Comparison of Some Statistical Methods of Probabilistic Forecasting of ENSO

S. J. Mason, Scripps Institution of Oceanography, smason@ucsd.edu
G. M. Mimmack, University of the Witwatersrand, mimmackg@ucfv.bc.ca

1. Introduction

In this paper, some statistical methods of producing probabilistic forecasts of monthly Niño-3.4 sea-surface temperature anomalies are tested. A carefully constructed retroactive forecast procedure was designed to estimate as closely as possible the operational skill of the model predictions. The performances of the models are compared to 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.

2. Data and methods

a. Predictors

Forecasts of monthly Niño-3.4 sea-surface temperature (SST) anomalies were produced at lead times of between 0 and 11 months. The forecasts were made using the first five principal component scores of antecedent monthly mean SSTs over the domain 25°N–25°S, 110°E–70°W. Data for the 50-yr period January 1951 to December 2000 were obtained from the Kaplan *et al.* (1998) dataset. Principal components were calculated from the correlation matrix of monthly anomaly data for the first 30 yr (1951–80), and for each month separately. The Niño-3.4 anomalies were grouped into five equiprobable categories over the training periods, defined as "La Niña", "cool", "normal", "warm", and "El Niño" conditions.

b. Models

Forecasts of Niño-3.4 anomalies were made using predictive discriminant analysis, canonical variate analysis, and various forms of generalized linear models. The performances of these probabilistic methods of forecasting were compared to forecasts from a multiple linear regression model. Probabilistic forecasts were obtained from the regression model in two ways: from the intersections of the prediction interval and the category boundaries; 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, here on "multiple regression" is used to refer to the first method, and "contingency table" to the second.) Full details of all the models used are provided by Mason and Mimmack (2001, *J. Climate*, in press).

The optimal combination of predictors for the discriminant analysis, generalized linear regression, multiple regression, and contingency table models was identified using a procedure based on the "maximum-posterior-probability/leave-one-out" method of variable selection (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. The goodness of fit was measured by calculating the ranked probability score over the training period. 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.

c. Model validation

Model performance was assessed using a retroactive forecast procedure so as to obtain realistic estimates of operational prediction skill. The training period was initially set as 30 years (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 to produce a set of 20 yr of retroactive predictions.

The models were validated using a variety of skill scores calculated with reference to strategies of random guessing and climatological probabilities. In addition, the performances of the various models 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.

3. Results

Retroactive ranked probability skill scores (RPSSs) were calculated for forecasts at separate lead times, and are shown in Fig. 1, where they are compared with scores for damped persistence and persistence forecasts. The RPSSs are calculated with reference to a strategy of forecasting the climatological probabilities of each of the five categories. At all lead-times beyond zero the scores for all the models, except for the contingency table, exceed the scores for the persistence forecasts. The inability to out-score persistence at the shortest lead-time is a ubiquitous problem with dynamical and statistical models (Goddard *et al.* 2001). While the skill scores for the persistence forecasts decrease to zero after only about 3 months, for most of the model forecasts skill is positive out to about 9- or 10-months lead-time. Skill scores for forecasts of persistence damped toward climatology are shown by the light gray bars in Figs. 1. 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 skilful forecasts at the longer lead times. 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.

Reliability diagrams were constructed for each of the five categories to compare differences in the reliability of the forecasts from the models (Fig. 2). Diagrams for damped persistence are shown in place of those 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 towards over-confidence 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 and minimal unconditional bias for all five categories. For the other models, 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. The poor reliability for some of the extremely high forecast probabilities may be partly a sampling problem since there are few occurrences of very high probabilities.

Brier skill scores were used to identify the dependence of forecast skill upon the outcome. The greater skill at predicting El Niño conditions compared to any of the other categories is clearly evident (Fig. 3). 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 partly because of an unconditional bias (forecast probabilities were consistently too high), but also because of poor forecast resolution as indicated by the reliability curves (Fig. 2). 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 over-forecasted the occurrence of negative anomalies.

4. 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. Forecast skill was demonstrated for all the models at lead times of between about 4 or 5 months and 10 months. 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.

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 sub-optimal. 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 different sensitivities to sampling errors (the degrees of freedom are greatest for the multiple regression model). These differences can be decreased by reducing the number of categories from five to three, which weakens the differences in model skill notably (results not shown).

References

Goddard, L., S. J. Mason, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane, 2001: Current approaches to seasonal to interannual climate predictions. *Int. J. Climatol.*, **21**, 1111–1152.

