Verification of Categorical Probability Forecasts

H. ZHANG AND T. CASEY

Bureau of Meteorology Research Centre, Melbourne, Victoria, Australia

(Manuscript received 11 January 1999, in final form 2 September 1999)

ABSTRACT

This paper compares a number of probabilistic weather forecasting verification approaches. Forecasting skill scores from linear error in probability space and relative operating characteristics are compared with results from an alternative approach that first transforms probabilistic forecasts to yes/no form and then assesses the model forecasting skill. This approach requires a certain departure between the categorical probability from forecast models and its random expectation. The classical contingency table is revised to reflect the ''nonapplicable'' forecasts in the skill assessment.

The authors present a verification of an Australian seasonal rainfall forecast model hindcasts for the winter and summer seasons over the period from 1900 to 1995. Overall skill scores from different approaches demonstrate similar features. However there are advantages and disadvantages in each of those approaches. Using more than one skill assessment scheme is necessary and is also of practical value in the evaluation of the model forecasts and their applications.

1. Introduction

Seasonal forecasts of climate variables such as rainfall and temperature are often presented as a probability of occurring within a certain category such as above or below average (two categories), or above, near, and below average (three-category tercile forecasts). The probability of occurrence in each forecast category is usually expressed as a percentage probability figure with the total probability in all categories adding to 100%. A significant shift of the probability away from its average is indicated by corresponding changes in the probabilities in other categories.

The ostensible reason for providing a probability value is because probabilistic forecasts have the advantage that they can convey the uncertainty associated with the forecasts in a quantitative way (Murphy 1977). However, any probability exceeding the average could be interpreted as meaning an event is likely to occur, when in fact a discrete amount of departure may need to be achieved between the forecast probability and the mean expectation value before any confidence can be attached to the forecast. It has been acknowledged that users of climate information will generally not alter their practices unless there is a significant shift in probabilities away from normal conditions (Hammer et al. 1996).

Basically, there are two ways of scoring probabilistic forecasts. One is to use some measure of the departure between forecasts and observations and the other is to do a conversion from the probabilistic form to a binary yes/no form to use a contingency table to score bit rates against misses and false alarm rates. Many skill scoring schemes have been developed and used in the verification of probabilistic forecasts in meteorology and comprehensive reviews appear in the literature (e.g., Bettge et al. 1981; Doswell and Flueck 1989; Wilks 1995). Generally, most scoring methods can be categorized as follows:

1) those that directly measure the departures of the forecasts from the actual observations, such as the root-mean-square error (rmse) or Brier scores (e.g., Stanski et al. 1989);
2) those that measure the departure between forecasts and observations in cumulative probability space such as the ranked probability score, or the linear error in probability space (LEPS) score (Ward and Folland 1991; Potts et al. 1996);
3) relative operating characteristics (ROC), which are based on signal detection theory and attempt to measure the relative ''signal'' and ''noise'' ratios contained in forecast information in the form of hits to misses ratios when measured against performance level (Egan 1975; Mason 1982); and
4) evaluation scores based on converting probability forecasts to binary (yes/no) forecasts and the generation of a contingency table from the hit and miss rates (Mason 1979; Gandin and Murphy 1992).

In this paper we present the comparative results of

Corresponding author address: Dr. H. Zhang, Bureau of Meteorology Research Centre, GPO Box 1289K, VIC 3001, Australia.
E-mail: h.zhang@bom.gov.au

TABLE 1. Classical 2 × 2 contingency table structure.

	Forecast: yes	Forecast: no
Observation: yes	A	B
Observation: no	C	D

applying a number of these different skill scores to a set of Australian seasonal rainfall hindcasts. In the next section we discuss a method of using probability thresholds to transform probabilistic forecasts to yes, no, and nonapplicable forecasts, and then assess the forecasting skill based on a revised contingency table. Section 3 presents the application of LEPS, ROC, and revised true skill statistic (TSS) scores to tercile categorical forecasts. In section 4 we present a comparative assessment of the scoring methods and an ensemble experiment to assess the statistical significance of some of the scoring schemes. Results are discussed in section 5.

2. Assessing probabilistic forecasts with contingency table

An important point in regard to the construction of contingency tables for skill scoring of probabilistic forecasts is the determination of threshold values to differentiate between yes and no binary scores. Forecasting skill then derived from the contingency table is of practical value in the sense that users of a forecast often have to make a yes/no decision to act on the information provided. As aforementioned, users generally would not alter their practices unless there is a significant shift in probabilities away from random expectation (Hammer et al. 1996). Therefore it is of help if we assess the model forecasting skill by classifying probability forecasts to yes/no, and nonapplicable forecasts with a certain amount of probability departure from its random expectation. For *m*-categorical forecasts in which the random expectation of the occurrence of each category is $1/m$, one can use a range of probability departures in the skill assessment. Here we define $(1/m)/m$ as a *significant departure.* For tercile categorical forecasts, it is about $0.33/3 = 11.1\%$. If the forecasted probability value is equal or greater than $1/m + 1/m^2$, then we classify this forecast as a yes forecast as there is a significant shift in probability space that such an event is more likely to occur. If the forecasted probability value is less than $1/m - 1/m^2$, then we classify the forecast as a no forecast as there is a significant shift in probability space that such an event is less likely to occur. If the forecasted probability value is between $1/m - 1/m^2$ and $1/m + 1/m^2$, then we classify the forecast as a nonapplicable forecast as there is no significant shift in probability space and the chance for the occurrence of each category is equal. Such a forecast is of very limited value in the users' decision-making and we therefore classify such a forecast as a nonapplicable forecast. The implication of choosing such an arbitrary threshold in the transformation of probabilistic forecasts and the skill assessment will be discussed later in the paper. A different approach will be introduced to compensate this drawback.

TABLE 2. A 3 × 2 contingency table for converting probabilistic forecasts to binary yes/no and nonapplicable forecasts. Nonapplicable forecasts are defined as in the text.

	Forecast: yes	Forecast: no	Nonapplicable forecast
Observation: yes	A	B	X
Observation: no	C	D	Y

A range of skill score measurements has been developed for the verification of binary forecasts (see Wilks 1995 for a review). Most of these skill scores are based upon a 2 × 2 contingency table that is constructed as shown in Table 1. The skill score is expressed as some ratio of the hits and misses with respect to the possible totals. Among these, the Heidke (1926) and TSS scores (Hanssen and Kuipers 1965) are commonly used in forecast verification. Compressing the information contained in the A, B, C, and D components in Table 1 into a single value inevitably loses some of the information contained in the contingency table (Woodcock 1976; Schaefer 1990), and introduces some deficiencies in the skill measurement. Woodcock (1976) showed that with the exception of the TSS, most skill measurement scores are affected by the mixture of events and nonevents in the trial. The TSS score gives the best estimates on an ''unequal'' trial basis because it is proportional to the frequency of events being forecast and gives equal emphasis to the ability to forecast events and nonevents.

