Revised "LEPS" Scores for Assessing Climate Model Simulations and Long-Range Forecasts

J. M. POTTS

Department of Statistics, IACR-Rothamsted, Harpenden, Hertfordshire, United Kingdom

C. K. FOLLAND

Hadley Centre for Climate Prediction and Research, Meteorological Office, Bracknell, Berkshire, United Kingdom

I. T. JOLLIFFE

Department of Mathematical Sciences, University of Aberdeen, Aberdeen, United Kingdom

D. SEXTON

Hadley Centre for Climate Prediction and Research, Meteorological Office, Bracknell, Berkshire, United Kingdom

(Manuscript received 14 October 1994, in final form 14 June 1995)

ABSTRACT

The most commonly used measures for verifying forecasts or simulations of continuous variables are root-mean-squared error (rmse) and anomaly correlation. Some disadvantages of these measures are demonstrated. Existing assessment systems for categorical forecasts are discussed briefly. An alternative unbiased verification measure is developed, known as the linear error in probability space (LEPS) score. The LEPS score may be used to assess forecasts of both continuous and categorical variables and has some advantages over rmse and anomaly correlation. The properties of the version of LEPS discussed here are reviewed and compared with an earlier form of LEPS. A skill-score version of LEPS may be used to obtain an overall measure of the skill of a number of forecasts. This skill score is biased, but the bias is negligible if the number of effectively independent forecasts or simulations is large. Some examples are given in which the LEPS skill score is compared with rmse and anomaly correlation.

1. Introduction

When assessing climate model simulations or predictions of fields of continuous variables such as surface pressure, the most commonly used measures of skill are root-mean-squared error (rmse) and anomaly correlation. These measures are related and each has disadvantages that are discussed in this paper. There are a number of existing systems for assessing forecasts that are given in the form of a number of categories whose prior probabilities are known. The aim of the linear error in probability space (LEPS) score is to provide a score that may be used to assess forecasts of both continuous and categorical variables and that does not have some of the problems associated with other scoring systems, such as rmse and anomaly correlation. This paper develops the LEPS scoring system introduced by Ward and Folland (1991) by deriving a new normalization method for the scores. Some important properties of the new scores are shown to be a marked improvement over the original LEPS scores. It is desirable to convert scores into measures of percentage skill. The behavior and limitations of skill scores based on LEPS are discussed, and some examples are given that are compared to standard and anomaly correlation.

2. Measures of similarity between meteorological fields

Let x_i, $i = 1 \cdots n$, denote a set of n observed values of a variable and y_i, $i = 1 \cdots n$, denote the corresponding forecast values. Then the mean-squared error (MSE) is defined as

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2, \qquad (1)$$

and rmse is the square root of this quantity.

The standard correlation coefficient in basic form is given by

Corresponding author address: Mr. C. K. Folland, Hadley Centre for Climate Prediction and Research, Meteorological Office, London Rd., Bracknell, Berkshire RG12 2SY, United Kingdom.

$$r_s = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{[\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2]^{1/2}}, \quad (2a)$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the observations and forecasts, respectively. Alternatively, the mean of the observed data may be subtracted from both the observations and the forecasts:

$$r_a = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{x})}{[\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{x})^2]^{1/2}}. \quad (2b)$$

In the case of time series at a single point, for which averaging over the entire series gives the climatological average at the point, r_a is the anomaly correlation coefficient. The following relationship exists between r_s and r_a (e.g., Ward and Folland 1991):

$$\frac{r_s}{r_a} = \left(1 + \frac{\text{BIAS}^2}{s_y^2}\right)^{1/2}. \quad (2c)$$

Here s_y^2 is the variance of the forecasts and BIAS $= \bar{y} - \bar{x}$.

If x_i, $i = 1 \cdots n$, and y_i, $i = 1 \cdots n$, are the observed and forecast values of a variable at each of a set of grid points, rather than time series of values at a single point, then one of a number of alternative forms of correlation coefficient is generally used. Since the climatological average of variables generally differs between grid points, often substantially, the ordinary correlation coefficient, in which the spatial means $\bar{x}$ and $\bar{y}$ are subtracted, has an expected value for independent meteorological fields that is greater than zero. To overcome this problem, the correlation coefficient can be calculated between the observed and forecast anomalies from climatological means obtained by averaging historical data, rather than between the actual values of the variable, giving

$$r_s' = \frac{\sum_{i=1}^{n} (x_i - \bar{c}_i - \bar{x} + \bar{c})(y_i - \bar{f}_i - \bar{y} + \bar{f})}{[\sum_{i=1}^{n} (x_i - \bar{c}_i - \bar{x} + \bar{c})^2 \times \sum_{i=1}^{n} (y_i - \bar{f}_i - \bar{y} + \bar{f})^2]^{1/2}} \quad (2d)$$

and

$$r_a' = \frac{\sum_{i=1}^{n} (x_i - \bar{c}_i - \bar{x} + \bar{c})(y_i - \bar{c}_i - \bar{y} + \bar{c})}{[\sum_{i=1}^{n} (x_i - \bar{c}_i - \bar{x} + \bar{c})^2 \times \sum_{i=1}^{n} (y_i - \bar{c}_i - \bar{y} + \bar{c})^2]^{1/2}}. \quad (2e)$$

Here $\bar{c}_i$, $i = 1 \cdots n$, are the climatological mean observed values of the variable at each grid point, $\bar{f}_i$, $i = 1 \cdots n$, are the corresponding climatological mean forecast values, and $\bar{c}$ and $\bar{f}$ are the spatial means of $\bar{c}_i$ and $\bar{f}_i$, respectively.

The correlation coefficients r_s' and r_a' are unable to detect systematic differences due to the addition of the same constant to the forecast value at each grid point. To overcome this, two different forms of correlation coefficient, in which the spatial means are not subtracted, are commonly used for comparing meteorological fields. The first form is analogous to the standard correlation coefficient r_s, but each observed value is referred to the climatological mean observed value rather than to the mean of the data, and each model forecast value is referred to the climatological mean forecast value:

$$r_s'' = \frac{\sum_{i=1}^{n} (x_i - \bar{c}_i)(y_i - \bar{f}_i)}{[\sum_{i=1}^{n} (x_i - \bar{c}_i)^2 \sum_{i=1}^{n} (y_i - \bar{f}_i)^2]^{1/2}}. \quad (2f)$$

In the remainder of the paper, when referring to meteorological fields, r_s'' is called the standard correlation. Alternatively, both observations and forecasts are subtracted from the observed climatological mean at each grid point (Miyakoda et al. 1972):

$$r_a'' = \frac{\sum_{i=1}^{n} (x_i - \bar{c}_i)(y_i - \bar{c}_i)}{[\sum_{i=1}^{n} (x_i - \bar{c}_i)^2 \sum_{i=1}^{n} (y_i - \bar{c}_i)^2]^{1/2}}. \quad (2g)$$

When meteorological fields are being compared, r_a'' is often called the anomaly correlation.

Standard correlation has the disadvantage that no account is taken of any systematic differences between the variance of the forecasts and that of the observations: multiplication of the forecast anomaly at each grid point by the same scale factor has no effect on the correlation coefficient. Anomaly correlation is affected by the variance of the forecasts in a complex way that depends on the size of the bias. The anomaly correlation coefficient also has the disadvantage that it is very sensitive to small differences between the forecasts and the observations when both are near the observed climatological average. Similar problems also arise for other forms of correlation when the denominator in the correlation coefficient is close to zero. Consider a comparison of a meteorological field with one that has been produced by perturbing the same data without bias. Suppose that $y_i = x_i + e_i$, where e_i is a set of random variables with mean zero and variance σ_e^2, which are distributed independently of the original anomalies $x_i - \bar{c}_i$. Let U denote the numerator in (2g) and V denote the square of the denominator. Then, treating the orig-

inal anomalies $x_i - \bar{c}_i$ as fixed, the expectations of U and V are

$$E[U] = \sum_{i=1}^{n} (x_i - \bar{c}_i)^2$$

$$E[V] = \sum_{i=1}^{n} (x_i - \bar{c}_i)^2 [\sum_{i=1}^{n} (x_i - \bar{c}_i)^2 + n\sigma_e^2].$$

As a rough approximation, the expectation of the anomaly correlation r_a'' is given by

$$E\left[\frac{U}{V^{1/2}}\right] \approx \frac{E[U]}{(E[V])^{1/2}} \approx \frac{1}{[1 + n\sigma_e^2 / \sum_{i=1}^{n} (x_i - \bar{c}_i)^2]^{1/2}}.$$

Thus, the expected value of r_a'' depends not only on σ_e^2 but also on the sum of the squared differences between the observations and the corresponding climatological averages. If this sum is close to zero, then r_a'' is close to zero even if σ_e^2 is small. On the other hand, the expected value of the MSE in this case is σ_e^2 regardless of the values of the observations.

Murphy (1988) demonstrated the following relationship between MSE and r_s:

$$\text{MSE} = \text{BIAS}^2 + s_x^2 + s_y^2 - 2s_x s_y r_s,$$

where

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{and} \quad s_y^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2.$$

By similar reasoning, Murphy and Epstein (1989) obtained

$$\text{MSE} = \text{BIAS}^2 + s_x'^2 + s_y'^2 - 2s_x' s_y' r_a',$$

where

$$s_x'^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{c}_i - \bar{x} + \bar{c})^2$$

$$\text{and} \quad s_y'^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{c}_i - \bar{y} + \bar{c})^2.$$

For our definition of standard correlation in (2f) we can define

$$s_x''^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{c}_i)^2 \quad \text{and} \quad s_{yf}''^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{f}_i)^2.$$

For a meteorological field, the bias varies between each grid point. Defining the local bias as $\text{BIAS}_i = (\bar{f}_i - \bar{c}_i)$, we have

$$\text{MSE} = s_x''^2 + s_{yf}''^2 - 2s_x'' s_{yf}'' r_s'' - \overline{\text{BIAS}_i^2} + \frac{2}{n}\sum_{i=1}^{n} (y_i - x_i)\text{BIAS}_i. \quad (3a)$$

A similar result may be obtained for anomaly correlation; this does not include bias terms:

$$\text{MSE} = s_x''^2 + s_y''^2 - 2s_x'' s_y'' r_a'', \quad (3b)$$

where

$$s_y''^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{c}_i)^2.$$

Rmse scores have the disadvantage that they can be reduced by damping the forecasts (Barnston 1992). For example, let z_i, $i = 1 \cdots n$, denote a set of forecasts. Consider modified forecasts of the form

$$y_i = \alpha z_i + (1 - \alpha)\bar{c}_i. \quad (4)$$

For these forecasts $s_y''^2 = \alpha^2 s_z''^2$ and $r_{xy}'' = r_{xz}''$, where

$$s_z''^2 = \frac{1}{n}\sum_{i=1}^{n} (z_i - \bar{c}_i)^2,$$

$$r_{xy}'' = \frac{\sum_{i=1}^{n} (x_i - \bar{c}_i)(y_i - \bar{c}_i)}{[\sum_{i=1}^{n} (x_i - \bar{c}_i)^2 \sum_{i=1}^{n} (y_i - \bar{c}_i)^2]^{1/2}},$$

$$\text{and} \quad r_{xz}'' = \frac{\sum_{i=1}^{n} (x_i - \bar{c}_i)(z_i - \bar{c}_i)}{[\sum_{i=1}^{n} (x_i - \bar{c}_i)^2 \sum_{i=1}^{n} (z_i - \bar{c}_i)^2]}.$$

Hence, from (3b)

$$\text{MSE} = s_x''^2 + \alpha^2 s_z''^2 - 2\alpha r_{xz}'' s_x'' s_z'',$$

which is minimized by setting

$$\alpha = \frac{s_x''}{s_z''} r_{xz}''. \quad (5)$$

If s_x'' and s_z'' are of similar magnitude, then α is approximately equal to r_{xz}''. Thus, the expected rmse may be reduced by using this unskilled strategy. However, assuming that the original forecasts came from a population with the same distribution as the observations, these modified forecasts will be, on average, less realistic than the original ones because they will be taken from a distribution that has a smaller variance than that of the observations.

