Comparison of Some Statistical Methods of Probabilistic Forecasting of ENSO

SIMON J. MASON
International Research Institute for Climate Prediction, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California

GILLIAN M. MIMMAC
Department of Statistics and Actuarial Science, University of the Witwatersrand, Johannesburg, South Africa

(Manuscript received 6 December 1999, in final form 17 July 2001)

ABSTRACT

Numerous models have been developed in recent years to provide predictions of the state of the El Niño–Southern Oscillation (ENSO) phenomenon. Predictions of the ENSO phenomenon are usually presented in deterministic form, but because of the inherent uncertainty involved probabilistic forecasts should be provided. In this paper, various statistical methods are used to calculate probabilities for monthly Niño-3.4 anomalies within predefined ranges, or categories. The statistical methods used are predictive discriminant analysis, canonical variate analysis, and various forms of generalized linear models. In addition, probabilistic forecasts are derived from a multiple regression model by using contingency tables and from the model’s prediction intervals. By using identical sets of predictors and predictands, the methods are compared in terms of their performance over an independent retroactive forecast period, which includes the 1980s and 1990s. The models outperform persistence and damped persistence as reference forecast strategies at some times of the year. The models have greatest skill in predicting El Niño, although La Niña is predicted with greater skill at longer lead times and with greater reliability. The forecasts for the ENSO extremes are reasonably well calibrated, and so the forecast probabilities are reliable estimates of forecast uncertainty.

All models are wrong, but some are useful. G. Box

1. Introduction

In recent decades there has been a proliferation of statistical models for making extended-range forecasts of the state of the El Niño–Southern Oscillation (ENSO) phenomenon over lead times of several seasons (Graham et al. 1987a,b; Barnett et al. 1988; Xu and von Storch 1990; Barnston and Ropelewski 1992; Keppenne and Ghil 1992; Penland and Magorian 1993; van den Dool 1994; Xue et al. 1994, 2000; Penland and Sardeshmukh 1995; Knaff and Landsea 1997; Tangang et al. 1997, 1998a,b; Berligner et al. 2000; Landman and Mason 2001). These statistical models almost exclusively produce deterministic forecasts and, despite the inherent uncertainty involved, surprisingly little attention has been directed to the importance of indicating forecast uncertainty (Berliner et al. 2000). While prediction intervals for regression-based approaches are generally provided for forecasts of ENSO, it is not clear whether these can be interpreted as adequate estimates of forecast uncertainty: the prediction intervals are obtained from the prediction variance, which is defined in terms of the mean square error of the model over the training period (Wilks 1995) rather than in terms of the inherent unpredictability of the current season. Alternatively, probabilistic forecasts can be obtained from deterministic forecasts using contingency tables, although usually with the loss of some sharpness in the probabilities (e.g., Pan and van den Dool 1998; Mason et al. 1999).

Deterministic forecasts of ENSO are not confined solely to statistical methods. Although seasonal forecasts of the atmosphere derived from dynamical models frequently are presented in probabilistic terms (Harrison 1995; Anderson 1996; Stockdale et al. 1998a; Mason et al. 1999; Goddard et al. 2001), forecasts for ENSO are still presented deterministically with only a few exceptions (e.g., Stockdale et al. 1998b). Given the strong theoretical basis of presenting long-lead forecasts probabilistically (Lorenz 1963, 1984, 1990; Epstein 1969a; Leith 1974; Shukla 1981; Murphy 1990, 1998; Mureau et al. 1993; Tracton and Kalnay 1993; Sivillo et al. 1997; Goddard et al. 2001), and given the poor performance of operational forecasts of ENSO (Barnston et al. 1999; Kerr 2000; Landsea and Knaff 2000), it is surprising...
that so few ENSO forecasts continue to be presented in probabilistic terms.

In this paper, some statistical methods of producing probabilistic forecasts are tested in the context of forecasting monthly Niño-3.4 sea surface temperature anomalies with lead times of up to 11 months. Careful assessments of the operational levels of forecast skill are made by using a retroactive forecast procedure. The performances of the models are compared with that of probabilistic forecasts obtained from multiple linear regression. The methods considered are predictive discriminant analysis, canonical variate analysis, and various forms of generalized linear models. Discriminant analysis has been widely used to produce seasonal forecasts of the atmosphere (Ward and Folland 1991; Hastenrath and Greischar 1993a,b; Casey 1995; Hastenrath et al. 1995; Lehmiller et al. 1997; Mason 1998; Mattes and Mason 1998; Mutai et al. 1998), but the other methods have received less attention.

2. Data and methods

A retroactive forecast procedure has been designed in which forecasts of monthly Niño-3.4 (5°N–5°S, 170°E–120°W) anomalies are made over the 20-yr period 1981–2000. The models are initially trained using data over the previous 30-yr period from 1951 but are updated every 5 yr to provide successively longer training periods. This retroactive procedure has been designed to estimate as closely as possible the operational skill of the model predictions (Armstrong 2001a). Details of the datasets used and of the forecasting procedure are provided in the sections below.

a. Sea surface temperature data

Because ENSO variability is to a large extent the result of ocean–atmosphere variability internal to the tropical Pacific (Neelin et al. 1998), skillful forecasts of central equatorial Pacific sea surface temperature anomalies, for example, can be developed relatively simply using only prior temperatures in the region as predictors (Barnston and Ropelewski 1992; Penland and Sardeshmukh 1995; Latif et al. 1998; Goddard et al. 2001). Although it may be possible to effect small improvements in skill by including variables such as surface wind stress, in the interests of keeping the number of candidate predictors to a minimum, the gridded monthly sea surface temperatures over the tropical Pacific were used as the only predictors. While all skill scores were calculated with reference to strategies of random guessing and climatological probabilities (section 2b), the performances of the various model forecasts were compared to a deterministic strategy of assuming the persistence of the monthly Niño-3.4 anomaly category, and to a strategy of “damped persistence.” For the persistence strategy, a probability of 100% was assigned to the observed category for the month from which the forecast was made. For the damped persistence strategy, probabilities of each of the five categories were defined by calculating the conditional probability of each category given the observed category for the month from which the forecast was made. Such “persistence forecasts” are known to outperform most other forecast strategies at lead times of less than about 3 months (Barnston et al. 1994; Latif et al. 1998; Goddard et al. 2001).

Sea surface temperature data for the 50-yr period January 1951–December 2000 were obtained from the Kaplan et al. (1998) dataset. The data are available at 5°×5° resolution, and only data for the domain 25°N–25°S, 110°E–70°W were retained, yielding 360 variables. Principal components were calculated from the correlation matrix of monthly anomaly data for the first 30 years, and for each month separately: namely, the period 1951–80. In most cases, the scree test (Jackson 1991) suggested that five unrotated principal components should be retained, explaining between about 60% and 70% of the variance. For the sake of consistency, five components were retained for all months.

Correlations between the sea surface temperatures and the principal component scores for May are shown in Fig. 1 as an example of typical loadings patterns. Given that the principal components loadings are unrotated, the spatial patterns should be interpreted with caution (Richman 1986). For all months, the first principal component represents sea surface temperature variability over the central Pacific Ocean, focusing on the Niño-3.4 region (top panel). The other components indicate off-equatorial variability, except for component two, which indicates variability in the eastern equatorial Pacific for May–September (second panel). Temporal scores for these five principal components were calculated for the 20-yr independent period, January 1981–December 2000, in addition to the 30-yr training period. The principal components were not redefined as the training period was updated.