After transforming categorical probabilistic forecasts to yes, no, and nonapplicable forecasts, we need to construct a 3 × 2 contingency table as shown in Table 2 to reflect the nonapplicable component in the contingency table.

The TSS score from this contingency table can then be written as

$$\mathrm{TSS} = \frac{N_{\mathrm{cm}} - N_{\mathrm{ccm}}}{N_{\mathrm{all}} - N_{\mathrm{cco}}},$$

with $N_{\mathrm{all}} = A + B + C + D + X + Y$, $P_{\mathrm{yes}} = (A + B + X)/N_{\mathrm{all}}$, $P_{\mathrm{no}} = (C + D + Y)/N_{\mathrm{all}}$, $N_{\mathrm{cm}} = A + D$, $N_{\mathrm{ccm}} = (A + C) \cdot P_{\mathrm{yes}} + (B + D) \cdot P_{\mathrm{no}}$, and $N_{\mathrm{cco}} = (A + B + X) \cdot P_{\mathrm{yes}} + (C + D + Y) \cdot P_{\mathrm{no}}$, where *A, B, C, D, X, Y* are the components in Table 2 summated over all categories.

Here, P_{yes} and P_{no} are the climatological probabilities (or random expectation) of the occurrence of yes and no events; N_{cm} is the number of correct forecasts from the forecast model; N_{ccm} is the number of correct forecasts that could be achieved by chance; N_{cco} is the number of observed events that can be correctly forecasted by chance; and N_{all} is the total number of observations. Therefore, $N_{\mathrm{cm}} - N_{\mathrm{ccm}}$ represents the number of correct forecasts after subtracting those achieved by chance;

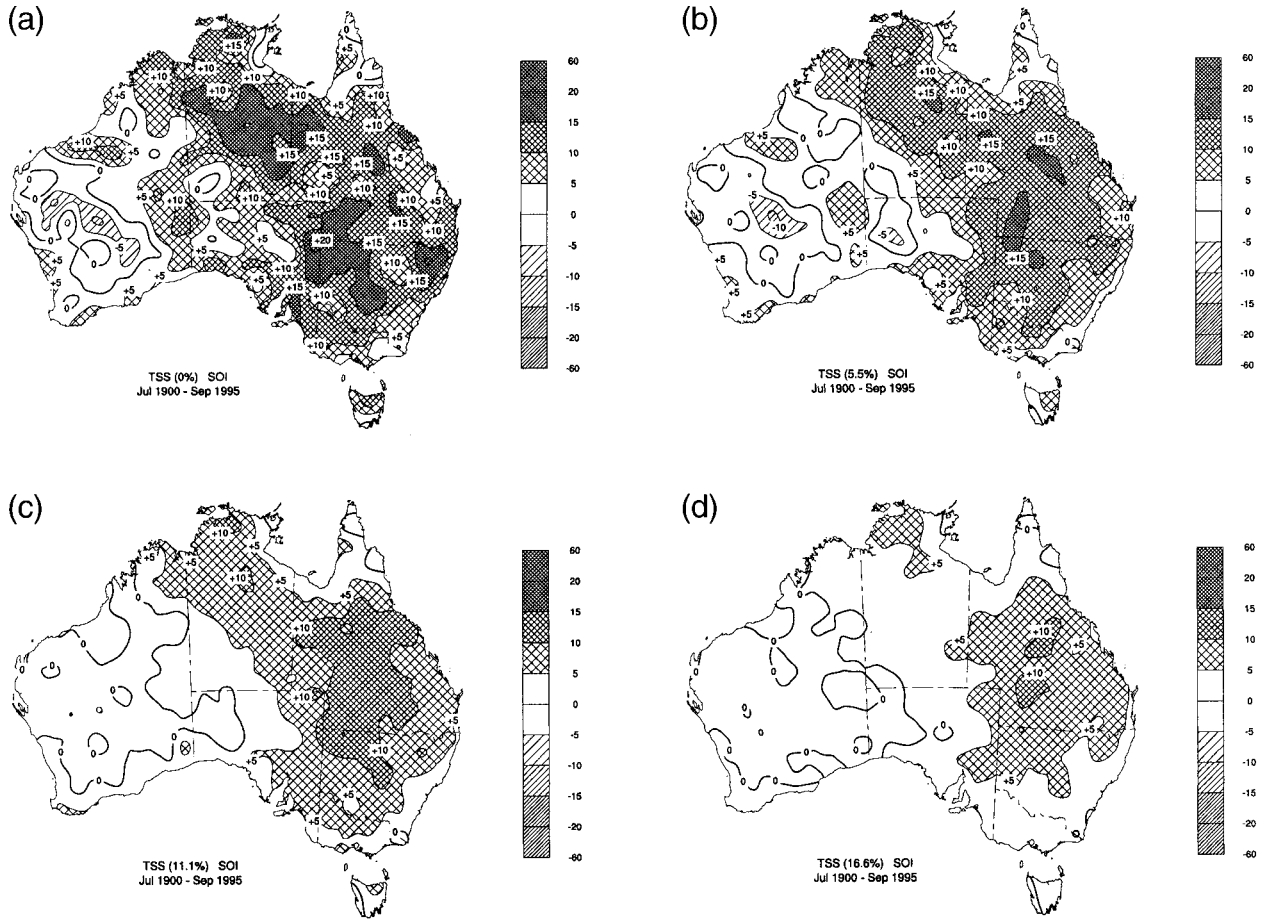

FIG. 1. Revised TSS score for the model hindcasts of Australian winter rainfall from 1900 to 1995: (a) with no probability departure, (b) with 5.5% probability departure, (c) with 11.1% probability departure, and (d) with 16.6% probability departure.

hence, it reflects the real model forecasting skill. In addition, $N_{\text{all}} - N_{\text{cco}}$ is the number of correct forecasts from a perfect model obtained by taking away the observed events that can be correctly forecasted by chance.