Skill scores are often constructed by taking some measure of accuracy of the forecast and comparing it with that of a corresponding reference forecast. For MSE with observed climatology as a reference forecast, a skill score (SS) may be constructed as follows:

$$\text{SS} = 1 - \frac{\text{MSE}(x, y)}{\text{MSE}(x, c)},$$

where $\text{MSE}(x, y)$ and $\text{MSE}(x, c)$ are the MSE between the forecasts and the observations and between the climatic averages and the observations, respectively.

Murphy (1988) gives a decomposition of SS that involves r_s^2 as well as $(r - s_y/s_x)^2$, which is a measure of the conditional bias, and $(\bar{x} - \bar{y})^2/s_x^2$, which is a measure of the unconditional bias. Murphy and Epstein (1989) give a similar decomposition based on r_a'. The following result may also be obtained:

$$\mathrm{SS} = r_a''^2 - \left(r_a'' - \frac{s_y''}{s_x''}\right)^2 .$$

However, if the strategy given in (4) and (5) for reducing the MSE is adopted, then the conditional bias is eliminated and SS reduces to $r_a''^2$.

3. Existing assessment systems for categorical forecasts

Forecasts of temperature and precipitation are frequently given in terms of equally probable categories, for example, corresponding to below, near, and above normal. Various scoring systems exist for categorical forecasts. The Heidke score (Heidke 1926) is defined as $(H - E)/(N - E)$, where H is the number of forecasts of the correct category, N is the total number of forecasts, and E is the number of forecasts of the correct category that would be expected by chance in the absence of any forecasting skill. This score is insensitive to the magnitude of errors, when there are more than two categories, although it is possible to modify the score to allow for different classes of error (Barnston 1992).

The Sutcliffe score (Freeman 1967) does penalize errors according to their severity and has the property that the expected score is zero if an unskilled forecasting strategy is used, such as always forecasting the same category or selecting a category at random; it is therefore equitable (Gandin and Murphy 1992). However, the Sutcliffe score does not have the property that the expected score is the same for each observed category. So this score can vary according to the fluctuations in recent climate and can give a false impression of skill, since it tends to give a slightly higher score during a run of near-average conditions than during a run of extremes. We shall describe scores that have both the property where the expected score is the same for a constant forecast of any category and the property where the expected score is the same for each observed category as "doubly equitable." The Folland–Painting (FP) scores (Folland et al. 1986) are doubly equitable and are based on the negative logarithm of the distance between the forecast and the observation, measured in the cumulative probability distribution of the observations. This quantity is normalized to create a doubly equitable scoring system. A problem associated with the FP system is that it can "bend back." Thus, when the number of equiprobable categories is increased to eight or more, scores for maximally incorrect forecasts can be slightly less negative than those for less erroneous forecasts. So, for octiles, if the first octile is observed, the score for a forecast of the seventh octile is −2.165, while that for a forecast of the eighth octile is only −2.151.

4. Aim of LEPS

The LEPS score aims to provide a doubly equitable scoring system that does not bend back and that may be used to assess long-range forecasts on a common basis, as well as to assess the skill in simulating or forecasting meteorological fields. The forecasts to be assessed may be either best estimates of a single value of a variable from a continuous distribution or best estimates of one of a set of predefined categories. The FP system can only assess forecasts issued in discrete categories as it becomes undefined when continuous variables are forecast. Thus, if the forecast and the observation occupy the same point in the cumulative probability distribution, their distance apart is zero and the nonnormalized FP score is $-\log(0)$ and infinite. It is useful to have a method of relating the "skill" of multiple regression forecasts of a single "best estimate" value, taken from a continuous distribution, to that of a forecast of a discrete category derived from the same value. A score that can assess forecasts of continuous variables can also assess climate model simulations or predictions of fields of continuous variables such as surface pressure. It may sometimes be more appropriate to measure the errors of such simulations according to their errors in the climatological probability distribution, as done by LEPS, than by using rmse or correlation. LEPS penalizes errors in the simulation of an excessively low surface pressure less than would rmse, but gives a greater relative penalty to a small error in a near-normal pressure simulation. Unlike correlation, LEPS also penalizes errors in the distance between the forecast and the observation. So within the rather different framework of the cumulative probability distribution, it combines the useful characteristics of rmse and correlation.

The LEPS scores of Ward and Folland (1991) suffer from a bending back problem, like the FP score, when many categories are used. This becomes most acute for continuous variables as this condition corresponds to an infinite number of categories. A better form of LEPS is needed, therefore, for continuous variables. Scores for categorical forecasts can then easily be obtained by finding the expected score averaged over all values in each pair of forecast and observation categories.

One score that is doubly equitable and that does not bend back for continuous variables is a form of the Gringorten score (Gringorten 1965) defined as

$$G = \begin{cases} -\ln[P_f(1 - P_v)] - 1 & \text{if} \quad P_f > P_v \\ -\ln[(1 - P_f)P_v] - 1 & \text{if} \quad P_f < P_v, \end{cases}$$

where P_f is the cumulative distribution function of the forecast and P_v is the cumulative distribution function

of the observation. However, LEPS aims to provide a scoring system that has a simpler and more intuitive basis as, in its basic form, it measures the error in a forecast according to the distance between the position of the forecast and the corresponding observation in units of their respective cumulative probability distributions.

5. Derivation of LEPS

The first part of the derivation follows the appendix of Ward and Folland (1991), except that a perfect forecast is given a score of a rather than a score of unity. For imperfect forecasts, a is reduced by a penalty equal to the modulus of the difference between the position of the observation in the cumulative probability distribution P_v and that of the forecast P_f: P_v and P_f are fractions with a lowest possible value of zero and highest value of unity. The score S_{fv} has the form

$$S_{fv} = a - |P_f - P_v|, \tag{6a}$$

where S_{fv} is normalized relative to chance to give an equitable score S'_{fv} so that the chance score is zero for any forecast or observation. The normalizing factor is the product of the averages of all possible scores for the given forecast $\bar{S}_f$ and all possible scores for the given observation $\bar{S}_v$ divided by the grand mean score $\bar{S}_{fv}$. Thus, the normalized score S' is given by

$$S'_{fv} = S_{fv} - \frac{\bar{S}_f \bar{S}_v}{\bar{S}_{fv}}. \tag{6b}$$

In the continuous case this is

$$S'_{fv} = S_{fv} - \frac{\int_0^1 S_{fv} dP_v \int_0^1 S_{fv} dP_f}{\int_0^1 \int_0^1 S_{fv} dP_f dP_v}.$$

Evaluating the first of these integrals, we obtain

$$\int_0^1 S_{fv} dP_v = \int_{P_v=0}^{P_v=P_f} [a - (P_f - P_v)] dP_v + \int_{P_v=P_f}^{P_v=1} [a - (P_v - P_f)] dP_v = a - \frac{1}{2} + P_f - P_f^2.$$

Similarly,

$$\int_0^1 S_{fv} dP_f = a - \frac{1}{2} + P_v - P_v^2.$$

The grand mean of S_{fv} over all forecasts and observations is given by

$$\int_0^1 \int_0^1 S_{fv} dP_f dP_v = \int_0^1 \left(a - \frac{1}{2} + P_v - P_v^2\right) dP_v = a - \frac{1}{3}.$$

This gives a normalized LEPS score:

$$S'_{fv} = a - |P_f - P_v| - \frac{\left(a - \frac{1}{2} + P_f - P_f^2\right)\left(a - \frac{1}{2} + P_v - P_v^2\right)}{\left(a - \frac{1}{3}\right)}. \tag{6c}$$

From now onward, explicit use of the subscript fv will be dropped from S'.

As mentioned above, Ward and Folland (1991) use $a = 1$, but this form of LEPS suffers from the disadvantage of bending back for large numbers of categories. To ensure that S' does not bend back, we require

$$\frac{\partial S'}{\partial P_v} > 0 \quad \text{and} \quad \frac{\partial S'}{\partial P_f} < 0 \quad \text{for all} \quad P_f > P_v$$

and

$$\frac{\partial S'}{\partial P_v} < 0 \quad \text{and} \quad \frac{\partial S'}{\partial P_f} > 0 \quad \text{for all} \quad P_f < P_v.$$

Consider the case $P_f > P_v$. Differentiating we obtain

$$\frac{\partial S'}{\partial P_v} = 1 - \frac{\left(a - \frac{1}{2} + P_f - P_f^2\right)(1 - 2P_v)}{\left(a - \frac{1}{3}\right)} > 0 \tag{7a}$$

and

$$\frac{\partial S'}{\partial P_f} = -1 - \frac{\left(a - \frac{1}{2} + P_v - P_v^2\right)(1 - 2P_f)}{\left(a - \frac{1}{3}\right)} < 0. \tag{7b}$$

Suppose that $P_v = 0$. Then in order to satisfy (7a) when $P_f = 0.5$ we require

$$1 - \frac{\left(a - \frac{1}{4}\right)}{\left(a - \frac{1}{3}\right)} > 0, \tag{8}$$

and to satisfy (7a) when $P_f = 1$ we require

$$1 - \frac{\left(a - \frac{1}{2}\right)}{\left(a - \frac{1}{3}\right)} > 0. \tag{9}$$

There is no finite value of a that simultaneously satisfies (8) and (9). Similar problems occur with Eq. (7b).

TABLE 1. LEPS S'' scored for terces.

Forecast	Observation T1	T2	T3
T1	0.89	−0.11	−0.78
T2	−0.11	0.22	−0.11
T3	−0.78	−0.11	0.89

TABLE 2. LEPS S'' scores for quints.

Forecast	Observation Q1	Q2	Q3	Q4	Q5
Q1	1.28	0.52	−0.20	−0.68	−0.92
Q2	0.52	0.56	0.04	−0.44	−0.68
Q3	−0.20	0.04	0.32	0.04	−0.20
Q4	−0.68	−0.44	0.04	0.56	0.52
Q5	−0.92	−0.68	−0.20	0.52	1.28

We therefore let $a \to \infty$. Set $P_v^2 - P_v + 1/2 = \Delta_1$ and $P_f^2 - P_f + 1/2 = \Delta_2$ in (6c):-

$$S' = a - |P_f - P_v| - (a^2 - a\Delta_1 - a\Delta_2 + \Delta_1\Delta_2)\left(a - \frac{1}{3}\right)^{-1}.$$

For large a,

$$S' \approx a - |P_f - P_v| - (a^2 - a\Delta_1 - a\Delta_2 + \Delta_1\Delta_2)\left(a^{-1} + \frac{1}{3}a^{-2}\right).$$

Letting $a \to \infty$ we obtain a new score:

$$S' = 1 - |P_f - P_v| + P_f^2 - P_f + P_v^2 - P_v - \frac{1}{3}. \quad (10)$$

The partial derivatives of this score are, for $P_f > P_v$,

$$\frac{\partial S'}{\partial P_v} = 2P_v > 0 \quad \text{and} \quad \frac{\partial S'}{\partial P_f} = -2 + 2P_f < 0, \quad (11a)$$

and for $P_f < P_v$,

$$\frac{\partial S'}{\partial P_v} = -2 + 2P_v < 0 \quad \text{and} \quad \frac{\partial S'}{\partial P_f} = 2P_f > 0. \quad (11b)$$

We call these two conditions the two states of P_f relative to P_v. Thus, this form of LEPS score does not change the sign of its gradient for a given state and satisfies the conditions that ensures that it does not bend back. The score in (10) can be seen from (6a) to be the LEPS score obtained using $a = 1$ as in Ward and Folland (1991) but with the new normalization

$$S' = S - \bar{S}_f - \bar{S}_v + \bar{S} \quad (11c)$$

instead of that given in (6b).