Forecasts of monthly Niño-3.4 sea surface temperature anomalies were produced at lead times of between 0 and 11 months. The forecasts were made using the principal component scores of antecedent monthly mean sea surface temperatures. The predictors consisted of data for only the most recent month available, and so no information on the recent evolution of sea surface temperatures was used. The statistical methods used in this paper require a categorical predictand, and so for each month, the Niño-3.4 anomalies were grouped into five equiprobable categories over the training periods. Over the 1951–80 training period, the six coldest January values, for example, were classified as indicating “La Niña” conditions, the next six coldest as “cool” conditions, the next six as “normal” conditions, the next six as “warm” conditions, and the warmest six as “El Niño” conditions. As the training period was updated to 1951–85 (see section 2b), the seven coldest January values were classified as indicating La Niña conditions,
Discriminant analysis has become a standard approach to probabilistic statistical climate prediction (Ward and Folland 1991; Hastenrath and Greischar 1993a,b; Casey 1995; Hastenrath et al. 1995; Lehmiller et al. 1997; Mason 1998; Mattes and Mason 1998; Mutai et al. 1998). The method is used to assign a case or observation to one of $n_g$ groups or categories on the basis of a set of $m$ independent variables (Wilks 1995). In our case the method is used to estimate the probability of the monthly Niño-3.4 anomaly indicating La Niña, cool, normal, warm, and El Niño conditions (as defined in the previous section), given the observed antecedent sea surface temperatures, which are incorporated in the
principal component time scores. These five posterior probabilities are inversely proportional to the Mahalanobis distances between the observation and the five \( m \)-dimensional group means (appendix A). A predictive approach to discriminant analysis (Aitchison et al. 1977; Moran and Murphy 1979; Huberty 1994) has been used in this paper. Further details on discriminant analysis and the difference between estimative and predictive approaches are provided in appendix A. Tests conducted to determine whether to use linear or quadratic discriminant analysis are discussed below in section 2i.

c. Canonical variate analysis

Canonical variate (or canonical discriminant) analysis provides an optimal way of discriminating among groups when there are many variables (Huberty 1994; Manly 1994). The primary objective is to create new variables, called canonical variates, from linear combinations of the original variables, to maximize the differences in the means of these new variables between the categories. The derivation of the canonical variates is explained in appendix B. Once the canonical variates have been defined, posterior probabilities for each category can be calculated in the same way as for discriminant analysis.

By maximizing the intercategory differences in the means of the new variables rather than maximizing the number of correct category allocations, canonical variate analysis can be interpreted as a form of descriptive discriminant analysis (Huberty 1994). As such, canonical variate analysis is not designed explicitly for prediction.

d. Generalized linear models

Generalized linear models have not been widely used in the climate literature. Models with multinomial errors can be used to relate probabilities of events, such as sea surface temperature anomalies exceeding predefined limits, and have a number of advantages over regression estimation of event probabilities (Wilks 1995). A brief introduction to the models used in this paper is provided in appendix C. Generalized models with multinomial errors were fit with parallel and nonparallel planes. Logistic, probit, and complementary log–log links were used. The logit model is equivalent to polytomous (polychotomous) logistic regression (Hosmer and Lemeshow 1989; Agresti 1990, 1996). Logistic regression and linear discriminant analysis are closely related, since the usual discriminant function model is the logistic regression function (Hosmer and Lemeshow 1989). While some use has been made of logistic regression in forecasting (Brelsford and Jones 1967), the application of polytomous logistic regression (Jones 1968), and of probit and complementary log–log models have been limited.

e. Multiple linear regression

To compare the performance of the probabilistic methods of forecasting described above with relatively simple methods of obtaining forecast probabilities, multiple linear regression was used. Probabilistic forecasts were obtained from the model in two ways. In the first method, the probabilities were calculated from the prediction intervals: the area beneath the prediction envelope was integrated between the category boundaries to give the forecast probabilities. In the second method, forecast probabilities were obtained from a contingency table that defines the frequency distribution for each category contingent upon the mean response of the prediction (Pan and van den Dool 1998). (For the sake of simplicity, hereinafter “multiple regression” is used to refer to the first method, and “contingency table” to the second.) For the first method, no variance adjustment, or “inflation” (von Storch 1999), was deemed necessary since the focus in this paper is on the prediction interval, which should not be unduly affected by the underestimation of the variance of the mean response. However, when calculating the contingency table, category boundaries for the predictions were redefined so that the mean response was equiprobable for each category over the training period. Similar results are obtained if no such variance adjustment is performed.

f. Model selection

The optimal combination of predictors for the discriminant analysis, generalized linear regression, multiple regression, and contingency table models was identified using a procedure that is similar to the “maximum-posterior-probability/leave-one-out” method of variable selection advocated by Huberty (1994). Model parameters were estimated using all possible combinations of one or two variables (from the five available principal components), and the set of predictors that provided the best cross-validated fit over the training periods was selected. The cross-validation window was defined as 5 yr to ensure that the categories remain equiprobable. The goodness of fit was measured by calculating the ranked probability score (Epstein 1969b; Daan 1985; Wilks 1995) over the training period. The number of variables was kept to a maximum of 2 in response to Lachenbruch’s (1968) recommendation to have a minimum sample size of

\[
\text{n} = 3 \times m_u \times n_g, \tag{1}
\]

where \( m_u \) is the number of variables used, and \( n_g \) is the number of groups. Given that \( n_g = 5 \), and the shortest training period used was 30 yr, \( m_u \) is constrained to be 2.

For the canonical variate analysis model, all five predictors were included, but the number of retained canonical variates was varied. The selection criteria for the canonical variate analysis model therefore differ slightly from those for the other models. A total of \( n_g - 1 \) ca-
nonical variates will explain all the between-group variance, and so the maximum number of canonical variates was 4, but to maintain some form of similarity to the other models, the maximum number of canonical variates retained was restricted to 2. The first variate was always retained; the second variate was included only if doing so resulted in an improvement in the cross-validated ranked probability score.

The model selection criteria used in this paper have the advantage of attempting to maximize the reliability of the forecast probabilities (a forecast system is said to be well calibrated when the forecast probabilities are reliable, i.e., they approximate the observed relative outcome of the event), without compromising their sharpness, or refinement. [See Murphy (1993) and Wilks (1995) for detailed discussions and definitions of these various attributes of forecast performance.] A stepwise procedure using the various standard tests for parameter significance (McCullagh and Nelder 1989; Agresti 1990, 1996; Huberty 1994) was not adopted because the objectives of such tests are not concerned explicitly with forecast calibration, and also for reasons of attempting to standardize model selection criteria as much as possible.

A disadvantage of optimizing the cross-validated skill scores is that the computational cost of testing all possible combinations of predictors can be high. The computational expense is greatest for the generalized linear regression models because their parameters have to be estimated using iterative procedures (Hosmer and Lemeshow 1989). However, the number of iterations required was usually less than 10, and initial estimates of parameters are not required for the iterative least squares algorithms used (McCullagh and Nelder 1989).

g. Retroactive forecasting

In order to obtain realistic estimates of operational prediction skill using the models tested in this paper, model performance was assessed using a retroactive forecast procedure in which the testing period was completely independent of the training period (van den Dool 1987). The training period was updated every 5 yr so that the five categories can remain equiprobable. The training period was initially set as 30 yr (1951–80), and retroactive predictions for the following 5 yr were then made using the optimal model. After this 5-yr period the model was retrained over the period 1951–85, possibly selecting different variables and a different number of retained variables, and predictions for 1986–90 were made. This procedure was repeated until a set of 20 yr of retroactive predictions had been made. At each stage, the definitions of the five categories were reset to ensure that the categories remained equiprobable a priori.

While the categories are defined as equiprobable over the training periods, this is not necessarily the case for the verifications over the independent period. For 1981–85, the verifications were categorized on the basis of the 1951–80 training period; the verifications for 1986–90 were categorized on the basis of the 1951–85 training period, and so on.

h. Model validation

Rather than reducing the probabilistic forecasts to deterministic forecasts by simply selecting the category with the highest posterior probability, the models have been validated using a variety of probabilistic performance scores. The ranked probability score (RPS) (Epstein 1969b; Murphy 1971; Daan 1985; Wilks 1995), the half-Brier score (Brier 1950; Murphy 1973; Wilks 1995), including Murphy’s (1973) decomposition, unconditional bias (Wilks 1995, 2000), relative operating characteristics (ROC) areas (Mason 1982; Harvey et al. 1992), and the version of the linear error in probability space (LEPS) score for probabilistic forecasts (LEPSPROB; Ward and Folland 1991) using the rescaled LEPS score coefficients for five categories (Potts et al. 1996), were calculated. These scores were calculated using the 20 yr of retroactive predictions, which should be sufficient to give reasonably accurate estimates of operational forecast performance (Montgomery and Peck 1992; Barnston et al. 1994; Wilks 1995).