Figure 1 shows the revised TSS scores for a set of Australian winter (July–August–September) rainfall hindcasts in tercile categories produced using a Southern Oscillation index (SOI) phase model (Zhang and Casey 1992; Casey 1995). The tercile categorical hindcasts cover the period from 1900 to 1995 and have been generated using cross validation to maintain statistical independence (Casey 1998).

In Fig. 1 sensitivity of the revised TSS scores to the setting of different probability thresholds between forecasts and their random expectation is illustrated. These results are consistent with Mason (1979), who noted that if the climatological mean is used as the probability threshold, the TSS score would be maximized. However, very limited confidence can be attached when transforming forecasts around their random expectations to yes/no forecasts despite the fact that good scores may be shown from the skill measurements. When the category probability thresholds are set to 44.4% for yes forecasts and 22.2% for no forecasts, the scores over central, western, and southern Australian region are much degraded. Figure 1 illustrates the advantage of requiring a departure in probability space when converting probabilistic forecasts to binary form even though some information may be lost by the increase of threshold. With a different probability threshold, it can provide important information to users such as the stability for the model forecasting skills and the model real skills in terms of simple concepts as hits, misses, and false alarms.

However, it should be noted that part of the information from probability forecasts will be lost during the transformation to binary form and its verification. When probabilities from forecasts satisfy the threshold, the transformation does not take account of the degree to which they exceed the criterion in assigning them to yes/no or nonapplicable classifications. The severity of the errors between categories, that is, whether they are one or two categories away, is not taken into account in this method of scoring. Barnston (1992) has also pointed out that scoring measurements based upon a

contingency table will be more informative and valuable if the severity of categorical errors is included.

3. LEPS, RPS, and ROC

The LEPS score was developed by Ward and Folland (1991) and has been refined by Potts et al. (1996). It evaluates model skill by penalizing errors in terms of the distance between forecasts and observations in cumulative probability space. It gives relatively more penalty when forecasting events around average values, but gives relatively higher scores and less penalty for forecasts of extreme events. Potts et al. (1996) showed that when used for the verification of categorical forecasts the LEPS scoring matrix is equitable in that the expected score for a constant forecast of any category is the same as the observational (climatological) distribution in each category, although this characteristic is degraded for the version that measures the percentage skills of forecasts (Potts et al. 1996).

Another cumulative probability measure is the ranked probability score (RPS; Wilks 1995). The RPS is essentially an extension of the rms error measure to the multicategorical forecast case. In the RPS, the squared errors are computed with respect to the cumulative probabilities of the forecasts and observations and the observation in binary 0/1 form. The RPS and LEPS scores share the idea that the skill measurement should penalize forecast errors in terms of the probability assigned to the events. However, the LEPS scoring matrix is calculated from the distance between the forecasts and observations in continuous cumulative probability space, while the RPS calculations of the observed probability is in 0/1 binary form depending on whether the observation or forecast values fall within the category or not.

The idea of ROC comes from quality control and signal detection theory where the quality of performance is assessed by the relation between hit and false alarm rates as the decision criterion varies (Swets 1973; Egan 1975; Mason 1982). The graph of hit rates against false alarm rates within a range of probability thresholds is called the relative operating characteristics. From Table 1 the hit rate is defined as $h = P\{\textit{event is predicted} \mid \textit{event occurs}\} = A/(A + B)$, and the false alarm rate $f = P\{\textit{event is predicted} \mid \textit{event does not occur}\} = C/(C + D)$. Clearly, the hit and false alarm rates are closely related to the threshold used in transforming from probabilistic forecasts to yes/no forecasts. The hit rate can be increased by reducing the probability threshold, but at the same time the false alarm rate is increased. Similarly, reducing the false alarm rate is at the expense of reducing the hit rate. Hence a sequence of hit rate and false alarm rate pairs can be generated by changing the probability threshold through the range from 0 to 1.

The ROC curve has the following properties. (i) A perfect model locates at the point (0, 1) in the coordinates of false alarm rate and hit rate. In this case the forecast model gives either 0% or 100% forecasts, and no false forecasts from the model. The worst forecasting model locates at the point (1, 0) in which the model gives either 0% or 100% probabilistic forecasts but no correct forecasts against observations. (ii) Constant value forecasts and random forecasts will locate on the straight line between (0, 0) and (1, 1). (iii) The shape of the ROC curve gives a total description of the skill of the model forecasts at all probability thresholds. A model with good skill will have its ROC curve lying above and to the left of the (0, 0) to (1, 1) diagonal and a model with bad skill compared with the random or constant forecast will be seen below and to the right of the diagonal. As the ROC score evaluates the model forecasts by investigating the relative model performance of hit and false alarm rates across the entire range of probability thresholds, an integrated measurement of the curve can provide a score that is independent of the threshold probability level chosen to transform a probability forecast to binary form.

The first way to quantify the ROC is to calculate the area beneath the ROC curve (Green and Swets 1966). The larger the area, the better the model skill. If the area is less than 0.5 of the whole (unit area), then the model is less skillful than a random or constant forecast. The other way to evaluate the model in terms of the ROC is to transform the hit and false alarm rates under the assumption that distributions of hit rate and false alarm rate belong to a Gaussian distribution with equal variance. A normal–normal transformation is done for the hit and false alarm rates but at the same (sample) variance. This linearizes the ROC curve and area calculation is simpler.

In this study, we employ the first approach. We calculated the ROC score from the categorical probability forecasts by (i) setting up a range of probability thresholds P_c from 0% to 100% in increments of 1% and assigning the probabilistic forecasts to yes/no forecasts if $P > P_c$, (ii) creating a 2×2 contingency table as in Table 1 summed for all categories, (iii) calculating hit and false alarm rates from the contingency table for every probability threshold in 1% increments from 0% to 100%, and (iv) ranking the pair of hit rate and false alarm rate by ascending order of false alarm rates and calculating the area by the numerical integration of the ranked values using the trapezoidal rule.

Figure 2 is an example of the results from the ROC score measurement that is calculated for the 3-month rainfall probability hindcast data from the SOI phase. Figure 2a shows the ROC curves over three locations where model forecasts are of skill. In contrast, Fig. 2b shows the ROC curves over another three locations where the model skill is very limited. It should be pointed out that in this case if the forecasts from the model were reversed, they would become skillful ''forecasts.''