From (10), the expected value of the score of a continuous variable for a correct forecast averaged over all equal values of P_f and P_v is 1/3. To produce a score with an expected value of 1 similarly calculated, this score needs to be multiplied by three. The revised, normalized, LEPS score, S'', is therefore defined as

$$S'' = 3(1 - |P_f - P_v| + P_f^2 - P_f + P_v^2 - P_v) - 1. \quad (12)$$

We call this the revised LEPS score, but the word "revised" will be dropped for the remainder of the paper. Note from (11a) and (11b) that the rate at which S'' changes with P_v asymptotically approaches zero for both $P_v = 0$ and 1, independently of the value of P_f.

The scores for forecasts issued in discrete categories are obtained by calculating the expected score for each combination of observed and forecast categories. For forecasts issued in discrete categories, the mean score for all categories for a high level of skill tends to increase as the number of categories increases. Thus, the average score for correct categories is 0.5 for two equiprobable categories, for terces it is 2/3, and for quints it is 0.8. In general, for n categories, the average LEPS score for correct categories is $1 - 1/n$. Table 1 shows

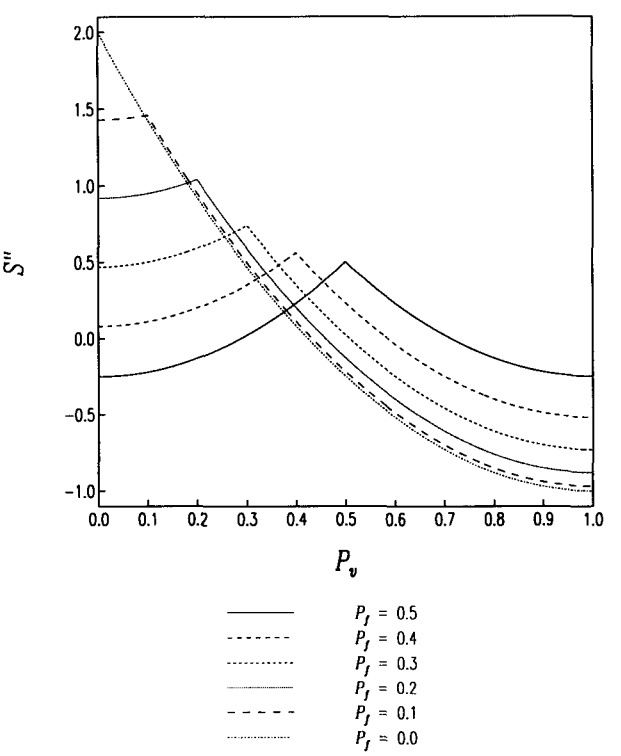


FIG. 1. The LEPS S'' score as a function of P_v for values of P_f at intervals of 0.1 between 0.0 and 0.5.

all scores for terces, rounded to two decimal places, while Table 2 shows the scores for quints.

6. Properties of LEPS

Figure 1 shows how S'' varies with P_v for $P_f = 0$, 0.1, 0.2, 0.3, 0.4, and 0.5. The curves for $P_f > 0.5$ are mirror images of those for $P_f < 0.5$. The maximum score, which occurs when $P_f = P_v = 0$ or $P_f = P_v = 1$, is 2, while the minimum score, which occurs when $P_f = 0$ and $P_v = 1$ or $P_f = 1$ and $P_v = 0$, is -1. The score for a correct forecast of a value near the center of the distribution is much less than that for a correct forecast of an extreme value. Because S'' is doubly equitable, the score is symmetrical so that the score for $P_f = A$ and $P_v = B$ is the same as that for $P_f = B$ and $P_v = A$.

Given one of the states $P_f > P_v$ or $P_f < P_v$, the partial derivative of the score with respect to P_v depends only on P_v. Hence the curves in Fig. 1 are parallel. As a consequence, provided that the state of P_f does not change, the change in S'' that occurs as a result of a given change in P_v is the same regardless of the value of P_f. This effect is visible in Table 2.

The variance of S'' for a random forecast and a given observation in the continuous case is, from (12),

$$\int_0^{P_v} (3P_f^2 + 3P_v^2 - 6P_v + 2)^2 dP_f + \int_{P_v}^1 (3P_f^2 + 3P_v^2 - 6P_f + 2)^2 dP_f = 4(-3P_v^4 + 6P_v^3 - 3P_v^2 + 0.2).$$

The overall variance of S'' for random forecasts and observations is

$$\int_0^1 4(-3P_v^4 + 6P_v^3 - 3P_v^2 + 0.2) dP_v = 0.4.$$

For rmse, the expected score is increased by issuing forecasts with a lower variance than the observations, but for LEPS the opposite is true. Here, S'' was derived in such a way that if the forecasts and the observations are independent, then the expected score is zero regardless of the variance of the forecasts. If forecasts are perfect, the maximum possible score is obtained where the variance of the forecasts is the same as that of the observations. However, if, as is usually the case, there is a positive correlation between the forecasts and the observations but the forecasts are not perfect, then, assuming that the same probability distribution is used for the forecasts as for the observations, the expected score can be slightly increased by issuing forecasts with a higher variance than the observations.

Suppose that the observations come from a normal distribution with zero mean and unit variance, that the population correlation coefficient between the forecasts and the observations is ρ, and that the forecasts also have a similar normal distribution but multiplied by a scale factor of σ. Then for a given observation x, the forecast has a normal distribution with mean $\sigma\rho x$ and variance $\sigma^2(1 - \rho^2)$. A close approximation to the expected score for values of ρ at intervals of 0.1 and for σ between 0 and 10 was obtained using the scores for 100 equiprobable categories. For each observed category, the conditional probability that the forecast lay in each category was calculated given that x lay at the midpoint of the observed category. Since the observation is equally likely to occur in any of the 100 categories, the probability of each combination of observed and forecast categories was obtained by dividing each conditional probability by 100. These probabilities were multiplied by the corresponding LEPS scores and added to give the expected score shown in Fig. 2. The differences between the maximum score and the score for $\sigma = 1$ are in general quite small. Although the optimal value of σ is very large for small values of ρ, as σ is increased the expected score becomes almost constant. This relatively small problem can be overcome by calculating the cumulative probability distribution of the forecasts and referring the forecasts to this distribution in order to reflect the larger variance of the forecasts compared to the observations. Equivalently, the forecast and observed values can be standardized by their respective standard deviations if the distributions are normal. This is readily done for model simulations for instance.

7. Estimating percentage "skill"

It is desirable to have a measure of overall skill over a range from 100% to -100%. The problem is to devise

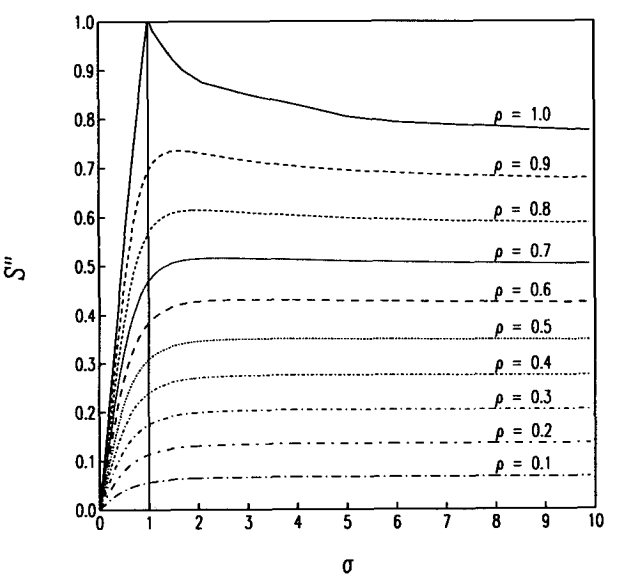


FIG. 2. The expected LEPS S'' score when the standard deviation of the forecasts σ is between 0 and 10, the observations have a standard deviation of 1, and the forecasts are assumed to have the same distribution function as the observations. The population correlation coefficient between the forecasts and the observations ρ takes values at intervals of 0.1 between 0.1 and 1.0.

TABLE 3. LEPS SK percentage skill terce table for a single forecast and a single observation.

	Observation			
Forecast	T1	T2	T3	Mean
T1	100.00	−100.00	−100.00	−33.33
T2	−14.29	100.00	−14.33	23.81
T3	−100.00	−100.00	100.00	−33.33
Mean	−4.76	−33.33	−4.76	−14.29

a technique that, as far as possible, does not lose the equitable character of the LEPS scores. A method is developed that succeeds for practical purposes as long as a sufficiently large ensemble of forecasts is assessed together. To achieve a skill range from 100% to −100%, average skill (SK) is defined for continuous, categorical, and probability forecasts as

$$\mathrm{SK} = \frac{\Sigma\, 100S''}{\Sigma\, S''_m}, \qquad (13)$$

where the summation is over all pairs of forecasts and observations. The definition of S''_m depends on whether the numerator is positive or negative. If it is positive, S''_m is the sum of the maximum possible scores given the observations, that is, the scores assuming all the forecasts were correct, because 100% skill is logically the result of forecasting the same category or value as is subsequently observed. If the numerator is negative, S''_m is the sum of the moduli of the worst possible scores given the observations. [The transformation in (13) was also used by Folland et al. (1986) to provide percentage skill scores for the FP scoring system.] For continuous forecasts, S''_m is readily calculated from (12) for positive values of S'' by setting $P_v = P_f$. For negative values, for a given P_v, the largest negative score is found from the value of P_f that is furthest away from P_v in the cumulative probability distribution. This will be the value of S'' corresponding to $P_f = 1$ or $P_f = 0$.

For a single forecast, SK is certainly not equitable. Consider the S'' scores for terces (Table 1). Average skill involves dividing the score by the maximum possible score, given the observation, if S'' is positive, and by the modulus of the minimum possible score, given the observation, if S'' is negative. Table 3 shows the resulting table for a single forecast. The expected mean skill for the whole table is −14.29%. The expected value of SK is 23.81% for a single forecast of terce 2 and −33.33% for a single observation of terce 2. Table 4 gives comparable values for quints. The overall expected skill score for quints is −4.58%, which is less negative than the expected score for terces.

Consider the expected value of SK aggregated over a pair of independent forecasts. For example, consider pairs of forecasts of terce 2. There are nine possible pairs of observations, each with equal probability. These are listed in Table 5, together with the corresponding values of SK. The average expected SK is 9.2% for the aggregate of two forecasts of terce 2, which is less than half the bias of 23.9% for one forecast. For aggregated pairs of observations of terce 2, a similar calculation (Table 6) gives a negative bias of −22.2%, two-thirds of the bias of −33.3% for a single observation of terce 2. The initial bias in SK for other categories declines in the same way.