The RPS and LEPSPROB scores provide overall indications of the quality of probabilistic forecasts. The RPS (Epstein 1969b; Murphy 1971; Daan 1985; Wilks 1995) is a squared distance measure comparing forecast performance with that achieved using a perfect deterministic forecast system. A perfect RPS has a value of 0.0, and the score generally decreases toward 0.0 as more probability is assigned to the category that actually occurred, although the exact value of the score depends also on the distribution of probability over the categories that did not verify. The RPS can be converted to a skill score, the ranked probability skill score (RPSS), in which a forecast system using the climatological probabilities of events or random guessing (or any other reference forecast strategy) has a score of 0.0, and skill is indicated by positive values up to a maximum of 1.0 for perfect skill (Wilks 1995). The RPSS is maximized when an event is correctly forecast with a high probability. As with other skill scores, a score of 0.0 is achieved when the forecast strategy has no skill relative to the reference forecast strategy, and a perfect deterministic strategy achieves a score of 100% (or 1.0). The LEPSPROB score similarly is high when an event is correctly forecast with a high probability but gives greater weight to forecasts of extreme categories. Because of the greater weight given to forecasts of extreme categories, the LEPSPROB is not a strictly proper scoring rule since the forecaster is encouraged to hedge; by increasing the forecast probabilities for the extreme categories a higher score can be achieved (appendix D). The RPSS was therefore used in preference to the LEPSPROB scores in the model selection procedure. The LEPSPROB scores
typically are smaller than LEPS scores (Ward and Folland 1991).

The half-Brier score and the ROC areas provide estimates of forecast skill for specific categories and so were calculated for the five categories separately. The half-Brier skill score was calculated using climatological relative frequencies as the reference strategy (Wilks 1995). The ROC area is not strictly an estimate of forecast skill but is a useful indicator of forecast quality (Mason and Graham 1999; Richardson 2000). ROC areas range between 0.0 and 1.0, with a score of 0.5 indicating zero forecast skill, and areas greater than 0.5 indicating potentially useful forecast strategies. The ROC area can give a perfect score even when the forecast probabilities do not correspond with the observed probability of an event, since over- and underconfidence are not penalized and neither are unconditional biases. Despite these apparent shortcomings of the ROC area, it is a first step toward estimating forecast value (Murphy 1973; Wilks 1995). To calculate the half-Brier skill score, the following formula was used:

$$\kappa = 1 - \frac{2m_c}{[m_{\text{max}}n(n-1)]}.$$  

where \(m_c\) is the total number of changed predictors at each new training step as compared with all previous steps, \(m_{\text{max}}\) is the maximum number of predictors selected (in our case \(m_{\text{max}} = 2\)), and \(n\) is the total number of training periods. Given that the independent testing period is 20 yr, and that the models were updated every 5 yr, four training periods are defined. The models were also trained over the full 50 yr, providing a fifth training period. The consistency index defined by Eq. (2) defines the ratio of the number of changed predictors to the maximum possible number of changes. It can range from 0.0, in which case different predictors are selected each time, to 1.0, in which case the same set of predictors is selected at each update. [Because of an insufficient number of candidate predictors, complete inconsistency in the selection of predictors is not possible. The minimum value of \(\kappa\) is 0.200, and so the value \(\kappa^* = 1 - (1 - \kappa)/(1 - 0.200)\) has been calculated to ensure that \(\kappa^*\) ranges from 0.0 to 1.0.] From repeated random selection of predictors, it was found that the median value of the consistency index was 0.188, the 90th percentile was 0.375, and the 99th percentile was 0.563. For the canonical variate analysis model, the two predictors with the largest absolute canonical loadings were selected.

\[\text{i. Assumptions}\]

In discriminant and canonical variate analysis it is usually assumed that the population of independent variables is multivariate normal (Lachenbruch et al. 1973; Clarke et al. 1979; Huberty 1994). The assumption of the (univariate) normality of the distribution of monthly subsets of the principal component scores was tested using Shapiro and Wilk’s \(W\) statistic (Royston 1982). The statistic was recalculated each time the training period was updated (section 2g). Out of the 5 series and 12 months, only two of the predictors had a distribution over the initial 30-yr training period that was significantly different from a normal distribution at a 95% level of confidence. These two predictors were the June values of the fourth principal component, and the July values of the fifth principal component. Over all training periods, nonnormality was indicated in a few additional instances, and was flagged in 6% of the models constructed, which is only slightly more than would be expected by chance. Rather than highlighting these few instances (Nicholls 2001), transformation of the data (to correct for departures from normality) was deemed unnecessary in all cases.

The multivariate normality of the predictors used was tested for each month using the sample multivariate measures of skewness and kurtosis (Mardia et al. 1979; Stephenson 1997). The tests were conducted for each selected combination of predictors (section 2f), and each time the training period was updated (section 2g). Measures of skewness differed significantly from 0 in only about 3% of the selected models, so it is reasonable to
assume that the distributions are approximately symmetric in each dimension. The measure of kurtosis was significantly different from that expected of a multivariate normal distribution in only 0.5% of cases.

Distributional assumptions concerning the monthly Niño-3.4 anomalies were made only for the multiple linear regression model: the data are treated as categorical by the other models. The distribution of the Niño-3.4 anomalies was not found to be significantly different from normal in any month or for any of the training periods. This result is in agreement with the conclusions of Burgers and Stephenson (1999).

In linear discriminant analysis the further assumption of equality of variance between the different categories is often adopted, but if this assumption is relaxed, the classification function becomes quadratic (Marks and Dunn 1974; Aitchison et al. 1977; Seber 1984; Huberty 1994; Manly 1994). Tests for the inequality of variance were conducted to determine whether to relax the assumption on a case-by-case basis. Equal variance was assumed except when the test statistic was significant beyond the 0.05 level since the relaxation of the assumption gives improved results only when there are marked differences in variance (Aitchison et al. 1977; Moran and Murphy 1979). Linear discriminant analysis is sensitive to the assumption of multivariate normality (Halpern et al. 1971; Lachenbruch et al. 1973; Hosmer et al. 1983; Hosmer and Lemeshow 1989), while quadratic discriminant analysis is reasonably robust to non-normality, except when the data are highly skewed (Clarke et al. 1979), which was not the case in this paper. Quadratic discriminant analysis was therefore conducted when the test statistic for kurtosis was significant beyond the 0.05 level, or, in the case of models with only one predictor, if the univariate test for normality was failed, regardless of the test for equal variance.

An advantage of the generalized linear models considered in this paper is that there are no distributional assumptions: neither normality nor equality of covariance matrices is required (Press and Wilson 1978), although tests for multicollinearity should generally be conducted (McCullagh and Nelder 1989). Because of the greater robustness of logistic regression it is often preferred over linear discriminant analysis (Press and Wilson 1978), despite being less efficient (Efron 1975). The predictors used in this paper are not necessarily orthogonal, despite the calculation of principal components, because subsets are used (e.g., by selecting all the January values), and because of the updating of the training periods. This problem could be avoided by calculating principal components for each month separately, and by updating the principal components as the training period is updated.

A possible additional advantage of the generalized linear models is that they exploit the ordinal nature of the five categories (appendix C), whereas discriminant and canonical variate analysis models do not. Although the ordering of the categories by the generalized linear regression models has the obvious advantage that the models “know” that the warm category is closer to the El Niño category than the cool category is, this ordering does impose the restriction of a monotonic relationship between each of the predictor(s) and the predictand.