Although the apparent advantage of the ROC score is in the assessment of probability forecasts independently of probability thresholds, it should be pointed out that when it is applied to the verification of more than

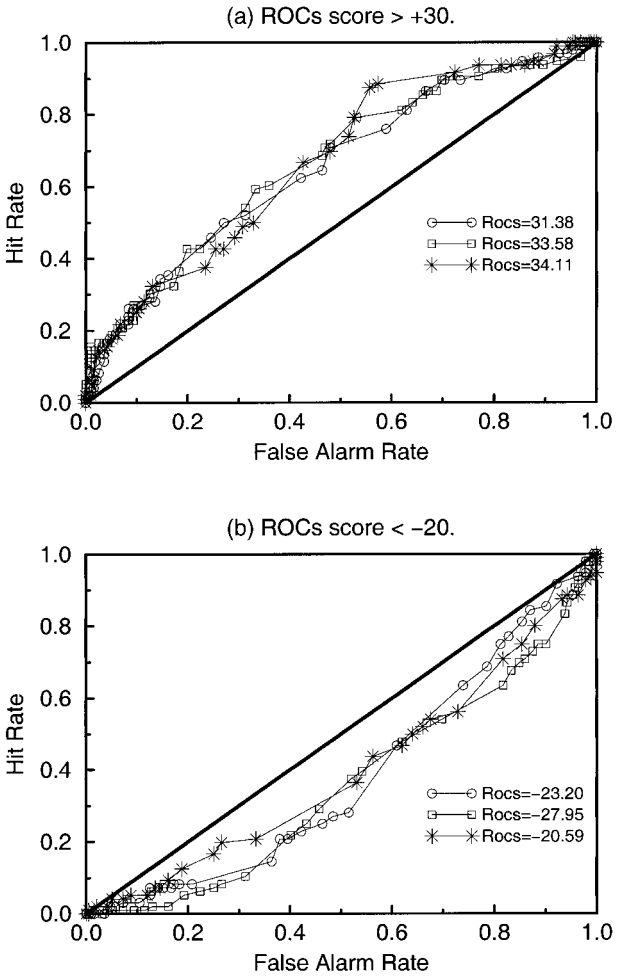

FIG. 2. ROC curve diagram showing the false alarm rate and hit rate through the range of probability thresholds from 0% to 100% at 1% increments. (a) ROC curves when the model skill is better than by chance; (b) ROC curves when the model skill is worse than chance. The diagonal from (0, 0) to (1, 1) in both diagrams is the ROC curve for climatological or random forecasts.

two category forecasts, it does not take account of the severity of errors across the categories. For instance, if tercile category one was observed, calculations of the false alarm rate do not distinguish between the severity of the error if terce two or terce three is forecast.

Clearly, from sections 2 and 3 we can see that different forecasting skill assessment approaches have their advantages and disadvantages. Some may have solid statistical bases but skill scores from those approaches may be difficult to interpret. Simple approaches can offer simple and straightforward skill scores but often have drawbacks in a statistical sense. Therefore, it is necessary to apply different skill assessment approaches in model forecasting evaluation.

4. Comparison of skill scores

In the previous sections we have discussed different approaches to the verification of categorical probability forecasts. Each approach provides valuable information about the skills of the probability model forecasts. The revised TSS score verifies model probability forecasts by converting them to yes/no binary form with a departure in probability space. It describes model skill based on a summary contingency table and results are easily interpreted. The ROC score compensates for the weakness of revised TSS in the sense that the revised TSS score is dependent on the prescribed probability threshold. ROC, on the other hand, evaluates the model performance across the entire domain of probability thresholds. Both skill measurements have the same unsatisfactory feature that they ignore the severity of the cross-category errors. Here we present results from both approaches together with the results from the LEPS and RPS for comparison and validation.

Figure 3 compares LEPS, ROC, TSS, and the RPS for the 96 years of Australian winter rainfall tercile hindcasts from an SOI phase model of Zhang and Casey (1992). The TSS results are with 11% probability departure. Overall, all scores show that the SOI phase model demonstrates good skill in the rainfall forecasts over the eastern part of the Australian continent during the winter season. Among the four skill measurements, the RPS shows the smallest areas of positive values where the model shows better skill than random forecasting. The ROC score shows a very similar pattern as the TSS score. Both of them have larger magnitudes of positive skill than either the LEPS score or RPS. This appears to be due to the fact that the ROC score and revised TSS score do not distinguish the severity of errors across categories while the LEPS score and the RPS give different penalties in terms of the categorical distance errors. In addition, the different magnitudes indicate that the statistical significance may be different in these skill scores and this will be discussed in the next section.

Figure 4 shows the correlations between the LEPS score and the TSS, RPS, and ROC scores calculated for the hindcasts of July–August–September and December–January–February seasons over the Australian region. The four score measurements in this study are in good agreement as seen in Fig. 3 and are highly correlated. The revised TSS and RPS scores have similar magnitudes to the LEPS score. Among the four skill scores, the ROC score has the largest magnitude. The revised TSS and ROC scores produce more negative skill scores than LEPS when the overall model skill is poor. In addition, the RPS does have a significant bias compared with LEPS. The differences between the LEPS, TSS, and ROC scores are partly due to the fact that the TSS and ROC scores do not take account of the magnitudes of probability when they satisfy the probability thresholds. In contrast, the RPS and the LEPS scores preserve the magnitudes of the probabilistic forecasts in their calculations of the difference between observations and forecasts and reward or penalize according to the severity of errors across the cat-