To illustrate the behavior of SK as the number of forecasts increases further, simulations were carried out for selected numbers of aggregated independent forecasts between 1 and 400. In each case, 100 000 simulations were carried out, as the standard deviation of SK values is very high compared to their means. Thus, for a constant forecast of terce 3, observations were randomly made of terces 1, 2, and 3. The mean percentage biases obtained are given below, together with their standard errors, which are shown in parentheses. The biases in SK for a single forecast and for constant forecasts of terce 1 or 3 and constant forecasts of terce 2 are calculated to be −33.38% (0.30) and 23.78% (0.17), respectively, which are close to their theoretical values of −33.33% and 23.81%; for 5 forecasts they are −6.08% (0.16) and 2.16% (0.05); by 25 forecasts the biases reduce to −1.89% (0.07) and 0.16% (0.02); by 100 forecasts to −0.95% (0.04) and −0.06% (0.01); by 400 forecasts biases are −0.41% (0.02) and −0.05% (<0.01). These results are displayed graphically in Fig. 3a. For constant forecasts of quint 1 or quint 5, quint 2 or quint 4, and quint 3 the biases for a single forecast are −21.24% (0.30), 3.01% (0.21), and

TABLE 4. LEPS SK percentage skill quint table for a single forecast and a single observation.

	Observation					
Forecast	Q1	Q2	Q3	Q4	Q5	Mean
Q1	100.00	92.86	−100.00	−100.00	−100.00	−21.43
Q2	40.63	100.00	12.50	−64.71	−73.91	2.90
Q3	−21.74	7.14	100.00	7.14	−21.74	14.16
Q4	−73.91	−64.71	12.50	100.00	40.63	2.90
Q5	−100.00	−100.00	−100.00	92.86	100.00	−21.43
Mean	−11.01	7.06	−15.00	7.06	−11.01	−4.58

TABLE 5. SK for pairs of forecasts of terce 2.

Combinations of terces observed	Numerator	Denominator	SK (%)
1 and 1	−0.22	1.56	−14.10
1 and 2	0.11	1.11	9.91
1 and 3	−0.22	1.56	−14.10
2 and 1	0.11	1.11	9.91
2 and 2	0.44	0.44	100.00
2 and 3	0.11	1.11	9.91
3 and 1	−0.22	1.56	−14.10
3 and 2	0.11	1.11	9.91
3 and 3	−0.22	1.56	−14.10
			Mean SK = 9.25%

14.18% (0.14), respectively; for 5 forecasts they are −5.23% (0.15), −1.20% (0.10), and 1.77% (0.04); for 25 forecasts they are −1.72% (0.07), 0.72% (0.04), and 0.10% (0.02); for 100 forecasts they are −0.79% (0.03), −0.41% (0.02), and −0.06% (0.01); and for 400 forecasts they are −0.37% (0.02), −0.21% (0.01), and −0.07% (<0.01). These results are shown in Fig. 3b. Thus, SK tends quickly at first, and then slowly, to zero as the number of forecasts rises. This is also seen for persistently observed terces or quints, although the detailed behavior for persistently observed categories is not the same as for persistently forecast categories.

Simulations were also carried out for terces, quints, and for continuous variables for the situation when both forecasts and observations are randomly sampled. The biases obtained for a single forecast were −14.25% (0.28), −4.49% (0.25), and −7.30% (0.21) for terces, quints, and continuous variables, respectively; for 5 forecasts they were −3.20% (0.13), −2.24% (0.12), and −3.72% (0.10); for 25 forecasts they were −1.29% (0.06), −1.09% (0.05), and −1.62% (0.05); for 100 forecasts they were −0.68% (0.03), −0.54% (0.03), and −0.81% (0.02); and for 400 forecasts they were −0.34% (0.01), −0.26% (0.01), and −0.44% (0.01). These results are shown in Fig. 3c.

TABLE 6. SK for pairs of observations of terce 2.

Combinations of terces observed	Numerator	Denominator	SK %
1 and 1	−0.22	0.22	−100
1 and 2	0.11	0.44	25
1 and 3	−0.22	0.22	−100
2 and 1	0.11	0.44	25
2 and 2	0.44	0.44	100
2 and 3	0.11	0.44	25
3 and 1	−0.22	0.22	−100
3 and 2	0.11	0.44	25
3 and 3	−0.22	0.22	−100
			Mean SK = −22.22%

For about 20 or more independent forecasts, biases in SK can usually be neglected and even for five independent forecasts biases may often not be serious. However, SK scores are not suitable for single forecasts as they are insufficiently equitable and the overall negative bias is rather large.

8. Some comparisons between the LEPS skill score and other measures

Figure 4a shows the relation between SK and the population correlation coefficient in a similar way to Fig. 2. The differences between the maximum LEPS score and that for $\sigma = 1$ are again in general quite small, and zero for perfect correlation. Figure 4b shows SK plotted against r_s'' for $\sigma = 1$ for 2000 simulated, independent pairs of fields consisting of three independent points. Also shown is the mean line obtained by simulating 10 000 pairs of dependent fields consisting of 500 independent points for various selected population correlation coefficients between the fields. Intermediate values were obtained by interpolation. The precise position of the mean line will depend to a small extent on the number of independent points in each field as SK is biased for small numbers of such points. In the case of 500 points the bias is negligible, but the actual average curve for three independent points differs slightly from the curve shown. The *scatter* will depend on the number of independent points, or, in the case of spatially correlated fields, on the equivalent number of independent points. Thus, for simulations of fields of 50 independent points (not shown) the scatter is much less about the mean line with no values near ±1. Note that if we define the equivalent number of independent points N as the number of independent points for which a statistic has the same variance as for a spatially correlated field of n points, then, for the mean, the equivalent number of independent points is given by

$$N \approx \frac{n}{1 + (n - 1)\bar{r}},$$

where $\bar{r}$ is the average correlation of every grid point with every other grid point (Parker et al. 1992). However, for the standard correlation coefficient r_s'', the following approximation to N is obtained (see appendix):

$$N \approx \frac{n}{1 + (n - 1)\overline{r^2}},$$

where $\overline{r^2}$ is the average of the squares of the correlations between all pairs of points.

Some advantages of LEPS over anomaly correlation and rmse are shown next. Figure 5a shows a hypothetical field of climatological average values of half-monthly mean sea level pressure over a 5 × 6 grid. In the following calculations of LEPS skill scores it was assumed that the forecasts and observations are normally distributed with means given by these climato-

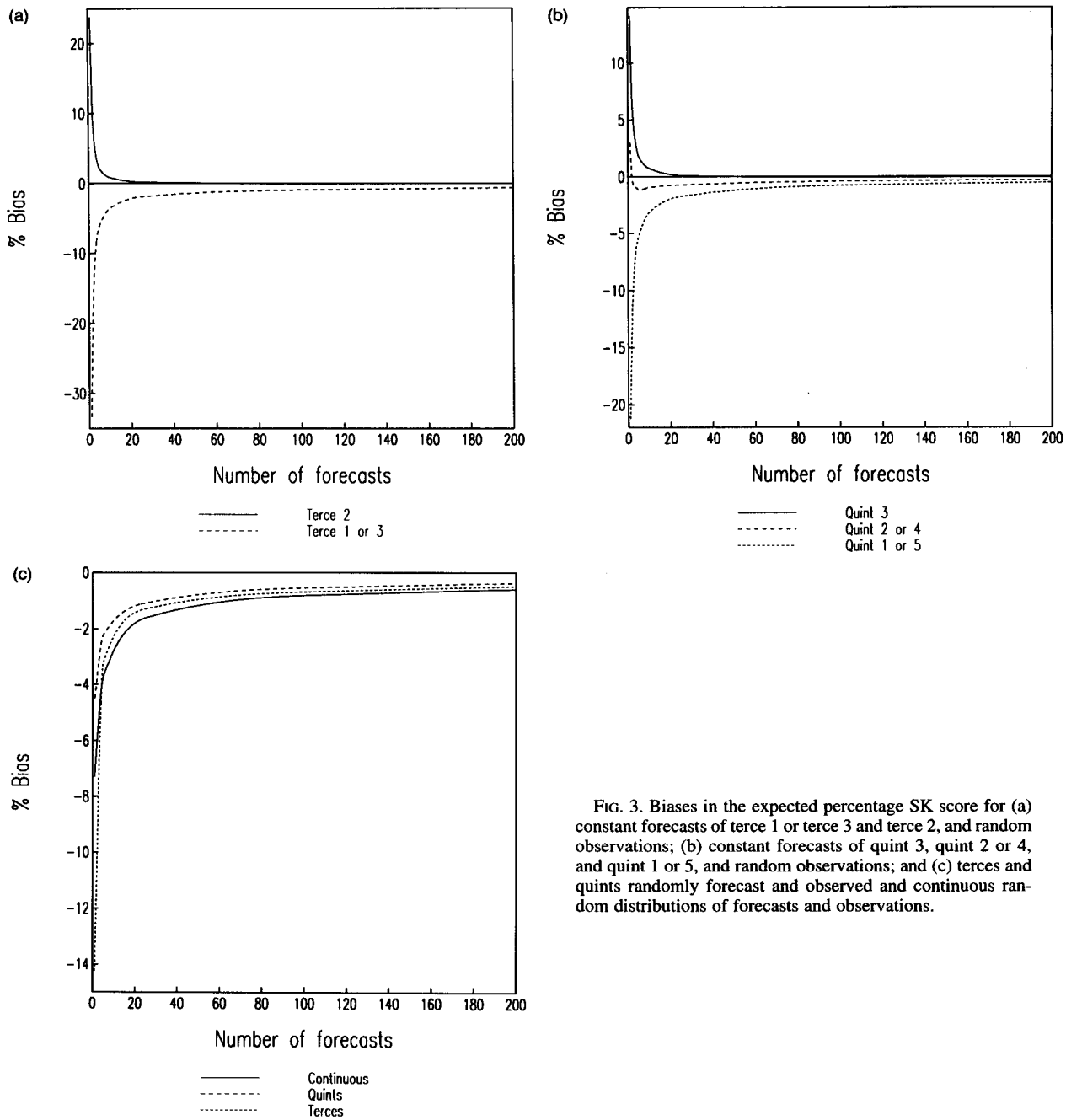


FIG. 3. Biases in the expected percentage SK score for (a) constant forecasts of terce 1 or terce 3 and terce 2, and random observations; (b) constant forecasts of quint 3, quint 2 or 4, and quint 1 or 5, and random observations; and (c) terces and quints randomly forecast and observed and continuous random distributions of forecasts and observations.

logical averages and a standard deviation of 9 mb at each grid point. Suppose that Fig. 5b represents an observed pattern over this grid and Fig. 5c represents the corresponding forecast pattern. The anomaly correlation between these two patterns is 0.69, the rmse is 6.41, and the LEPS skill score is 35%.