3. Results

a. Selection of predictors

In addition to assessing and comparing the skill levels of the various models, it is worth considering whether the models select similar predictors, and whether there is any consistency in the selection of predictors as the training period is updated. All the models selected the first and fourth principal components as predictors more frequently than the other components. These two principal components were each selected about 35% and 30% of the time, respectively. The other components were selected about 10%–15% of the time. It seems reasonable that principal components one and two should be selected as predictors of the Niño-3.4 index, since both components are indicative of different stages of ENSO development (Fig. 1). Reasons for the frequent selection of the fourth component are less clear.

The consistency index averaged across the 12 lead times and 12 seasons is larger than the value that would be exceeded by chance only 1% of the time for all eight models. Given that the training periods have many years in common, it is not surprising that the consistency indices are so high. Of greater interest than the absolute values of the indices, however, are the relative values. The canonical variate analysis and contingency table models are the least consistent in the selection of predictors, while the multiple regression model is most consistent, but differences between the models are minor.

b. Retroactive forecast performance

Retroactive RPSSs have been calculated for forecasts at separate lead times, and are shown in Figs. 2 and 3, where they are compared with scores for damped persistence and persistence forecasts. In Fig. 2 the RPSSs are calculated with reference to a strategy of random guessing. At all lead times beyond 0 the scores for all the models, except for the contingency table, exceed the scores for the persistence forecasts. The inability to outscore persistence at the shortest lead time is a ubiquitous problem with dynamical and statistical models (Latif et al. 1998; Goddard et al. 2001). While the skill scores for persistence decrease to about 0 at lead times of 11 months, the scores for the models remain strongly positive.

In Fig. 3, the RPSSs are calculated relative to a strategy of forecasting the climatological probabilities of each of the five categories. In this case, the skill of persistence forecasts decreases to 0 after only about 3 months and continues to decrease at longer lead times.
The persistence forecasts assign 100% of the probability to one category, and so even a one-category miss can be penalized heavily. For most of the model forecasts, however, skill scores remain positive out to about 9- or 10-months lead time. Of all the models, the skills for the proportional odds and multiple regression model remain positive for the longest lead times, while again the contingency table model performs least well. The multiple regression model provides the best forecasts at the longest lead times.

It is evident that the climatological forecast probabilities are more difficult to beat than the persistence of
the anomaly category (cf. Figs. 2 and 3). Skill scores for forecasts of persistence damped toward climatology were therefore calculated (section 2a), and are shown by the light gray bars in Figs. 2 and 3. These forecasts provide the highest skill for lead times less than about 4 months, outscoring all the models and the simpler persistence forecasts, but the models provide more skillful forecasts at the longer lead times.

The LEPSPROB skill scores indicate less promising results, although the scores for the models are positive at all lead times (Fig. 4). However, the simple persistence forecasts achieve much larger scores than the mod-
The extremely high LEPSPROB scores for the persistence forecasts are an effect of the way LEPSPROB encourages the forecaster to hedge by assigning a probability of 1.0 to the category that is considered most likely (appendix D). The LEPSPROB score therefore favors the strategy with the sharper probabilities, even if those probabilities are not totally reliable. As for the RPS scores, the LEPSPROB scores for the contingency table indicate the poor performance of this approach as compared with that of the other models.

To compare differences in the reliability of the forecasts from the models, reliability diagrams were constructed for each of the five categories and are shown in Fig. 5. Diagrams for persistence were not constructed.
Fig. 5. Reliability diagrams for retroactive forecasts at increasing lead times of La Niña (solid thin line), cool (dashed thin line), normal (dotted line), warm (dashed thick line), and El Niño (solid thick line) conditions for the 20-yr period Jan 1981–Dec 2000. Forecasts at all lead times and for all months are pooled. The histograms indicate the frequency of forecasts with probabilities in the ranges 0.0–0.05, 0.05–0.15, 0.15–0.25, . . . , 0.95–1.0. The y axes range to 1250. The top histogram is for El Niño conditions, the second top for warm conditions, and the rest as indicated.
since only probabilities of 0% and 100% were used and because the forecast probabilities are necessarily unreliable given that hit rates are imperfect. Instead, diagrams for damped persistence were constructed and are shown in place of reliability diagrams for the contingency table, which has been shown above to perform relatively poorly. The reliability curves indicate good reliability for forecasts of La Niña by all the models, and reliability is exceptional for the canonical variate analysis model. The forecast probabilities for this category are reasonably sharp, although less so than for forecasts of El Niño. The forecasts for El Niño show good reliability also, although there is a tendency toward overconfidence when forecast probabilities are high. The only exception is for damped persistence, which has a slight negative unconditional bias.

The multiple regression model provides good reliability, including minimal unconditional bias, for all five categories, even though prediction intervals are usually too wide (Chatfield 2001). For the other models, however, the reliability of forecasts for the three intermediate categories is not as high as for La Niña and El Niño. In most cases, however, the curves are upward sloping for probabilities below about 50%, indicating that the models are able to provide more reliable indications of diminished probabilities of the intermediate categories than of increased probabilities. For some of the extremely high forecast probabilities, the poor reliability suggested by the diagrams may be partly a sampling problem since there are few occurrences of very high probabilities.

Although the reliability plots cannot be constructed for the different lead times because of insufficient sample size, indications of the reliability of the forecast probabilities at increasing lead times can be assessed using the Murphy (1973) decomposition of the half-Brier score, which can be calculated with fewer forecasts than are required for reliability curves. Unfortunately, since the reliability score of the decomposition is partly a function of forecast refinement (or sharpness), comparison of the scores is difficult. Besides, differences in the reliability scores between models, lead times, and categories are small. Instead, differences in the half-Brier scores are dominated by the resolution term. The half-Brier skill scores are shown in Fig. 6. The greater skill at predicting El Niño conditions compared to any of the other categories is clearly evident. Notable skill at predicting La Niña conditions is evident also, and exceeds that of El Niño at lead times of greater than about 6 months. There is only very weak skill for the other categories. Forecasts for the cool category are poor at all lead times, partly because of an unconditional bias (forecast probabilities for this category were consistently too high), but also because of poor forecast resolution as indicated by the reliability curves (see Fig. 5). Bias statistics indicate that this positive bias was at the expense of negative biases for the El Niño category at all lead times. The models were thus unable to indicate completely successfully the predominance of El Niño conditions over the independent period, and overforecasted the occurrence of negative anomalies.

Unconditional biases and imperfect reliability curves can effect poor half-Brier and ranked probability skill scores, without necessarily detracting from forecast value (Richardson 2000). ROC areas can give a good indication of forecast resolution without requiring perfect reliability (Mason and Graham 1999) or penalizing unconditional bias, and so in some sense provide a less stringent measure of forecast skill than the half-Brier score. The ROC areas confirm the ability to forecast the El Niño category more successfully than the other categories at lead times less than about 5 months, the La Niña category for longer lead times (Fig. 7). In addition, some skill is suggested for the intermediate categories, although at short lead times the damped persistence forecasts cannot be outscored. As with the RPSSs, differences in the ROC areas between the models are much less than differences between lead times. In general, however, the ROC areas for the multiple regression model are the highest at the longer lead times, and for the intermediate categories.