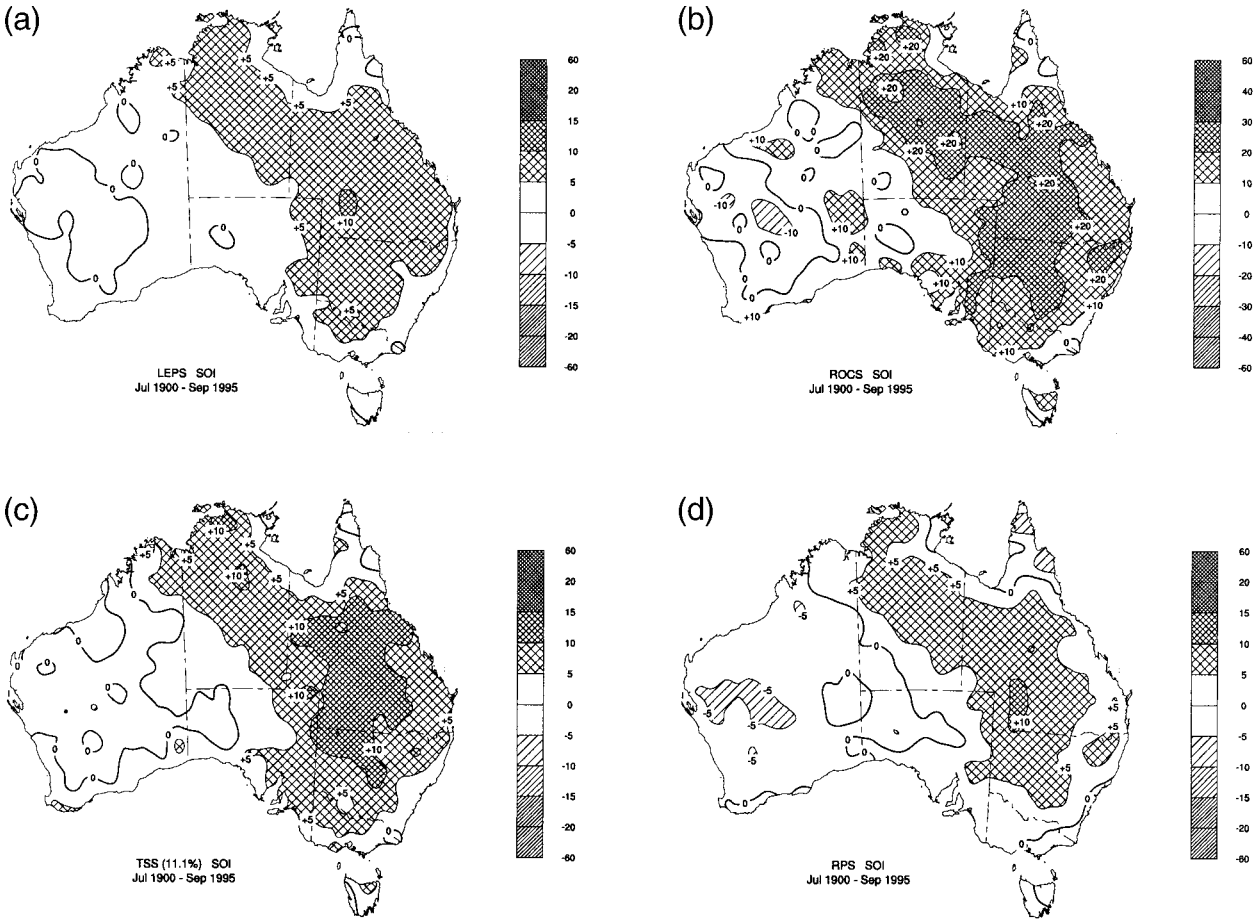

FIG. 3. The LEPS, RPS, ROC, and revised TSS scores for the model hindcasts of Australian winter rainfall from 1900 to 1995.

egories. From Figs. 3 and 4 it is seen that the scales for the skill scores are different for each of the four schemes and the ROC score is generally larger than the other scores. Thus, verifying model forecasts using different skill scoring methods also requires a knowledge of the statistical properties of the skill scores in order to establish the statistical significance of the skills and also to compare skill scores from different scoring schemes.

We have devised a quasi-random ensemble experiment to establish some statistical features of the different skill scores used in this study. Instead of generating random tercile categorical forecasts we keep the hindcasts from the SOI model intact but reorder the hindcasts in a cyclic way through the 96 yr of data for the period from 1900 to 1995. We quasi-randomize the data by reordering the observations 95 times by shifting the observations 1 yr ahead at a time and providing the last point with the first at each iteration. This gives a total number of 95 quasi-random observational time series, all of which retain the autocorrelation characteristic of the original series. Using these quasi-random observations to verify the forecasts, we obtain 95 skill scores for each scoring scheme. To further enhance the robustness of the test, the 95 random skill scores at every grid point over the Australian continent are combined, giving a total number of 98 606 random skill score population.

Figure 5 illustrates the frequency distribution of the four skill scores from the quasi-randomized data. All the skill scores for the randomized data display approximately normal distribution characteristics. The LEPS, revised TSS, and ROC scores all have their central mean skill near zero, but the RPS consistently shows a bias (and in all seasons) with a mean of about -5.1. Such a negative value is related to the feature of RPS as it penalizes forecast errors in terms of probability of the forecasting event. This explains why the RPS consistently shows less positive skill in Fig. 3. More instructively, results from Fig. 5 show that the standard deviations of the skill score distributions are quite different. The LEPS score has the smallest standard deviation with a value of about 2.3, while the standard deviation of the revised TSS score and the RPS are around 4.2. The ROC score has the largest standard deviation among the four skill scores with a value of about 8.8. For a given skill score, the larger the standard deviation, the less the statistical significance of that skill score achieved by the model.

