. This definition is consistent with that of Murphy and Winkler (1992).

content can be provided by the resolution score (which is strictly proper), this term cannot distinguish between positive and negative sloping reliability curves.

In conclusion, it is recommended that the Brier skill score with climatology as the reference forecast strategy [Eq. (5)] not be used as a lone measure of forecast skill because of the possibility that negative skill scores may hide the fact that the forecast system does contain useful information, especially if the sharpness of the forecasts is high. A similar recommendation can be made for the ranked probability skill score because of its simple relationship to the Brier skill score. This weakness of the widely used version of the Brier skill score is a side effect of its lack of equitability. Although from some perspectives equitability is likely to be an undesirable feature of a probabilistic scoring rule, the downside of the lack of equitability is that the information content of nonclimatological forecast probabilities may be discarded under the somewhat arbitrary condition of resolution being less than reliability. The possibility of using randomly assigned probabilities with the same marginal distribution as the set of forecast probabilities under consideration as a reference strategy instead of climatological probabilities was considered [Eq. (11)]. By explicitly considering the sharpness of the probabilities issued, this skill score has the desirable feature of having an expected score of 0 when nonclimatological forecast probabilities are issued, but at the loss of propriety. Regardless of which measures are used, this paper has highlighted the need to consider a set of scoring measures because of inherent weaknesses in any single measure of forecast performance.

*Acknowledgments.* This note was funded by Cooperative Agreement AN07GP0213 from the National Oceanic and Atmospheric Administration (NOAA). The views expressed herein are those of the author and do not necessarily reflect the views of NOAA or any of its subagencies. The comments of L. Goddard, A. G. Barnston, and anonymous referees are gratefully acknowledged, as are valuable discussions with F. J. Doblas-Reyes, L. Ferranti, A. Ghelli, R. Hagedorn, F. Vitart, and M. S. J. Harrison.

## REFERENCES

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78,** 1–3.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8,** 985–987.

Gandin, L. S., and A. H. Murphy, 1992: Equitable scores for categorical forecasts. *Mon. Wea. Rev.*, **120,** 361–370.

Gerrity, J. P., 1992: A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev.*, **120,** 2707–2712.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting,* **2,** 285–293.

Jolliffe, I. T., and D. B. Stephenson, 2003: Introduction. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 1–12.

Kumar, A., A. G. Barnston, and M. P. Hoerling, 2001: Seasonal predictions, probabilistic verifications, and ensemble size. *J. Climate,* **14,** 1671–1676.

Livezey, R. E., 2003: Categorical events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 77–96.

Mason, I. T., 2003: Binary events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 37–76.

Murphy, A. H., 1966: A note on the use of probabilistic predictions and the probability score in the cost-loss ratio decision situation. *J. Appl. Meteor.*, **5,** 534–537.

——, 1969: On the ''ranked probability score.'' *J. Appl. Meteor.,* **8,** 988–989.

——, 1971: A note on the ranked probability score. *J. Appl. Meteor.,* **10,** 155–156.

——, 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

——, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

——, and E. S. Epstein, 1967: A note on probability forecasts and ''hedging.'' *J. Appl. Meteor.,* **6,** 1002–1004.

——, and R. L. Winkler, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting,* **7,** 435–455.

Potts, J. M., 2003: Basic concepts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 13–36.

——, C. K. Folland, I. T. Jolliffe, and D. Sexton, 1996: Revised ''LEPS'' scores for assessing climate model simulations and long-range forecasts. *J. Climate,* **9,** 34–53.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 137–163.

Ward, N. M., and C. K. Folland, 1991: Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea surface temperatures. *Int. J. Climatol.,* **11,** 711–743.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.