The sum of the squared differences between the observations and the climatological averages is approximately equal to that between the forecasts and the climatological averages for these two patterns. Damped forecasts y_i' were therefore formed as follows (Fig. 5d):

$$y_i' = 0.69y_i + 0.31\bar{c}_i,$$

where y_i, $i = 1 \cdots 30$, are the original forecasts at each grid point. Comparing these damped forecasts with the observations, the anomaly correlation remains the same; the rmse reduces to 5.44, implying that the

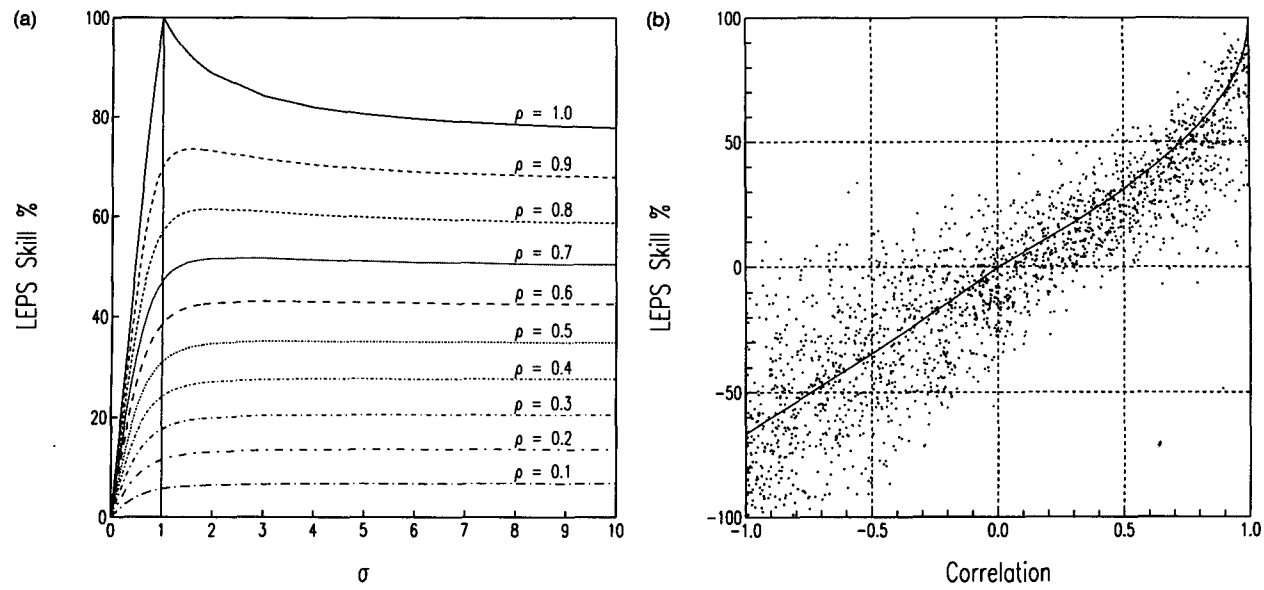


FIG. 4. (a) The expected LEPS skill when the standard deviation of the forecasts σ is between 0 and 10: the observations have a standard deviation of 1 and the forecasts are assumed to have the same distribution function as the observations. The population correlation coefficient between the forecasts and the observations ρ takes values at intervals of 0.1 between 0.1 and 1.0. (b) Simulated values of LEPS skill versus standard correlation r_s'', where observations and forecasts have equal standard deviations, for fields consisting of three equivalently independent values, together with the mean line for fields containing a large number of independent values.

damped forecast is better than the original one; and the LEPS skill score is 30%, implying that the damped forecast is slightly worse than the original one. Visual inspection of the patterns suggests that the damped forecast is probably not quite as good as the original one because it underestimates the intensity of the high pressure region.

Now suppose that the observed and forecast anomalies are each multiplied by a scale factor of 0.1, so that the new observations (Fig. 5e) x_i'' are given by

$$x_i'' = 0.9\bar{c}_i + 0.1x_i,$$

where x_i, $i = 1 \cdots 30$, are the original observations and the new forecasts (Fig. 5f) y_i'' are given by

$$y_i'' = 0.9\bar{c}_i + 0.1y_i.$$

Because all the observed and forecast anomalies have been multiplied by the same scale factor, the anomaly correlation is still only 0.69 even though these two patterns are very similar. The increased skill is, however, reflected in both the rmse, which has fallen to 0.64, and the LEPS skill score, which has risen to 85%.

9. LEPS and correlation scores for several climate model simulations

We first study a time series of simulated and observed data for northeast Brazil where the skill is relatively high. LEPS scores and LEPS skill are compared with r_s and r_a. LEPS skill and anomaly and standard correlation are then compared for gridpoint simulations where the skill is very low. Finally we compare simulated and observed global fields of seasonal rainfall where the skill is very variable. In all cases we use continuous distributions of both the observed and modeled data.

a. Northeast Brazil rainfall time series

We first compare observed rainfall in northeast Brazil in 1949–85 during the March–May wet season with that simulated by the atmospheric part of the Hadley Centre 19-level 2.5° latitude × 3.75° longitude climate model. The Hadley Centre model is a component of the U.K. Meteorological Office unified climate and weather forecasting model (Cullen 1993). Note that observed data after 1985 are too sparse to analyze. The model was forced with observed sea surface temperature and sea ice extent from the Meteorological Office Global Sea Ice and Sea Surface Temperature dataset (D. E. Parker 1992, personal communication). The model was run four times for this period, forced with the same time-varying sea surface temperature and sea ice extent data, but restarted from four different sets of initial atmospheric conditions appropriate to 1 October (taken from recent years). The initial ocean surface data were always those for 1 October 1948. Since the predictability of the atmosphere from the initial atmospheric conditions alone is only about two weeks, any common signals in the integrations averaged over any specific period of time after the first two weeks must come from the ocean surface data or variations in other

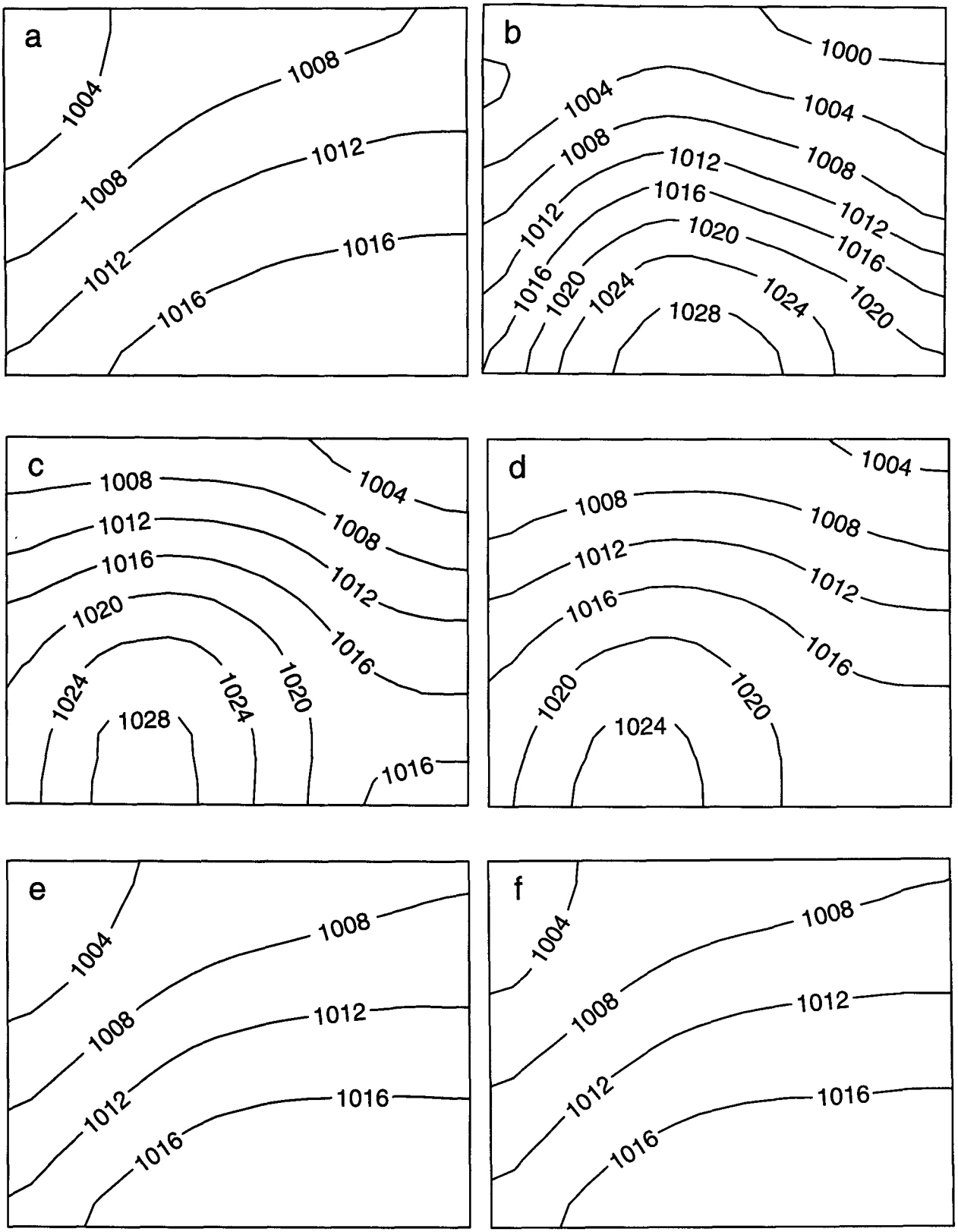


FIG. 5. Hypothetical half-monthly average mean sea level pressure. (a) Climatological average. (b) Observed pattern. (c) Forecast pattern. (d) Damped forecast pattern. (e) Observed pattern close to the climatological average. (f) Forecast pattern close to the climatological average.

boundary conditions, like calculated soil moisture, that result from the influence of the ocean surface data on the model atmosphere. The remaining differences between the runs are due to the unpredictable internal variability of the model. If the ocean surface fields have a strong influence, however, the unpredictable component will be quite small. Northeast Brazil is expected to be an area where the ocean surface temperature has a strong influence on climate, particularly rainfall, on timescales of three months or more. A large empirical influence of sea surface temperature on the northeast Brazil wet season rainfall is shown by Ward and Folland (1991). The observed rainfall time series is an average of values estimated for two grid boxes 2.5° latitude × 3.75° longitude centered at 3.75°S, 39.375°W and 6.25°S, 39.375°W. Model grid boxes have the same size but are offset by half a grid box, so corresponding values were obtained by averaging model boxes centered on for 5.0°S, 37.5°W and 5.0°S, 41.25°W. The reason for the difference in the grids is that the observed gridded data were created to match the grid of a different climate model that was used at the Hadley Centre until recently; regridded rainfall data are not yet available. All modeled values are based on the average of the ensemble of four runs.