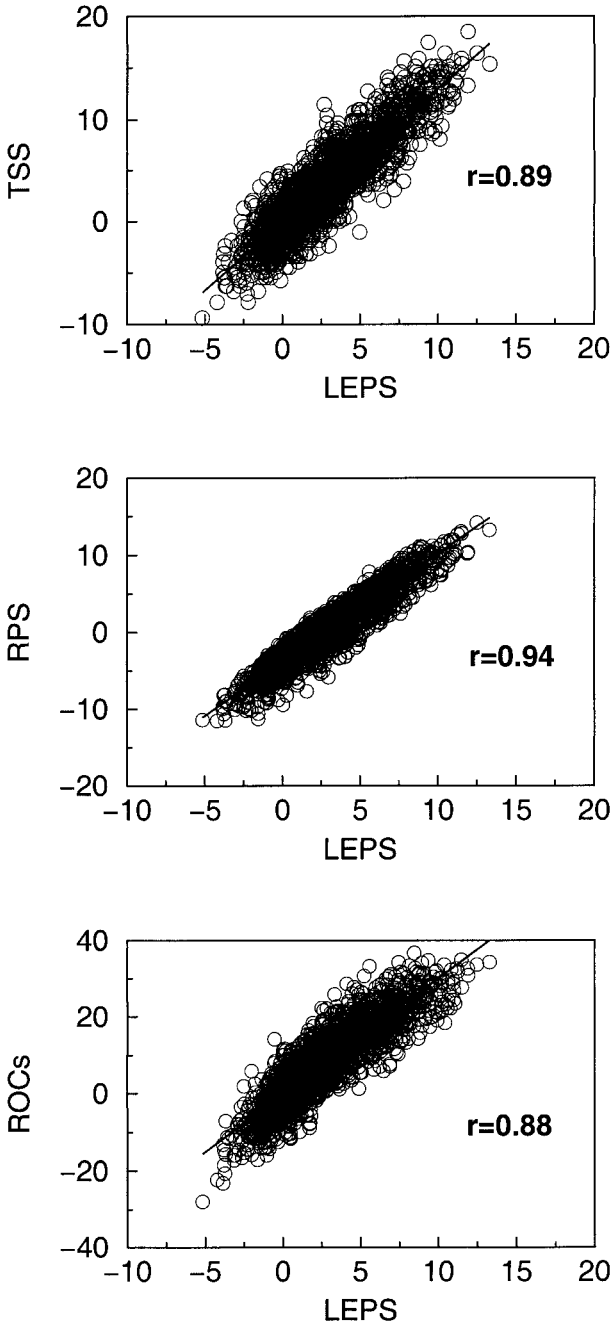

FIG. 4. Scatter diagrams showing correlation between the LEPS score and the other scores shown in Fig. 3.

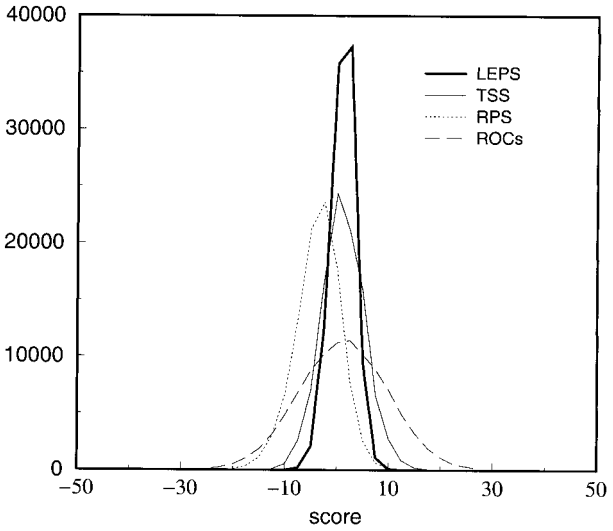

FIG. 5. Statistical skill score distributions derived from a quasi-randomized ensemble of hindcast data (see text).

As the distribution of skill scores has an approximately normal distribution, a significance level can be attached to the skill scores. For example, if the LEPS score from a model forecast is around 5.0, which is twice as large as the standard deviation of the scores found from the randomized hindcasts, we can state at the 95% confidence level that the skill score is statistically significant as there is only about a 5% probability that such skill can be achieved from a random (but serially correlated) set of forecasts.

As a further illustration, Fig. 6 shows the areas of the LEPS, ROC, and revised TSS skill scores of the July–August–September hindcasts from the SOI phase model where the areas of skill score value exceeding its two standard deviations are shaded for comparison. As seen earlier in Fig. 3, the spatial patterns of these statistically significant skill scores are in good agreement. All the skill score results indicate this model has significant forecasting skills over the east and north of the Australian continent. However, the areas of positive skill scores, which are statistically significant at the 95% confidence level, are marginally larger in the LEPS score than in the ROC and TSS scores. Despite the fact that the random experiment may only partially reflect the statistical properties of the skill scoring schemes examined in this study, these results highlight the importance of investigating the statistical features of differently formulated skill scoring schemes.

5. Discussion

a. Skill assessment and model development

It is generally recognized that when different prediction techniques are available, optimal combination of forecasts from those separate schemes provides higher skill than that achieved by any of the individual models (e.g., Fraedrich and Leslie 1987; Casey 1995). The weights of model forecasts in the combination are calculated by minimizing the mean-square errors. Casey (1995) indeed showed that the mean-square errors (mses) or half-Brier score (Wilks 1995) is reduced after combination of two different forecasts.

However, it has been pointed out that mean-square errors can be reduced by damping probability forecasts

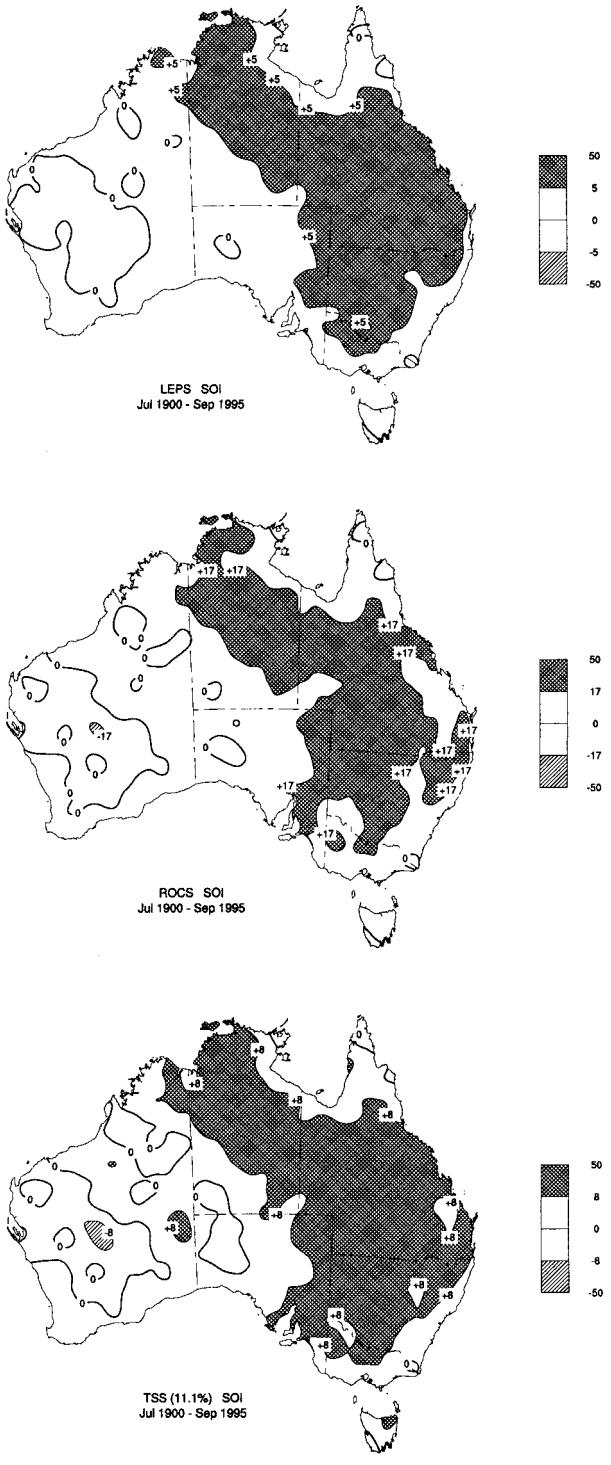

FIG. 6. LEPS, ROC, and TSS skill scores of the SOI phase model hindcasts for 1900 to 1995. Areas where skill scores are statistically significant at the 95% level are shaded.