The ROC areas indicate a more positive picture of model skill compared to the half-Brier skill scores. This difference is probably an indication that the forecast probabilities are imperfectly reliable, but that the probabilities are at least reasonably well resolved. The reliability curves in Fig. 5 support this interpretation. It may be possible to improve on the reliability of the forecast probabilities by considering the level of agreement between the models, as considered in the following section.

c. Combined forecast performance

A combined forecast was calculated by simply averaging the forecast probabilities from the various models (except for the contingency table model). No attempt was made to weight the probabilities from the different models by a measure of model skill, since it has been shown above that ranked model performance is sensitive to the precise skill measure used, and can be conditional upon the actual outcome (Mason and Graham 1999), and differences in model skill are relatively minor anyway. Weighting is more effective when there are large differences in model skill (Armstrong 2001b). Reliability diagrams for the combined forecasts were constructed, as shown in Fig. 8. The combined forecasts show some improvements in reliability for the normal and warm categories, except when compared to the multiple regression model (Fig. 5). The improvements in reliability of the combined forecast probabilities compared to the forecasts from the probabilistic models (top 6 panels of Fig. 5) are not simply a result of including the reliable forecast probabilities from the multiple regression model in the averaging. Small improvements
Fig. 6. Half-Brier skill scores for retroactive forecasts at increasing lead times of La Niña (black bars), cool (dark gray bars), normal (gray bars), warm (light gray bars), and El Niño (white bars) conditions for the 20-yr period Jan 1981–Dec 2000.

in reliability are achieved if the probabilities from the six probabilistic models are averaged (not shown).

d. Seasonal dependence of forecast skill

A seasonal variation in the predictability of ENSO has been noted previously, with lower skill being generally observed for predictions that extend through boreal spring (Balmaseda et al. 1995; Webster 1995; Davey et al. 1996; Moore and Kleeman 1996; Latif et al. 1998; Goddard et al. 2001). The RPSSs were therefore calculated for each month separately, using the combined forecast probabilities from the seven models (again the contingency table is excluded), and comparing the fore-
casts with a strategy of climatology. The scores for 6 months are shown in Fig. 9, where they are compared with the skill of damped persistence and simple persistence forecasts. The seasonal dependence of skill is clearly apparent for both the model and the persistence forecasts. Forecast skill for Niño-3.4 at any time of year drops rapidly for forecasts extending through about April and May. Forecasts for January and March are therefore skillful at long lead times (top panels), whereas forecasts for May and July are skillful only at short lead times (middle panels).

Although detailed comparisons of the skill scores should be avoided because of the small number of forecasts involved, the model forecasts outscore persistence
and damped persistence only at certain times of the year. While the models beat persistence at all times of year for lead times longer than about 3 months, damped persistence is a much tougher standard to beat; forecasts for March, for example, are more skilful than damped persistence only for lead times of longer than 10 months. The strength of the models is most clearly evident for May and July, when the combined model forecasts beat damped persistence at all lead times.

4. Discussion and summary

A detailed validation of a selection of probabilistic statistical models for predicting ENSO has been presented. The models considered are discriminant analysis, canonical variate analysis, various forms of generalized linear regression, and two methods of converting multiple linear regression model predictions to probabilistic forecasts, namely from the prediction intervals, and by using contingency tables. The intention in this paper was not to construct the ideal model for forecasting ENSO, but rather to demonstrate a number of useful alternative statistical methods for generating probabilistic forecasts. The methods have application beyond that of forecasting ENSO, and could be used for forecasts at a range of timescales (e.g., Carter and Elsner 1997).

The models were trained over a 30-yr training period and then used to produce 20 yr of retroactive forecasts of monthly Niño-3.4 sea surface temperature anomaly categories, with the models being updated every 5 yr. Five categories of anomalies were defined ranging from “La Niña,” through “cool,” “normal,” “warm,” to “El Niño.” Probabilities for each of the categories over the 20-yr retroactive period January 1981–December 2000 were calculated. This period includes the early 1990s when the inherent predictability of ENSO is generally considered to have been weak (Chen et al. 1995; Balmaseda et al. 1995; Ji et al. 1996; Goddard and Graham 1997; Kirtman and Schopf 1998; Landman and Mason 2001). The retroactive forecasting procedure has been designed to mimic an operational procedure as closely as possible, and the skill estimates are considered to be accurate estimates of operational skill.

Forecast skill was demonstrated for all the models at lead times of between about 4 or 5 months and 10 months, but with a seasonal dependence to the skill. In comparison with a reference strategy of forecasting climatological probabilities, skill scores for the models are highest for forecasts of January–March temperatures, but damped persistence provides slightly better forecasts at fairly long lead times for this time of year. The models are able to beat persistence and damped persistence at all lead times for the May–July period and to provide useful forecasts for September–November from as early as January. While these skill levels are apparently weak when compared with some early estimates of the predictability of ENSO (e.g., Barnston et al. 1994), they do compare favorably with more recent and realistic estimates of operational skill (Barnston et al. 1999; Landsea and Knaff 2000; Landman and Mason 2001). A properly designed retroactive forecasting procedure should provide more realistic estimates of forecast skill than a cross-validated procedure: cross-validated estimates of skill can be inflated if the cross-validation window is insufficiently wide or if there are trends and changes in dynamics in the system. The retroactive forecasting procedure used in this paper was designed to minimize artificial skill as much as possible. In addition, the validation period considered here involves the early 1990s when predictability is considered to have been relatively weak (Chen et al. 1995; Balmaseda et al. 1995; Ji et al. 1996; Goddard and Graham 1997; Kirtman and Schopf 1998; Landman and Mason 2001). It is of interest to note that all the models in this paper indicated high probabilities of El Niño conditions for December 1997 as early as January 1997 and generally strengthened these probabilities as the year progressed. However, the rapid onset of La Niña conditions in May 1998 was less successfully indicated.

Forecast skill is apparently conditional upon the ENSO state: most of the skill achieved is a result of the predictability of El Niño events, although the skill at forecasting La Niña events is also high, and forecast probabilities are more reliable than for El Niño. Only weak evidence of an ability to forecast intermediate conditions could be identified, and that at short lead times. The lack of skill in forecasting intermediate conditions has been attributed in part to questions of definition rather than to physical–dynamical considerations (van den Dool and Toth 1991). Because the categories are defined as equiprobable only over the training period, and not over the verification period, apparent differences in the predictability of the different ENSO states may

Fig. 8. Reliability diagram for retroactive combined forecasts at increasing lead times of La Niña (solid thin line), cool (dashed thin line), normal (dotted line), warm (dashed thick line), and El Niño (solid thick line) conditions for the 20-yr period Jan 1981–Dec 2000. Forecasts at all lead times and for all months are pooled. The histograms indicate the frequency of forecasts with probabilities in the ranges 0.0–0.05, 0.05–0.15, 0.15–0.25, . . . , 0.95–1.0. The y axes range to 1700. The top histogram is for El Niño conditions, the second top for warm conditions, and the rest as indicated.
be artificial, at least in part. However, the consistency of differences in scores between categories for the various reference forecasting strategies used in this paper suggests that predictability is conditional upon the ENSO state to some extent (cf. van den Dool and Toth 1991). While it would be ideal to be able to forecast all ENSO states reliably, in most applications it is forecasts of the extreme events that are potentially most beneficial (Easterling 1986; Mjelde et al. 1993).

There are no obvious reasons for preferring any one of the models considered in this paper, except that deriving probabilities from multiple regression by using a contingency table appears to be suboptimal. Instead, greater skill can be obtained from a multiple regression model by using prediction intervals. This approach uniquely gave reliable forecasts for all five categories, and achieved the highest skill scores for lead times longer than about 6 months. The apparent inferiority of the probabilistic methods over the multiple regression model may be largely attributable to sensitivity to sampling errors. By categorizing the predictand, the degrees of freedom are reduced, and so model parameters are subject to greater sampling error. It is for this same reason that the forecast probabilities derived from a multiple regression model by using a contingency table are less reliable than those derived using the prediction intervals. The degrees of freedom can be increased by reducing the number of categories from five to three. When all analyses were repeated using a three-category system, the differences in model skill were weakened notably (results not shown).