Figure 6a shows the modeled and simulated rainfall in millimeters, while Fig. 6b shows the same standardized values, where standardization was done for the whole period and individually for each series. There is a high value of $r_s = 0.87$ between the simulated and observed values in Fig. 6a, but the modeled data has a much smaller $\sigma = 99$ mm compared to the observed $\sigma = 213$ mm. There is also a bias of -143 mm in the modeled compared with observed rainfall, the observed mean being 581 mm. The rmse of the simulations is 136 mm, which is markedly larger than the simulated standard deviation. Because of the bias and small simulated standard deviation, r_a is reduced to 0.49 [Eq. 2c]. Standardization in Fig. 6b removes the bias, equalizes the standard deviations, and gives an rmse of only 0.51 standard units, that is, now only half the standard deviation of the simulations. Also r_s and r_a both equal 0.87. Figure 6c shows the individual LEPS scores and their corresponding maxima and minima for Figs. 6a and 6b. Modeled and observed data in Fig. 6a are both referred to the observed cumulative probability distribution; corresponding LEPS scores are called LEPSOB. In Fig. 6b, scoring of the standardized modeled data against the standardized distribution is effectively the same as assessing it against the model's own cumulative probability distribution (the model and observed distributions are insignificantly different from normal). These scores are called LEPSOBMD. Maximum and minimum LEPS scores in Fig. 6c depend only on the position of the observation in the observed cumulative probability distribution. This is the same for

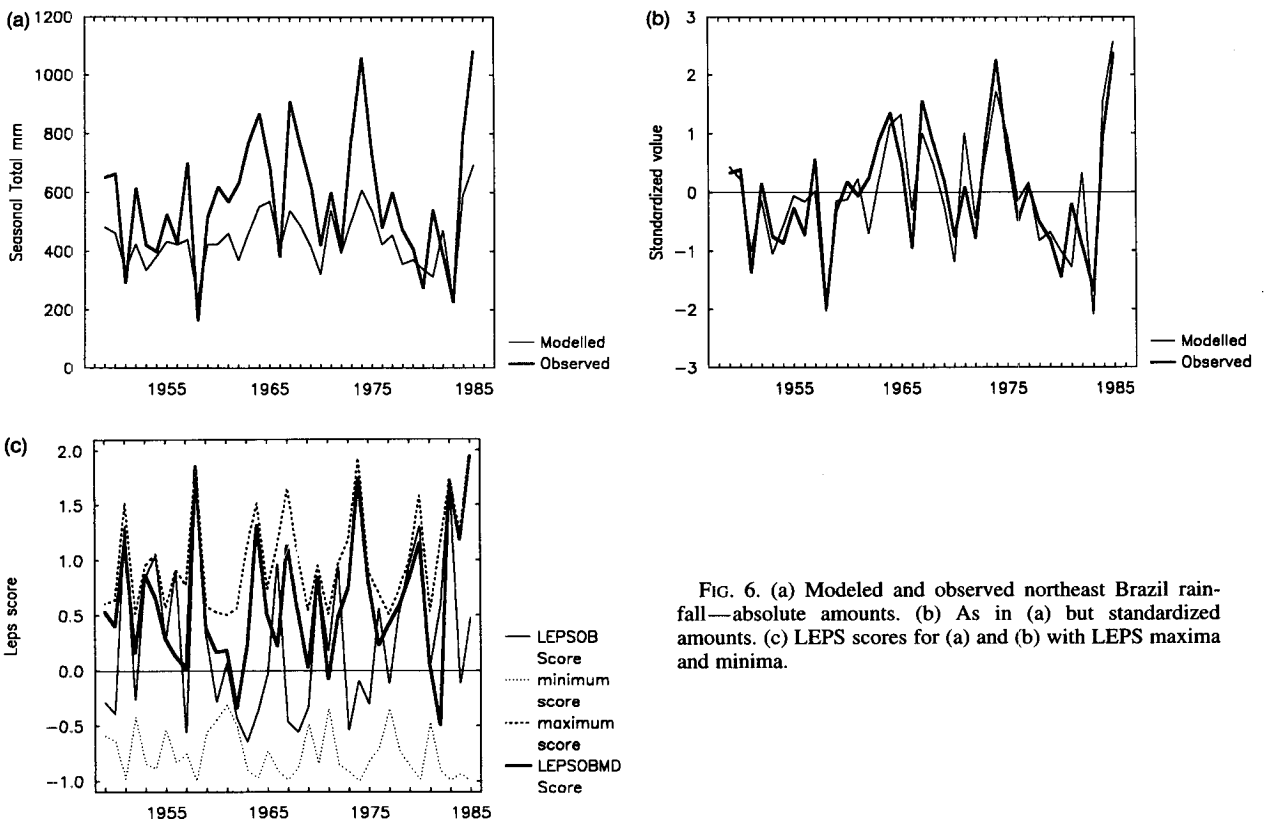


FIG. 6. (a) Modeled and observed northeast Brazil rainfall—absolute amounts. (b) As in (a) but standardized amounts. (c) LEPS scores for (a) and (b) with LEPS maxima and minima.

LEPSOB and LEPSOBMD because standardization of the observed distribution in Fig. 6b is a linear transformation of that underlying Fig. 6a. The highest maxima correspond to the most extreme observations, as expected from Fig. 1. The LEPSOB skill of 24.5% is much less than the LEPSOBMD skill of 61.3%. The former score would be the appropriate one to use when considering the need to improve the ability of the model to truly physically simulate northeast Brazil rainfall. The latter is more appropriate for estimating the potential seasonal predictability of March–May northeast Brazil rainfall as such standardized values can always be calculated in advance of the rainfall season and reinterpreted in terms of true expected rainfall. The difference in the two LEPS scores arises because many modeled values in the LEPSOB calculation are low in the observed cumulative probability distribution because of negative bias and low variance. Bias affects both anomaly correlation and LEPSOB, but a reduced standard deviation only influences anomaly correlation if bias is present. This is emphasized by recalculating Fig. 6a but with no bias. Both r_s and r_a are again 0.87, but the LEPSOB skill is 37.3%, still much below the value of LEPSOBMD (61.3%). Thus, unlike anomaly correlation, LEPS is sensitive to forecast variance independently of forecast bias. From Fig. 4a, the σ value as defined there is 0.46, and a correlation of 0.87 corresponds to a LEPSOB skill of about 40%, close to that calculated.

Figure 6c shows the unbiased LEPS scores for each simulation (LEPS skill is too biased for single forecasts). It can be seen that LEPSOBMD simulations for 1962 and 1982 were unskillful, and several other years had poor skill even with the advantage of standardization of modeled values. This kind of information is not easy to obtain except in a cumulative probability framework. Note that the high average LEPSOBMD skill is due to the large weight given to scores for a correct or near-correct simulation of an observed extreme. As Fig. 6c indicates, the extremes are mostly very well simulated.

b. *Gridded surface pressure fields in the North Atlantic–European region (low skill case)*

Here we compare observed and simulated fields for a low-skill situation where the influence of the ocean surface on simulated climate is low. Thus, we expect the internal nonlinear variability of the model to be dominant. The same observed anomaly fields of two-month average pressure at mean sea level (PMSL) were compared between 1949 and 1990 with two of the model runs used for Northeast Brazil simulations extended to 1990, giving 504 pairs of observed and simulated two-month average fields. Thus, the model was restarted in the same way as in section 9a. Anomalies were calculated from 1951 to 1980 averages. The region 65°–35°N, 40°W–20°E was chosen for this test as atmospheric variability here is only a little influenced by SST (D. Rowell 1993, personal communication); average skill is low, allowing a comparison of LEPS and correlation in this situation. Each field was sampled at 63 grid points to fit observed PMSL data on a 5° latitude × 10° longitude grid. Simulations in the first run were almost uncorrelated with those in the second run, with little serial correlation, so time series of all 504 values were created where the second run follows the first. Four types of correlation were looked at: r'_s, r'_a, r''_s, and r''_a, together with three different ways of calculating LEPS skill:

a. LEPSOB—model and observed PMSL data at a grid point are referred to the observed cumulative probability distribution of PMSL for 1951–80 at that grid point.

b. LEPSOBNB—as for LEPSOB but the mean differences between the individual model grid point and the observed values for 1951–80 (the local biases) are removed.

c. LEPSOBMD—as for LEPSOB but model data at each grid point are referred to the local model 1951–80 climatology.

For ready comparison with correlation, LEPS skills are expressed on a fractional scale from −1 to +1 in the remainder of section 9.

Table 7a shows the average and standard deviation for all 504 values of the four types of correlation and the three LEPS skill scores. Average LEPS skill scores are (a) those for each field calculated separately using Eq. (13) and (b) the grand average obtained by first adding the LEPS scores for all grid points in the 504 fields before using Eq. (13) to calculate LEPS skill. Calculation (a) gives a small negative bias near −3% because of the small number of equivalently independent grid points in individual model and observed fields (see Fig. 3c). It can be seen that all LEPS skill scores have mutually similar standard deviations that are substantially less than those for correlation. This can be understood by referring to Figs. 4a and 4b, where for

TABLE 7a. LEPS skill (expressed as a fraction) and correlation for North Atlantic–European area two-month average PMSL anomaly fields, 1949–1990.

	Average	Standard deviation
r'_s	0.02	0.41
r'_a	0.03	0.37
r''_s	0.02	0.40
r''_a	0.02	0.38
LEPSOB (a)	−0.01	0.28
(b)	0.02	
LEPSOBNB(a)	−0.01	0.28
(b)	0.02	
LEPSOBMD(a)	−0.02	0.29
(b)	0.01	

TABLE 7b. Correlation between LEPS skills and correlation measures for data in Table 7a.

LEPS skill	r'_s	r'_a	r''_s	r''_a
LEPSOB	0.75	0.75	0.80	0.89
LEPSOBNB	0.84	0.83	0.89	0.79
LEPSOBMD	0.75	0.74	0.81	0.76

TABLE 7c. Intercorrelation between LEPS skills.

	LEPSOB	LEPSOBNB	LEPSOBMD
LEPSOB	1	0.84	0.78
LEPSOBNB		1	0.86
LEPSOBMD			1

large correlations the LEPS skill is less except for $\rho = 1$, though very small LEPS and correlation scores are similar and nominally identical for $\rho = 0$.

Table 7b shows standard correlations [Eq. (2a)] between the LEPS skill and correlation scores for the 504 individual fields, and Table 7c shows the correlation between the different LEPS skill scores. The correlations are generally high. The high correlation between r''_a and LEPSOB arises as in section 9a because both penalize local bias whereas r''_s does not. However, r''_a is not *always* less than r''_s; it can happen that the modeled values for a specific field of gridpoint values are on average nearer the local observed climatologies rather than the local model climatologies even if appreciable longer-term biases occur. This contrasts with r'_s and r'_a, where r'_a is always equal to or less than r'_s.

Figure 7a shows all 504 values of LEPSOB skill plotted against r''_s (correlation = 0.80 from Table 7a). The scatter, and therefore the correlation, depends partly on the number of independent points equivalent to the 63 grid points in the sense that the standard correlation coefficient has the same variance. This is assessed (see section 8 and appendix) to be close to 7 grid points. The simulated line is that calculated for Fig. 4b, so Fig. 7a is similar to Fig. 4b except that the scatter is less because Fig. 4b assumed only three equivalently independent points.