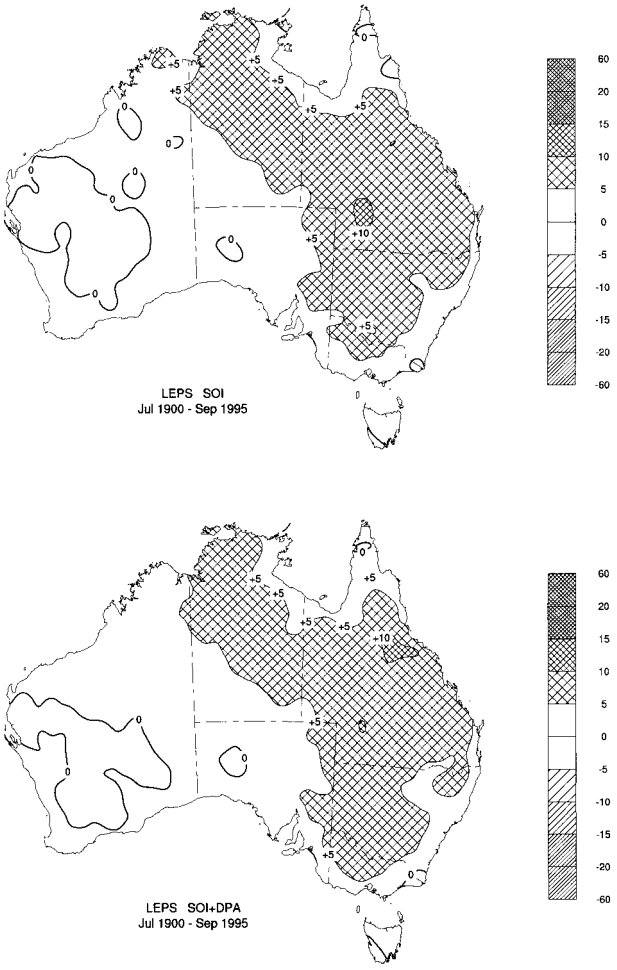

FIG. 7. Comparison of LESP scores for SOI phase model (SOI) hindcasts for 1900–95 Australian winter rainfall with those from the combination of SOI phase model and a Darwin pressure phase model.

(Murphy 1988; Potts et al. 1996). Combining techniques are in some sense damping the forecasts by introducing other variables to explain the total variance; therefore, the model skill may not be improved if other skill scores are calculated. Figure 7 shows LEPS scores calculated from the hindcasts of the SOI model and from optimal combination of the SOI phase model with another seasonal forecasting model that uses the Darwin surface pressure as the predictor. Using the optimal combination approach of Casey (1995), it was shown that mses are reduced by combining forecasts, compared with the individual models. However, comparing Figs. 7a and 7b, it is found that the positive LEPS scores are, in contrast, degraded in the optimal combination hindcasts compared with LEPS scores of the SOI phase model. This degrading is seen over the east coastlines and over southeast Australia regions. This is largely because LEPS not only measures the difference between forecasts and observations in magnitudes, but also gives higher scores and less penalty for forecasting rare

events. Clearly, forecasts damped toward the climatology from the optimal combination are not favored by LEPS skill scoring and thus there is no significant improvement in forecasting skill. This is a clear example demonstrating that the approach used in the model skill evaluation could affect the model development and improvement, thus underlining the importance of applying more than one skill assessment in practice.

b. Discussion

Mse scores can be evaluated using a binary representation of the observations, assigning 0 for no and 1 for yes depending on whether the observation falls within a given category or not. However, mean-square-error measures of skill can have a number of drawbacks. Barnston (1992) and Potts et al. (1996) have provided examples showing that if mse is used as a skill measurement, then this can be artificially improved by damping the forecasts through linear combination with a random variable, or equivalently, with the climatological or observational distribution. This feature can act as a disincentive to forecasting extreme values in any scheme that is based purely on this measure as a guide to performance. Schemes that produce forecasts clustered about the climatological expectation will score better on this measure than those that tend to forecast more extreme departures. Also, mean-square-error scores are not always appropriate for multiple category forecasts because information about the way the probability distributions shift toward the extremes within a particular category is lost (Stanski et al. 1989).

Assessment of skill by constructing contingency tables using yes/no binary forecast verification values has been used for many years. However, there is continuing debate on the strengths and weaknesses of these methods (see, e.g., Woodcock 1976; Schaefer 1990; Wilks 1995) as some of these skill scores have particular properties that are designed for specific applications such as the verification of rare events. This simplification inevitably loses some information about the overall character of the forecasts and can introduce deficiencies in the measurement of skill.

Most, if not all, of these schemes have a number of deficiencies when applied to multiple-category forecasts. Gandin and Murphy (1992) pointed out that many skill scores used to evaluate categorical forecasts of discrete variables are inequitable in the sense that constant forecasting of events near the mean produces better scores than constant forecasting of extremes. This is because for more or less randomly distributed phenomena there is a tendency to cluster about the mean, so the likelihood of a forecast being correct when close to the mean is higher than that for an extreme forecast. This has led to the devising of so-called equitable scoring matrices for categories, which penalize forecasts that are close to the mean and reward those that are farther toward the extremes. These depend on some assumptions about the underlying probability density structure. Ward and Folland (1991) developed the LEPS score, which uses the principle of the equitable scoring matrix. Mason (1982) introduced the ROC score into the assessment of meteorological forecasts. It is particularly suitable for the assessment of probabilistic predictions in that it is capable of measuring how much signal as distinct from noise is included in the information provided, in terms of the likelihood ratio between success and failure of the prediction.