Greater sampling errors in the estimation of param-
meters for the probabilistic models because of the weakened degrees of freedom in comparison with those of the regression model, are not the only source of greater forecast errors. Sampling errors in the parameter estimates when calculating the cross-validated forecast probabilities may translate into occasional suboptimal selections of predictors, as reflected in the weaker consistency scores of the probabilistic models (section 3a). If Niño-3.4 is used as the only predictor of itself in place of the five principal components used in this paper, notable differences in model skill again diminish. While the principal components used as predictors in this paper are simplistic (in the sense that subsurface data and the evolutions of surface temperatures over previous months are ignored), the problems arising because of sampling errors demonstrate that careful preselection of predictors is advisable.

Sampling errors can be compensated for to some extent by using a selection of statistical models. Just as in seasonal climate forecasting, there is advantage in considering output from more than one general circulation model (Mason et al. 1999; Goddard et al. 2001), so also the reliability of the Niño-3.4 forecasts can be improved by considering forecasts from a range of models. A simple average of the forecast probabilities may be sufficient, as long as there are no major differences in the skill of the models. More sophisticated methods of forecast combination, and/or careful and detailed diagnostics of the strengths and weaknesses of each model enabling subjective forecast combination, may effect further improvements in forecast quality (Krishnamurti et al. 1999, 2000). However, further research clearly is required to improve the reliability of long-lead forecast probabilities (Wilks 2000).

The results of the probabilistic models considered in this paper indicate that these methods may deserve further attention in climate research. Although probabilistic forecasts obtained from the prediction intervals of a multiple regression model were shown to be slightly superior in quality, this can be attributed, at least in part, to differences in degrees of freedom. The forecasts from the probabilistic models were clearly superior to those from the multiple regression model when forecast probabilities were derived from a contingency table. The use of contingency tables for converting deterministic to probabilistic forecasts is reasonably common. The results in this paper suggest that there may be simple alternative and more reliable methods that can be applied.

Acknowledgments. This paper was funded in part by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NOAA). The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its subagencies. Initial funding for the research was provided in part by Technology Research and Investigations, a division of the Technology Group of Eskom. The comments of D. Wilks, F. W. Zwiers, and two anonymous referees contributed to a substantial improvement in this paper. Helpful discussions with A. Barnston, N. Ward, N. Graham, and D. Unger are gratefully acknowledged. G. Morgan of the Numerical Algorithms Group provided valuable assistance. The assistance of W. Landman, E. Klopper, and A. Bartman of the South African Weather Bureau; of L. Unganai of the Zimbabwe Meteorological Service; and of S. Crimp of the Queensland Department of Natural Resources, in experimenting with and commenting on some of the models when applied to seasonal precipitation forecasting has contributed to improvements in the model designs.

APPENDIX A

Discriminant Analysis

Let the population vector of means of the \( m \) independent variables, \( \mathbf{x} \) (the sea surface temperature principal component scores), for group \( j \) [where \( j \in \{\text{“La Niña,” “cool,” “normal,” “warm,” or “El Niño”}\} \) be \( \mu_j \), and the variance covariance matrix be \( \Sigma_j \). The posterior probability that a case belongs to group \( j \), given \( \mathbf{x} \), can then be obtained from Bayesian theory (Aitchison et al. 1977):

\[
p(j | \mathbf{x}, \theta) \propto p(j) \times p(\mathbf{x} | j, \theta),
\]

where \( \theta \) represents the set of unknown parameters \( \mu_1, \mu_2, \ldots, \mu_m, \Sigma_1, \Sigma_2, \ldots, \Sigma_m \), and \( p(j) \) is the prior probability of group \( j \). If it is assumed that the independent variables are multivariate normal, and that the parameters \( \theta \) can be estimated from a sample, \( \mathbf{z} \), so that \( \hat{\theta} = \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m, \bar{S}_1, \bar{S}_2, \ldots, \bar{S}_m \), where \( \bar{\mathbf{x}}_j \) represents the sample mean, and \( \bar{S}_j \) is the sample variance-covariance matrix of group \( j \), then

\[
p(\mathbf{x} | j, \mathbf{z}) = N_m(\mathbf{x}, \bar{S}_j) = \left[ 2\pi |\bar{S}_j|^{1/2} \exp \left( -\frac{d_j^2}{2} \right) \right]^{-1},
\]

where \( d_j^2 = (\mathbf{x} - \mu_j)' \bar{S}_j^{-1} (\mathbf{x} - \mu_j) \).

The posterior probability can then be calculated from Eqs. (A1) and (A2), substituting \( \bar{x}_j \) for \( \mu_j \), and \( \bar{S}_j \) for \( \Sigma_j \) in Eq. (A3). Equations (A1)–(A3) represent the estimative approach to discriminant analysis, but a predictive approach, in which a posterior distribution for \( \theta \) is first obtained, has been recommended (Aitchison et al. 1977). Equation (A2) then becomes

\[
p(\mathbf{x} | j, \mathbf{z}) = St_m \left[ \nu_j, \bar{x}_j, \left( 1 + \frac{1}{n_j} \right) \bar{S}_j \right],
\]

where \( St_m \) represents an \( m \)-dimensional Student-type density function, with \( \nu_j \) degrees of freedom, and where \( n_j \) is the number of cases in each category. For the case of unequal variance-covariance matrices, \( \nu_j = n_j - 1, \)
and classification functions become quadratic. If the matrices can be assumed equal, \( v_j = n - n_g \), and \( \mathbf{S}_j \) is replaced by the pooled variance–covariance matrix, and the classification functions are linear.

**APPENDIX B**

**Canonical Variate Analysis**

Canonical variates are linear combinations of variables defined to maximize differences between the means of each group. Given a set of \( n \) observations on \( m \) variables, \( \mathbf{X} \) (\( \mathbf{X} \) is \( n \times m \)), which are classified into \( n_g \) groups, differences between the groups can be described by considering the differences in the \( m \)-dimensional means. New variables \( \mathbf{Z} \) can be defined that are linear combinations of \( \mathbf{X} \),

\[
\mathbf{Z} = \mathbf{A}'\mathbf{X},
\]

which maximize the differences between the groups. Specifically, the weights, \( \mathbf{A} \), known as canonical variate loadings, are defined so that the new variables maximize the between-group variance.

The weights \( \mathbf{A} \) can be obtained in a manner similar to that for the canonical loadings in canonical correlation analysis (for further details see Gittins 1985; Tatsuoka 1988; Rencher 1995). Assume a set of indicator variables \( \mathbf{Y} \) that define the group membership of each observation. If the \( i \)th observation is a member of group \( j \) then let \( y_{ij} = 1 \), otherwise let \( y_{ij} = 0 \). By allowing observations that are members of the \( j \)th group to be 0 on all indicator variables, the group membership of all observations can be indicated by \( n_g \) variables.

Thus \( \mathbf{Y} \) is an \( n \times (n_g - 1) \) matrix of 1s and 0s. If the columns of \( \mathbf{X} \) are centered, its variance–covariance matrix can be defined as \( \mathbf{S}_y \) (dimension \( m \times m \)), while the cross-products matrix of \( \mathbf{Y} \) (the columns of \( \mathbf{Y} \) are not centered) is defined as \( \mathbf{S}_x \) [dimension \( (n_g - 1) \times (n_g - 1) \)]. The covariance between \( \mathbf{X} \) and \( \mathbf{Y} \) is given by the \( m \times (n_g - 1) \) matrix \( \mathbf{S}_{xy} \).

Because \( \mathbf{Y} \) is a set of indicator variables, the elements of \( \mathbf{S}_{xy} \) are simply

\[
s_{ij} = \begin{cases} n/n & \text{for } i = j \\ 0 & \text{for } i \neq j, \end{cases}
\]

that is, \( \mathbf{Y} \) is a diagonal matrix with the diagonals indicating the prior probability of each group. The elements of \( \mathbf{S}_{yy} \) are

\[
s_{ij} = \frac{n}{n} \sum_{k \in \{j\}} x_{ik},
\]

that is, the sum of \( x \), in group \( j \) weighted by the prior probability.