Figure 7b is a plot of LEPSOB skill and standard correlation for 120 consecutive fields of the second run. The main difference between the two measures is the lack of values of LEPSOB above 0.5, partly compensated by a lack of very negative values. This results mostly from the relationship shown in Fig. 4a, where for equal standard deviations of the observations and the forecasts, a correlation of say 0.7 is expected to be equivalent to an unbiased LEPS skill of about 0.47, though with wide individual variations as indicated by Fig. 4b. The arrowed pair [season 6 (November–December), 1967, second run] shows a particularly big difference with $r''_s = 0.67$ and LEPSOB = 0.07. The corresponding anomaly correlation $r''_a = 0.17$, so most of the difference compared to the mean relationship in Fig. 4a is due to large numbers of local biases. Thus, the anomaly correlation and LEPSOB fit Fig. 4a quite well. By contrast, in season 3 of 1976, r''_a and r''_s are similar at 0.53 and 0.40, while LEPSOB is only 0.04. This largely arises because the standard deviations of observed and modeled pressure vary systematically over this late spring field. Relatively small absolute errors in the southern part of the field have negative LEPS scores (low standard deviations) and are treated on an equal standardized basis with larger errors in the north. However, in the correlation calculations larger positive covariances in the north override generally smaller neg-

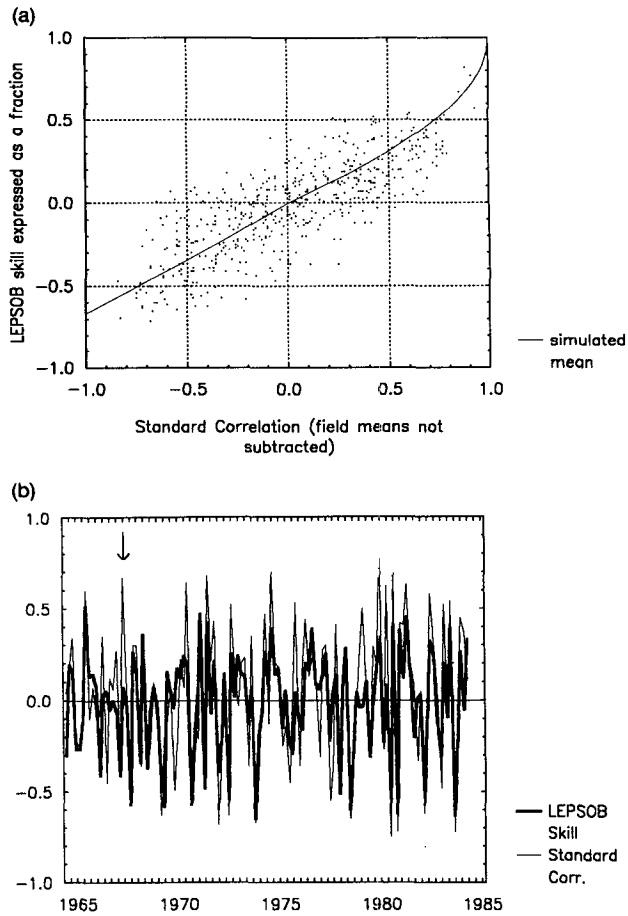


FIG. 7. (a) LEPS skill LEPSOB (observations and modeled values both referred to the observed cumulative probability distribution at each grid point) vs standard correlation r''_s. LEPSOB skill is expressed as a fraction. Scores for 504 cases, each consisting of 63 correlated grid points of modeled and observed pressure at mean sea level in the North Atlantic–European region, are shown. (b) Time series of typical 120 consecutive values of LEPSOB vs standard correlation r''_s.

ative covariances in the south that are nevertheless climatologically as important. This problem can be seen from Eqs. (2f) and (2g). Although the local climatological average is allowed for, differences between these and the local simulated values are lumped together, whereas LEPS effectively disaggregates the differences and assesses their relative importance first. This is clearly a further advantage of LEPS.

c. *Gridded global rainfall (strong geographical variation in skill)*

Finally we show for completeness some typical results that can be expected on a global scale, where the model skill varies greatly geographically. We illustrate only the bare results; further analysis will be completed elsewhere. Figures 8a–d show maps of correlation and LEPS skill for four simulations of seasonal mean rainfall for boreal spring (March–May) based on the 42 seasons in 1949–90. Each seasonal simulation was averaged over the four runs. The observed data are observations over land only, derived from the latest version of the Climatic Research Unit rainfall dataset (Hulme 1992). Skill measures are standard correlation r_s and anomaly correlation r_a, and LEPSOB and LEPSOBMD. Values are assessed on the model 2.5° × 3.75° grid and the climatological averages and LEPS cumulative distribution curves are calculated over 1949–90. The global mean skill at this space and time resolution is very low (weighted average $r_s = 0.09$, $r_a = 0.05$, LEPSOBMD = 0.03, and LEPSOB = 0.0), but there is a very strong variation in skill geographically, even within the Tropics. The high LEPSOBMD skill in northeast Brazil can be seen, together with the strong sensitivity to the skill measure used as discussed in section 9a. (The only other differences between these results and those for Northeast Brazil in section 9a are the use of single grid boxes and that LEPSOBMD uses empirical distributions rather than normal distributions.) The global standard deviations in skill are highest for r_s, 0.21, but fairly similar for r_a, 0.13; LEPSOB, 0.13; and LEPSOBMD, 0.12. Note that local values of LEPS ensemble mean skill have very little bias in the sense of Fig. 3c as they are calculated from 42 seasonal values (expected bias near −1%). The chief features are the systematically wider ranges of r_s than r_a and of LEPSOBMD than LEPSOB. For positive values of LEPSOBMD, values of LEPSOB are always less. Note that this is only expected to be true everywhere if the observed and modeled cumulative probability distributions are calculated for the whole period of the analysis, as here. As discussed above, the LEPSOBMD values are more appropriate to assessing the local utility of the model than LEPSOB, which better assesses its absolute skill.

Figure 8e shows box and whisker plots of the four skill measures based on values at each point that summarize the basic differences between the four skill measures as applied to such a typical global meteorological field. This confirms the larger standard deviation of r_s than its LEPS equivalent, LEPSOBMD, which is also deducible from Fig. 4a.

10. Conclusions

The LEPS score or skill score may be used to assess forecasts of continuous and categorical variables, using a common framework. It is doubly equitable and does not suffer from the problem of bending back, which is exhibited by some other scores of this kind. In this respect, the score derived in this paper is an improvement on the earlier form of LEPS discussed in Ward and Folland (1991). Rmse can be reduced by damping forecasts, while anomaly correlation remains the same when forecast anomalies are multiplied by a constant scale factor unless bias is present; the LEPS score is independently sensitive to bias and forecast variance if the latter is less than observed. This will most often be the case in real forecasts or simulations, so LEPS has advantages over both correlation measures. By definition, LEPS skill scores are less sensitive to outliers than correlation but more sensitive to changes in values near the center of the cumulative probability distribution. The form of LEPS used here is slightly transformed from the version used in Ward and Folland (1991), which had the advantage that the basis of LEPS scores was exactly equal to the linear error in probability space. However, as tables of LEPS scores here and in Ward and Folland show [compared in more detail in Folland (1992)], the two forms of LEPS are very similar numerically. So the LEPS scores derived here are *almost* based on the "linear error in probability space." LEPS scores can be used to assess the skill of individual forecasts or simulations, which correlation cannot do for a univariate time series. This is particularly valuable for highlighting unskillful, or negatively skillful, forecasts in time series that may not be immediately obvious by inspection.

The skill score version of LEPS is quite easy to interpret and enables a large number of forecasts to be aggregated. For example, it may be used to assess the percentage skill of forecasts of a meteorological field that are made at grid points. This version of LEPS is no longer equitable, but the bias reduces as the number of independent forecasts increases and can be ignored if the effective number of independent point forecasts exceeds about 20, and is small for as few as five independent forecasts. A matter not dealt with in this paper is the estimation of the statistical significance of LEPS and LEPS skill scores. This will be discussed elsewhere.

In this paper we have concentrated on LEPS scores and emphasized their advantages over other scores. However, we have also noted some drawbacks, and it is clear that a "perfect" score has yet to be devised and may even be unattainable. Con-

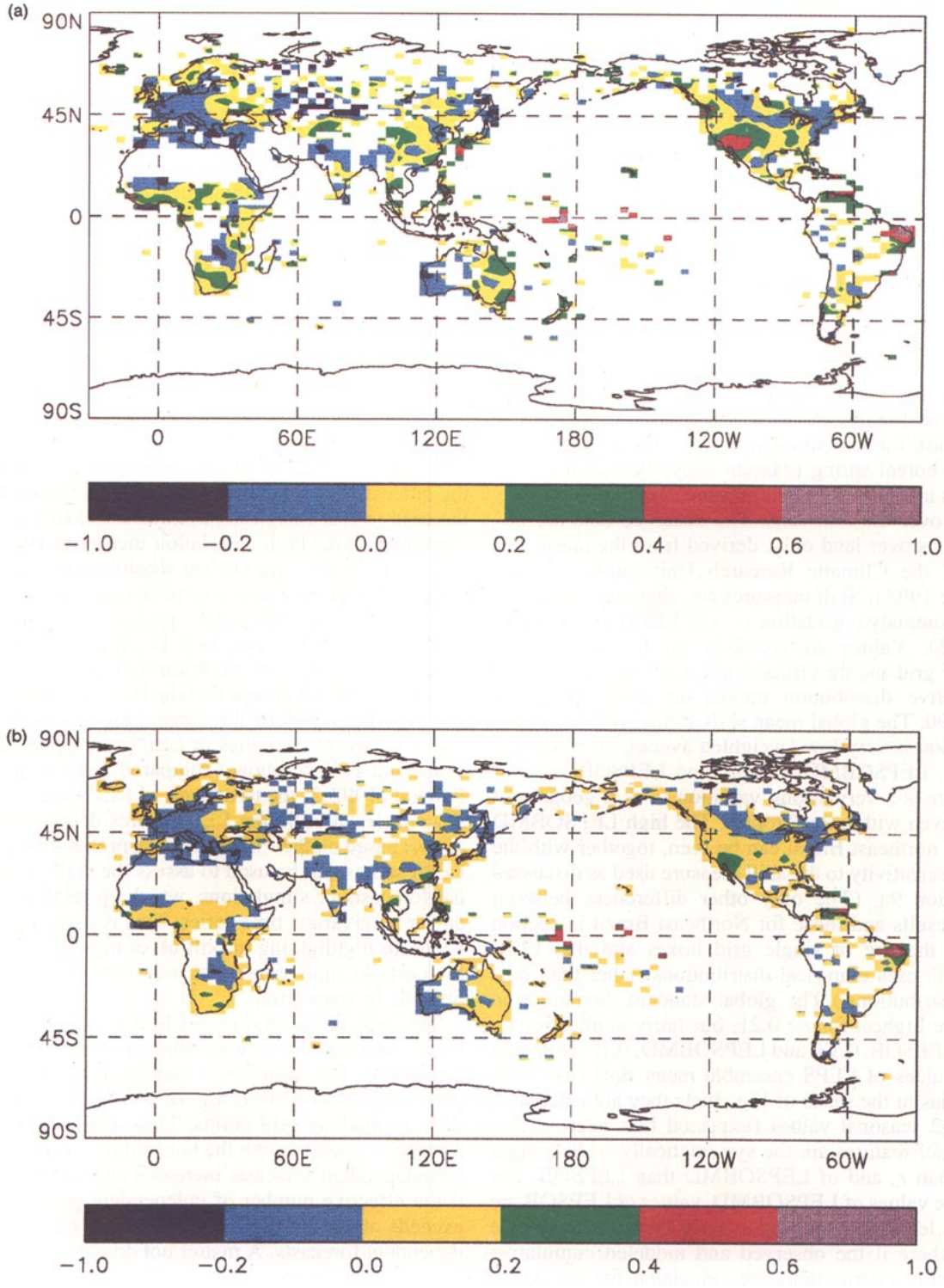
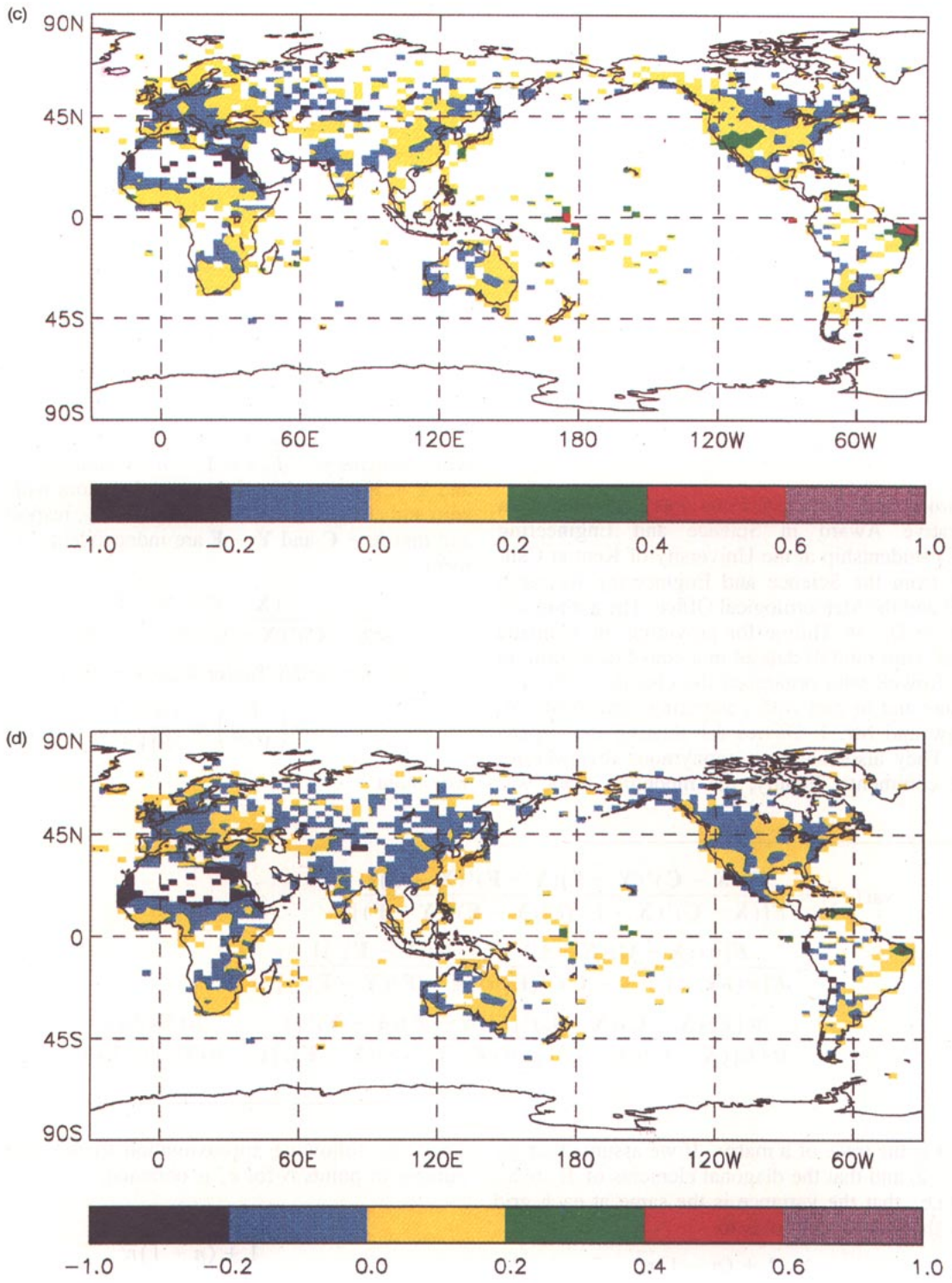