In this paper our attention has been focused on the verification of tercile categorical probability forecasts, but we believe that a number of the concepts discussed here can be generalized to multicategory forecasts. First we have proposed an approach to evaluating model forecasts by converting probabilistic forecasts to binary yes/no forecasts with departures in probability space and then using a 3×2 contingency table to determine the model forecasting skills. We have revised the true success statistic score to take account of forecasts that are not significantly different from the random probability mean. The sensitivity of the TSS score to the probability threshold is discussed. This is of considerable value in the application of probabilistic-type forecasts. The relative operating characteristic score evaluates the model forecasting skills by investigating the hit and false alarm rates for varying probability thresholds. A practical approach is established for the calculation of the ROC score. As both the revised TSS and the ROC score do not penalize the errors in terms of their severity between each of the categories, the LEPS score and the ranked probability score have also been calculated for the purpose of intercomparison.

Model skills measured by the LEPS, ROC, TSS, and RPS are compared using verifications of an Australian seasonal rainfall forecast model. Overall similar distribution patterns of model skill over the Australian region are seen in the results from the four skill measurements. Among the four skill scores, the ROC score has the largest magnitudes and the RPS shows the smallest areas of positive skill. The scores from all these measurements are highly correlated but there is some bias between them. We have discussed the importance of investigating the statistical characteristics of any skill measurement scheme. Results from a series of random experiments conducted in this study suggest that the significance of the skill scores is related to the variance of the skill scores. Better skill score measurements should have a smaller variation of skill scores for random forecast data and this is an important issue that should be considered when a comparison of different skill score measurements is made.

The advantages and disadvantages of several different skill scores are discussed. In this study, we have presented four different approaches to verifying probabilistic forecasts in forms often found in weather and climate predictions. The revised TSS score converts probabilistic forecasts to yes/no forecasts with a departure

from the mean in probability space and then measures the skills from a modified contingency table that takes account of the number of nonapplicable forecasts with departures below the probability thresholds. The revised TSS score is of particular value if a yes/no decision is required from the probabilistic forecast at a prescribed probability threshold. The ROC score can give information about model performance at all threshold levels compared with random forecasts. However, the false alarm rates in the ROC score calculation are also equally weighted, that is, there is no regard to the difference in the size of the error between quantile categories. LEPS assesses the model skills by measuring the distance between forecasts and observations in the cumulative probability space. Nevertheless, it lacks information about the significance of the model skill for yes/no decision making. There are also some drawbacks as mentioned by Potts et al. (1996). Therefore different skill scores may be expected from the same forecasts if the skill score measures used in the assessment are different.

Skill scoring measurements have implications for the development of forecast models. An example has been shown in this study to demonstrate that improvements of the forecasting model in terms of one skill score result might not be seen if the model is verified by a different skill measurement. Developing a skill scoring scheme that overcomes the weakness of existing schemes often inevitably introduces other problems as pointed out by Potts et al. (1996) and no one standard scoring system has yet been developed that is appropriate for all types of forecasts. Considering the complexity of forecast verification, it is doubtful whether a universal scoring scheme exists. Rather, it seems desirable that a number of different scoring techniques should be applied in order to obtain an objective assessment of any given forecast scheme (Murphy 1991; Lee and Passner 1993; Huntrieser et al. 1997). Because of the complexity of objective scoring, it also appears there is no advantage in doing category forecasts and then scoring with elaborate methods to verify the forecasts, if no more useful information is conveyed to the user by presenting the forecast information in multiple categories rather than by the easily interpreted and easily scored probability of exceeding the median.

Acknowledgments. This study was undertaken when both authors worked at the Australian National Climate Centre and the Climate Group at the Bureau of Meteorology Research Centre. Thanks to Neville Nicholls, Beth Ebert, Frank Woodcock, and Ian Mason for reviews of early drafts and constructive advice from the BMRC Climate Group discussion.

REFERENCES

Barnston, A. G., 1992: Correspondence among the correlations, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Wea. Forecasting,* **7,** 699–709.

Bettge, A. G., D. P. Baumhefner, and R. M. Chervin, 1981: On the verification of seasonal climate forecasts. *Bull. Amer. Meteor. Soc.,* **62,** 1654–1665.

Casey, T. M., 1995: Optimal linear combination of seasonal forecasts. *Aust. Meteor. Mag.,* **44,** 219–224.

——, 1998: Assessment of a seasonal forecast model. *Aust. Meteor. Mag.,* **47,** 103–113.

Doswell, C. A., III, and J. A. Flueck, 1989: Forecasting and verifying in a field research project: DOPLIGHT 87. *Wea. Forecasting,* **4,** 97–109.

Egan, J. P., 1975: *Signal Detection Theory and ROC Analysis.* Academic Press, 277 pp.

Fraedrich, K., and L. M. Leslie, 1987: Combining predictive schemes in short-term forecasting. *Mon. Wea. Rev.,* **115,** 1640–1644.

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

Green, D. M., and J. A. Swets, 1966: *Signal Detection Theory and Psychophysics.* Wiley, 455 pp. (Reprinted by Robert E. Krieger Publishing Co., 1974.)

Hammer, G. L., D. P. Holzworth, and R. Stone, 1996: The value of skill in seasonal climate forecasting to wheat crop management in a region with high climatic variability. *Aust. J. Agric. Res.,* **47,** 717–737.

Hanssen, A. W., and W. J. A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Meded. Verh.,* **81,** 3–15.

Heidke, P., 1926: Berechnung des erfolges und der gute der windstarkevorhersagen im sturmwarnungsdientst. *Geogr. Ann.,* **8,** 301–349.

Huntrieser, H., H. H. Schiesser, W. Schmid, and A. Waldvogel, 1997: Comparison of traditional and new developed thunderstorm indices for Switzerland. *Wea. Forecasting,* **12,** 108–125.

Lee, R. R., and J. E. Passner, 1993: The development and verification of TIPS: An expert system to forecast thunderstorm occurrence. *Wea. Forecasting,* **8,** 271–280.

Mason, I., 1979: On reducing probability forecasts to yes/no forecasts. *Mon. Wea. Rev.,* **107,** 207–211.

——, 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

Murphy, A. H., 1977: The value of climatological categorical and probabilistic forecasts in the cost–loss ratio situation. *Mon. Wea. Rev.,* **105,** 803–816.

——, 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.,* **116,** 2417–2424.

——, 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.,* **119,** 1590–1601.

Potts, J. M., 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.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. World Weather Watch Tech. Rep. 8, WMO/TD No.-358, 114 pp.

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science,* **182,** 990–1000.

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

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

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purpose. *Mon. Wea. Rev.,* **104,** 1209–1214.

Zhang, X.-G., and T. M. Casey, 1992: Long-term variation in the Southern Oscillation and relationships with Australian rainfall. *Aust. Meteor. Mag.,* **40,** 211–225.