Huberty, C. J., 1994: *Applied Discriminant Analysis*. Wiley, 466 pp.

Kaplan, A., M. A. Cane, Y. Kushnir, A. Clement, M. B. Blumenthal, and B. Rajagopalan, 1998: Analyses of global sea surface temperature 1856–1991. *J. Geophys. Res.*, **103**, 18 567–18 589.

Pan, J., and H. van den Dool, 1998: Extended-range probability forecasts based on dynamical model output. *Wea. Forecasting*, **13**, 983–996.

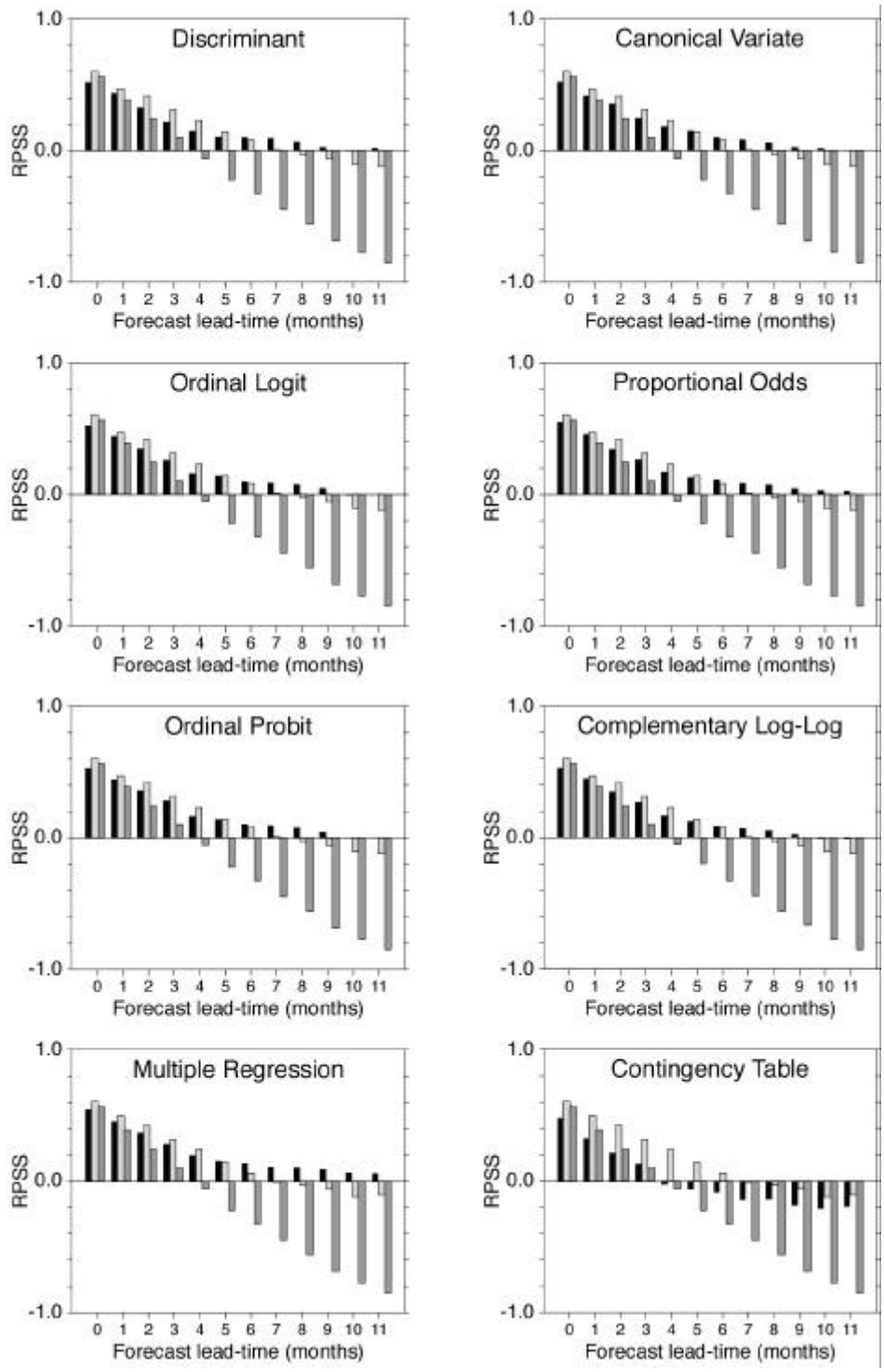

Fig. 1. Ranked probability skill scores for retroactive forecasts at increasing lead times of monthly Niño-3.4 sea surface temperature anomaly categories for Jan 1981–Dec 2000. The skill scores are calculated with reference to a strategy of forecasting climatology. The black bars represent the scores for the models, and the dark (light) gray bars are for forecasts of persisted anomaly categories (damped toward climatology).