FIG. 8. (a) Standard correlation r_s between mean modeled and mean observed rainfall for the March–May during 1949–90 for areas with sufficient data coincident with the model 2.5° lat × 3.75° long grid. (b) As in (a) but for anomaly correlation r_a. (c) As in (a) but for LEPSOBMD skill expressed as a fraction. (d) As in (c) but for LEPSOB. (e) Box and whisker plots of individual grid box values of r_s and r_a, LEPSOB skill and LEPSOBMD skill. The range in the box is ±1 standard deviation of the skill score about its mean (central line), and the whisker measures the range between the largest maximum and lowest minimum skill values.



structing a score that avoids particular undesirable properties seems inevitably to introduce other problems. In many applications it is advisable to compute more than one score as, together, they may provide information that any one, alone, could not portray (Murphy 1991).

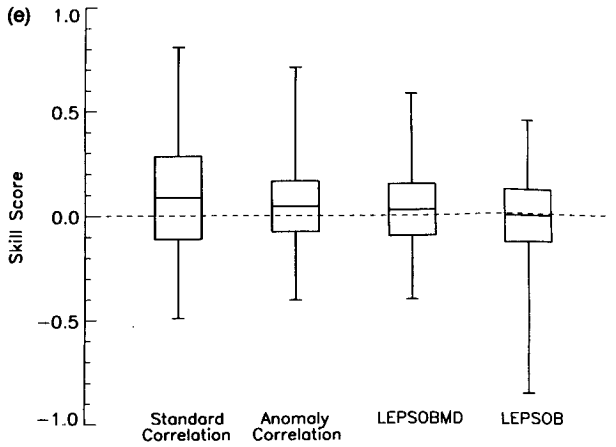


FIG. 8. (*Continued*)

Acknowledgments. J. M. Potts was supported by a Cooperative Award in Science and Engineering (CASE) studentship at the University of Kent at Canterbury from the Science and Engineering Research Council and the Meteorological Office. The authors are grateful to Dr. M. Hulme for providing the Climatic Research Unit rainfall dataset in a convenient form, to Dr. D. Rowell who organized the climate model experiments and helped with computing, and to Ms. A. Renshaw and Mr. J. Davies for running the experiments. They also thank an anonymous reviewer for helpful comments. Finally, the modeling work was funded by Department of the Environment Contract PECD 7/12/37.

APPENDIX

Estimation of the Equivalent Number of Independent Points for the Correlation Coefficient between Spatially Correlated Fields

Clifford and Richardson (1985) give an approximation to the variance of the correlation coefficient between two independent, normally distributed spatial processes. The form of correlation coefficient that they consider corresponds to r_s' as the sample means are subtracted. It is assumed that the two processes have a constant mean at each point. A similar expression is now given for the variance of r_s''.

Let $\mathbf{X} - \mathbf{C}$ denote the $n \times 1$ vector with elements $x_i - \bar{c}_i$, $i = 1 \cdots n$, and $\mathbf{Y} - \mathbf{F}$ denote the $n \times 1$ vector with elements $y_i - \bar{f}_i$, $i = 1 \cdots n$. Assume that $\mathbf{X} - \mathbf{C}$ and $\mathbf{Y} - \mathbf{F}$ are multivariate normal vectors with mean zero and covariance matrices $\Sigma_\mathbf{X}$ and $\Sigma_\mathbf{Y}$, respectively, and that $\mathbf{X} - \mathbf{C}$ and $\mathbf{Y} - \mathbf{F}$ are independent. In vector form

$$r_s'' = \frac{(\mathbf{X} - \mathbf{C})'(\mathbf{Y} - \mathbf{F})}{[(\mathbf{X} - \mathbf{C})'(\mathbf{X} - \mathbf{C})(\mathbf{Y} - \mathbf{F})'(\mathbf{Y} - \mathbf{F})]^{1/2}}.$$

Using the first-order Taylor approximation

$$\mathrm{var}\left(\frac{U}{V^{1/2}}\right) \approx \frac{\mathrm{var}(U)}{E[V]},$$

we obtain

$$\begin{aligned}
\mathrm{var}(r_s'') &\approx \frac{E[(\mathbf{X} - \mathbf{C})'(\mathbf{Y} - \mathbf{F})(\mathbf{Y} - \mathbf{F})'(\mathbf{X} - \mathbf{C})]}{E[(\mathbf{X} - \mathbf{C})'(\mathbf{X} - \mathbf{C})]E[(\mathbf{Y} - \mathbf{F})'(\mathbf{Y} - \mathbf{F})]} \\
&\approx \frac{E[\mathrm{tr}((\mathbf{X} - \mathbf{C})(\mathbf{X} - \mathbf{C})'(\mathbf{Y} - \mathbf{F})(\mathbf{Y} - \mathbf{F})')]}{E[\mathrm{tr}((\mathbf{X} - \mathbf{C})(\mathbf{X} - \mathbf{C})')]E[\mathrm{tr}((\mathbf{Y} - \mathbf{F})(\mathbf{Y} - \mathbf{F})')]} \\
&\approx \frac{\mathrm{tr}\{E[(\mathbf{X} - \mathbf{C})(\mathbf{X} - \mathbf{C})']E[(\mathbf{Y} - \mathbf{F})(\mathbf{Y} - \mathbf{F})']\}}{\mathrm{tr}\{E[(\mathbf{X} - \mathbf{C})(\mathbf{X} - \mathbf{C})']\}\,\mathrm{tr}\{E[(\mathbf{Y} - \mathbf{F})(\mathbf{Y} - \mathbf{F})']\}} \approx \frac{\mathrm{tr}(\Sigma_\mathbf{X}\Sigma_\mathbf{Y})}{\mathrm{tr}(\Sigma_\mathbf{X})\,\mathrm{tr}(\Sigma_\mathbf{Y})},
\end{aligned} \tag{A1}$$

where tr is the trace of a matrix. If we assume that $\Sigma_\mathbf{X} = \Sigma_\mathbf{Y} = \Sigma$ and that the diagonal elements of Σ are all equal (i.e., that the variance is the same at each grid point), then (A1) simplifies to

$$\mathrm{var}(r_s'') \approx \frac{1 + (n-1)\overline{r^2}}{n},$$

where $\overline{r^2}$ is the average of the squares of the correlations between all pairs of points. In the absence of any spatial correlation

$$\mathrm{var}(r_s'') = \frac{1}{n}.$$

Thus, the following approximation to the equivalent number of points N for r_s'' is obtained:

$$N \approx \frac{n}{1 + (n-1)\overline{r^2}}.$$

REFERENCES

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

Clifford, P., and S. Richardson, 1985: Testing the association between two spatial processes. *Statist. Decisions,* **2** (Suppl.), 155–160.

Cullen, M. J. P., 1993: The unified forecast/climate model. *Meteor. Mag.*, **122,** 81–94.

Folland, C. K., 1992: *LEPS scores for assessing climate model simulations and long-range forecasts.* CRTN 33 (unpublished). [Available from the National Meteorological Library, London Rd, Bracknell, Berkshire RG12 2SZ, United Kingdom.]

——, A. Woodcock, and L. Varah, 1986: Experimental monthly long-range forecasts for the United Kingdom. Part III: Skill of the monthly forecasts. *Meteor. Mag.*, **115,** 377–395.

Freeman, M. H., 1967: The accuracy of long-range forecasts issued by the Meteorological Office. *Weather,* **22,** 72–76.

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

Gringorten, I. I., 1965: A measure of skill in forecasting a continuous variable. *J. Appl. Meteor.*, **4,** 47–53.

Heidke, P., 1926: Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst. *Geografiska Annaler,* **8,** 301–349.

Hulme, M., 1992: A 1951–80 global land precipitation climatology for the evaluation of general circulation models. *Clim. Dyn.*, **7,** 57–72.

Miyakoda, K., G. D. Hembree, R. F. Strickler, and I. Shulman, 1972: Cumulative results of extended forecast experiments. Part I: Model performance for winter cases. *Mon. Wea. Rev.*, **100,** 836–855.

Murphy, A. H., 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.

——, and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.*, **117,** 572–581.

Parker, D. E., T. Legg, and C. K. Folland, 1992: A new daily Central England Temperature series, 1772–1991. *Int. J. Climatol.*, **12,** 317–342.

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.