Defining an additional set of weights \( \mathbf{B} \) that define linear combinations of the indicator variables, canonical variate analysis involves finding the matrix \( \mathbf{A} \), such that

\[
\begin{bmatrix} \mathbf{A}' & 0 \\ 0 & \mathbf{B}' \end{bmatrix} \begin{bmatrix} \mathbf{S}_x & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{S}_y \end{bmatrix} \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{\Gamma}' \\ 0 & \mathbf{I}' \end{bmatrix},
\]

(B1)

where \( \mathbf{I} \) is an identity matrix of order \( n \), and where \( \mathbf{\Gamma} \) is an \( m \times (n_g - 1) \) matrix such that

\[
\mathbf{\Gamma}' = \begin{bmatrix} \text{diag}(r_k) \\ 0 \end{bmatrix} \text{ for } k = 1, \ldots, n_g - 1,
\]

where \( r_k \) is a canonical correlation coefficient, defined in more detail below. The matrix \( \mathbf{A} \) can be obtained from Eq. (B1) as the set of solutions to the eigenproblem

\[
(\mathbf{S}_x^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}' - r_k^2\mathbf{I})\mathbf{a}_k = 0.
\]

(B2)

Equation (B2) is identical to the canonical correlation analysis problem in which \( \mathbf{Y} \) is the set of dependent variables. However, instead of imposing the constraint \( \mathbf{a}'\mathbf{S}_y\mathbf{a} = 1 \), used in canonical correlation analysis (von Storch and Zwiers 1999), the canonical variates are scaled so that the within-group variance is 1.

That Eq. (B2) maximizes the between-group variance can be demonstrated as follows. Premultiplying Eq. (B2) by \( \mathbf{S}_{xx}^{-1} \) gives

\[
(\mathbf{S}_{yy}'\mathbf{S}_{yy}^{-1} - r_k^2\mathbf{I})\mathbf{a}_k = 0.
\]

(B3)

Here \( \mathbf{S}_y \) represents the total variance and covariance of \( \mathbf{X} \), and \( \mathbf{S}_{yy}^{-1} \) represents the between-groups variance. The aim in canonical variate analysis can be stated as to maximize the ratio of between-group to total variance, \( f(\mathbf{a}) \),

\[
f(\mathbf{a}) = \frac{\mathbf{a}'\mathbf{S}_y\mathbf{S}_{yy}^{-1}\mathbf{S}'_y\mathbf{a}}{\mathbf{a}'\mathbf{S}_y\mathbf{a}},
\]

(B4)

where the between-group and total variances of \( z \) are given by the quadratic forms \( \mathbf{a}'\mathbf{S}_y\mathbf{S}_{yy}^{-1}\mathbf{S}'_y\mathbf{a} \) and \( \mathbf{a}'\mathbf{S}_y\mathbf{a} \), respectively. Equation (B4) can be maximized by solving

\[
\frac{\partial f(\mathbf{a})}{\partial \mathbf{a}} = 2[(\mathbf{S}_y^{-1}\mathbf{S}_{yy}^{-1}\mathbf{S}'_y\mathbf{a})(\mathbf{a}'\mathbf{S}_y\mathbf{a}) - (\mathbf{a}'\mathbf{S}_y\mathbf{S}_{yy}^{-1}\mathbf{S}'_y\mathbf{a})(\mathbf{S}_y\mathbf{a})]/(\mathbf{a}'\mathbf{S}_y\mathbf{a})^2 = 0,
\]

which simplifies to

\[
[\mathbf{S}_y^{-1}\mathbf{S}_{yy}^{-1}\mathbf{S}'_y - f(\mathbf{a})\mathbf{S}_y]\mathbf{a} = 0.
\]

(B5)

Setting \( f(\mathbf{a}) = r^2 \), Eq. (B5) is identical to (B3). Thus, the ratio of between-group to total variance is equivalent to the squared canonical correlation.

It can be demonstrated that canonical variate analysis is equivalent to descriptive discriminant analysis. For further details, the interested reader is referred to the texts cited above, where significance tests for the canonical correlations are described in addition.

In practice, the canonical variates usually are derived from QR decomposition or singular value decomposition (Mardia et al. 1979; Muirhead 1982), although if \( \mathbf{X} \) is prefiltered using principal components analysis,
becomes an identity matrix, and the solution to Eq. (B3) can be obtained reasonably simply.

**APPENDIX C**

**Generalized Linear Models**

The form of the classic linear regression model can be generalized to include a wide range of other model forms. Generalized linear regression (McCullagh and Nelder 1989) permits the fitting of a model that is linear in its parameters, but may have nonnormally distributed errors, and permits a function, \( g() \), known as a link function, to be defined to transform the values of the predictand:

\[
g(y) = \eta = \alpha + \beta'x. \tag{C1}
\]

By defining an appropriate link function and error distribution, these generalized linear models can be used, inter alia, to model probabilities of events. Regression estimation of event probabilities (Wilks 1995), which involves classic linear regression, usually is considered inadequate because the predictor, as a probability, is constrained between 0 and 1, and because the predictand does not follow a normal distribution. Assuming for the moment a simple situation in which there are only two categories (such as above- and below-median temperature), the predictand is defined as the outcome of a Bernoulli trial, in which, for example, \( y = 1 \) if temperature is above the median, and \( y = 0 \) otherwise. The objective is to fit a regression model defining how the probability of temperature being above-median changes contingent upon the independent variables \( x \). A transformation function \( g() \) is defined to map the unit interval of the probability onto the unbounded range of the linear equation. Three link functions are commonly used [setting \( p = p(y = 1 | x) \) for simplicity]: logit [the logit model is the inverse of the logistic function (Hosmer and Lemeshow 1989)]

\[
\ln\left(\frac{p}{1-p}\right) = \alpha + \beta'x; \tag{C2}
\]

probit

\[
\Phi^{-1}(p) = \alpha + \beta'x, \tag{C3}
\]

where \( \Phi \) represents the normal distribution function; and complementary log–log

\[
\ln[-\ln(1-p)] = \alpha + \beta'x. \tag{C4}
\]

All three link functions define monotonic relationships between \( \eta \) and \( p \). The logit and probit models define s-shaped curves symmetric about \( \eta = 0 \) (\( p = 0.5 \)).

When there are more than two categories and the categories can be ordered, as is the case in this paper, it becomes inappropriate to calculate the value of \( p \) in Eqs. (C2)–(C4) for each category separately because it would normally be assumed that there are monotonic relationships between the independent and the dependent variables. Instead, the \( n_e \) categories can be ordered, and Eq. (C1) is expressed in terms of cumulative probabilities, as follows:

\[
g[p(y > j|x)] = \eta = \alpha_j + \beta'x
\]

\[
j = 1, \ldots, n_e - 1, \tag{C5}
\]

where \( y \in \{1, \ldots, n_e\} \). For example, sorting the five temperature categories in ascending order, \( y = 1 \) if the precipitation is in the La Niña category, \( y = 2 \) if the precipitation is in the cool category, \ldots, and \( y = 5 \) if the precipitation is in the El Niño category. In the case of the logit link, Eq. (C5) becomes

\[
\ln\left[\frac{p(y > j|x)}{p(y \leq j|x)}\right] = \alpha_j + \beta'x
\]

\[
j = 1, \ldots, n_e - 1, \tag{C6}
\]

for the probit link, Eq. (C5) becomes

\[
\Phi^{-1}[p(y > j|x)] = \alpha_j + \beta'x,
\]

\[
j = 1, \ldots, n_e - 1, \tag{C7}
\]

and for the complementary log–log link, Eq. (C5) becomes

\[
\ln[-\ln(1-p(y > j|x))] = \alpha_j + \beta'x \quad j = 1, \ldots, n_e - 1. \tag{C8}
\]

The \( n_e - 1 \) equations defined in (C5) are regression planes. Substitution of any \( x \) vector of independent variables in Eq. (C6), for example, yields \( n_e - 1 \) values that estimate the probability ratios defined in the left-hand side of the equation. From the fact that the probabilities of the three outcomes must sum to 1, the required probabilities \( p(\text{La Niña}), p(\text{cool}), p(\text{normal}), p(\text{warm}), \text{and } p(\text{El Niño}) \), can be obtained. In this paper, the resultant models are called ordinal logit [Eq. (C6)], ordinal probit [Eq. (C7)], and complementary log–log models [Eq. (C8)].

A limitation of Eq. (C5) is that it is possible for the regression planes to cross, in which case a negative probability for an intermediate category would be obtained. As an alternative, Eq. (C5) often is simplified to

\[
g[p(y > j|x)] = \eta = \alpha_j + \beta'x
\]

\[
j = 1, \ldots, n_e - 1, \tag{C9}
\]

where \( \alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_{n_e-1} \). For different values of \( j \), the differences in the right-hand sides of Eq. (C9) are solely in the constant \( \alpha_j \) which implies that, for the logit link, the log–odds ratios are modeled as parallel planes. While constraining the planes to be parallel does guarantee that the regression planes do not cross, it is restrictive since it assumes the same rate of change in the log odds for different levels of \( j \). Equation (C9) using a logit link is known as the proportional odds model (Hosmer and Lemeshow 1989; McCullagh and Nelder 1989). Corresponding forms for the probit (proportional
APPENDIX D

Proof That LEPSPROB is Not a Strictly Proper Scoring System

A strictly proper score is defined as one that encourages the forecaster to make his/her probabilities equal to his/her true beliefs (Murphy and Epstein 1967). Given \( n_\ell \) exhaustive and mutually exclusive categories, let the forecaster’s true beliefs concerning the probabilities of the possible outcomes be defined as \( p_j \) for \( j = 1, \ldots, n_\ell \), where \( p_j \geq 0 \), and

\[
\sum_{j=1}^{n_\ell} p_j = 1. \tag{D1}
\]

Let the probabilities of the possible outcomes that are actually issued by the forecaster be defined as \( r_j \) for \( j = 1, \ldots, n_\ell \), where \( r_j \geq 0 \), and

\[
\sum_{j=1}^{n_\ell} r_j = 1. \tag{D2}
\]

For a strictly proper scoring system, the forecaster will be encouraged to set \( r_j = p_j \) for all \( j = 1, \ldots, n_\ell \).

The LEPSPROB score is defined as

\[
\sum_{j=1}^{n_\ell} c_{ij} r_j, \tag{D3}
\]

where \( c_{ij} \) is the LEPS coefficient for category \( j \) given that the verification was for category \( k \). These coefficients are dependent only upon the number of categories, and on the prior probabilities of each category (Ward and Folland 1991; Potts et al. 1996). The forecaster’s expected LEPSPROB score, \( E(\text{LS}) \), is maximized when

\[
\frac{\partial E(\text{LS})}{\partial r_j} = \sum_{k=1}^{n_\ell} p_k \sum_{j=1}^{n_\ell} c_{ik} c_{jk}^{-1} \frac{\partial c_{jk}^{-1}}{\partial r_j} = 0, \tag{D4}
\]

for any \( i \), and subject to the constraints in Eqs. (D1) and (D2). Equation (D4) is independent of \( r \), and so the appropriate hedge is to set \( r_j = 1 \) for the category for which \( \sum_{i=1}^{n_\ell} p_i c_{ij} c_{ik}^{-1} \) is largest and \( r_j = 0 \) for the others. Since \( c_{ik} \) for the extreme categories is greater than for the intermediate categories, \( \sum_{i=1}^{n_\ell} p_i c_{ik} c_{ik}^{-1} \) may be larger for an extreme than for an intermediate category even if \( p_i \) is greater for the intermediate category. The forecaster can therefore hedge by being bullish about the possibility of an extreme.

As an example, let the forecaster’s true beliefs about the probabilities for an \( n_\ell = 5 \) category system be 0.35, 0.40, 0.15, 0.07, and 0.05. Assuming equal priors, the LEPS coefficients are as defined in Table D1 (Potts et al. 1996). The LEPSPROB scores would be 0.430, 0.644, −0.029, −0.631, and −0.429, for a verification in category \( k = 1, \ldots, 5 \), respectively. Appropriately the score is highest if the verification is category 2, which was the category with the highest posterior probability (but note that a better score is achieved if the verification is category 5 than if it is category 4). The expected score is 0.347.

If all probability is assigned to category 2, the LEPSPROB scores would be 0.406, 1.000, 0.125, −0.786, and −0.531, for a verification in category \( k = 1, \ldots, 5 \), respectively. The expected score is 0.407, which is slightly greater than the expected score of 0.347 given that \( r_j = p_j \) for all \( j = 1, \ldots, n_\ell \). The forecaster is thus able to achieve a higher score by assigning a probability of 1.0 to the category considered most likely, rather than forecasting his/her true beliefs.

However, the forecaster can achieve an even larger score by assigning a probability of 1.0 to category 1. In this case, the LEPSPROB scores would be 0.929, −0.625, −1.214, and −0.719, for a verification in category \( k = 1, \ldots, 5 \), respectively. The expected score is 0.521, which is greater than the expected score of 0.490 given that a probability of 1.0 is assigned to category 2. The forecaster is thus able to achieve the highest score by assigning a probability of 1.0 to the extreme category.

## REFERENCES


Clarke, W. R., P. Lachenbruch, and B. Brof®tt, 1979: How nonnor-
Chen, D., S. E. Zebiak, A. J. Busalacchi, and M. A. Cane, 1995: An
Harvey, L. O., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992:
Hastenrath, S., and L. Greischar, 1993a: Further work on the prediction
Harrison, M. S. J., 1995: Long-range seasonal forecasting since 1980:
Epstein, E. S., 1969a: Stochastic dynamic prediction. 
Krause, V., 1969: A scoring system for probability forecasts of ranked categories. 
Harrison, M. S. J., 1995: Long-range seasonal forecasting since 1980: Empirical and numerical prediction out to one month for the United Kingdom. 
Hastenrath, S., and L. Greischar, 1993a: Further work on the prediction of northeast Brazil rainfall anomalies. 
——, 1993b: Changing predictability of Indian monsoon rainfall anomalies. 
——, J. van Heerden, 1995: Prediction of the summer rainfall over South Africa. 
Ji, M., A. Leetmaa, and V. E. Kousky, 1996: Coupled model predictions of ENSO during the 1980s and the 1990s at the National Centers for Environmental Prediction. 
Jones, R. H., 1968: A nonlinear model for estimating probabilities of k events. 
——, ———, and ———, 2000: Multi-model ensemble forecasts for weather and seasonal climate. 
Lachenbruch, P. A., 1968: On expected probabilities of misclassi®cation in discriminant analysis, necessary sample size, and a relation with the multiple correlation coef®cient. 
Leitn, C. E., 1974: Theoretical skill of Monte Carlo forecasts. 
Marks, S., and O. J. Dunn, 1974: Discriminant functions when covariance matrices are unequal. 
——, J. van Heerden, 1995: Prediction of the summer rainfall over South Africa. 
——, and ———, 1995: Prediction of the summer rainfall over South Africa. 
——, J. van Heerden, 1995: Prediction of the summer rainfall over South Africa. 
——, and ———, 1995: Prediction of the summer rainfall over South Africa. 
——, J. van Heerden, 1995: Prediction of the summer rainfall over South Africa. 
——, and ———, 1995: Prediction of the summer rainfall over South Africa. 
——, J. van Heerden, 1995: Prediction of the summer rainfall over South Africa. 
——, J. van Heerden, 1995: Prediction of the summer rainfall over South Africa. 
——, J. van Heerden, 1995: Prediction of the summer rainfall over South Africa. 
——, J. van Heerden, 1995: Prediction of the summer rainfall over South Africa. 