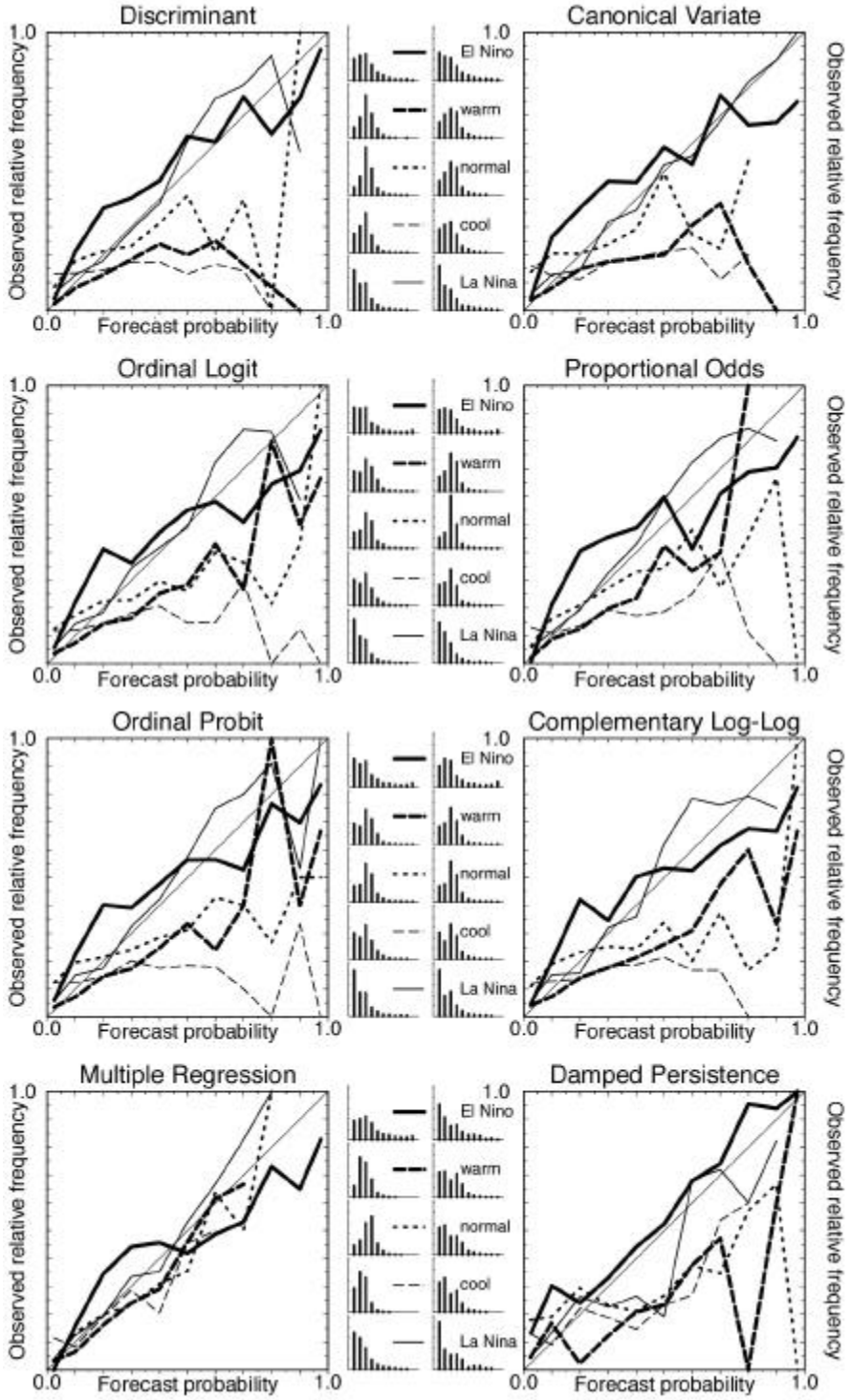

Fig. 2. Reliability diagrams for retroactive forecasts at increasing lead times of La Niña (solid thin line), cool (dashed thin line), normal (dotted line), warm (dashed thick line), and El Niño (solid thick line) conditions for the 20-yr period Jan 1981–Dec 2000. Forecasts at all lead times and for 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.

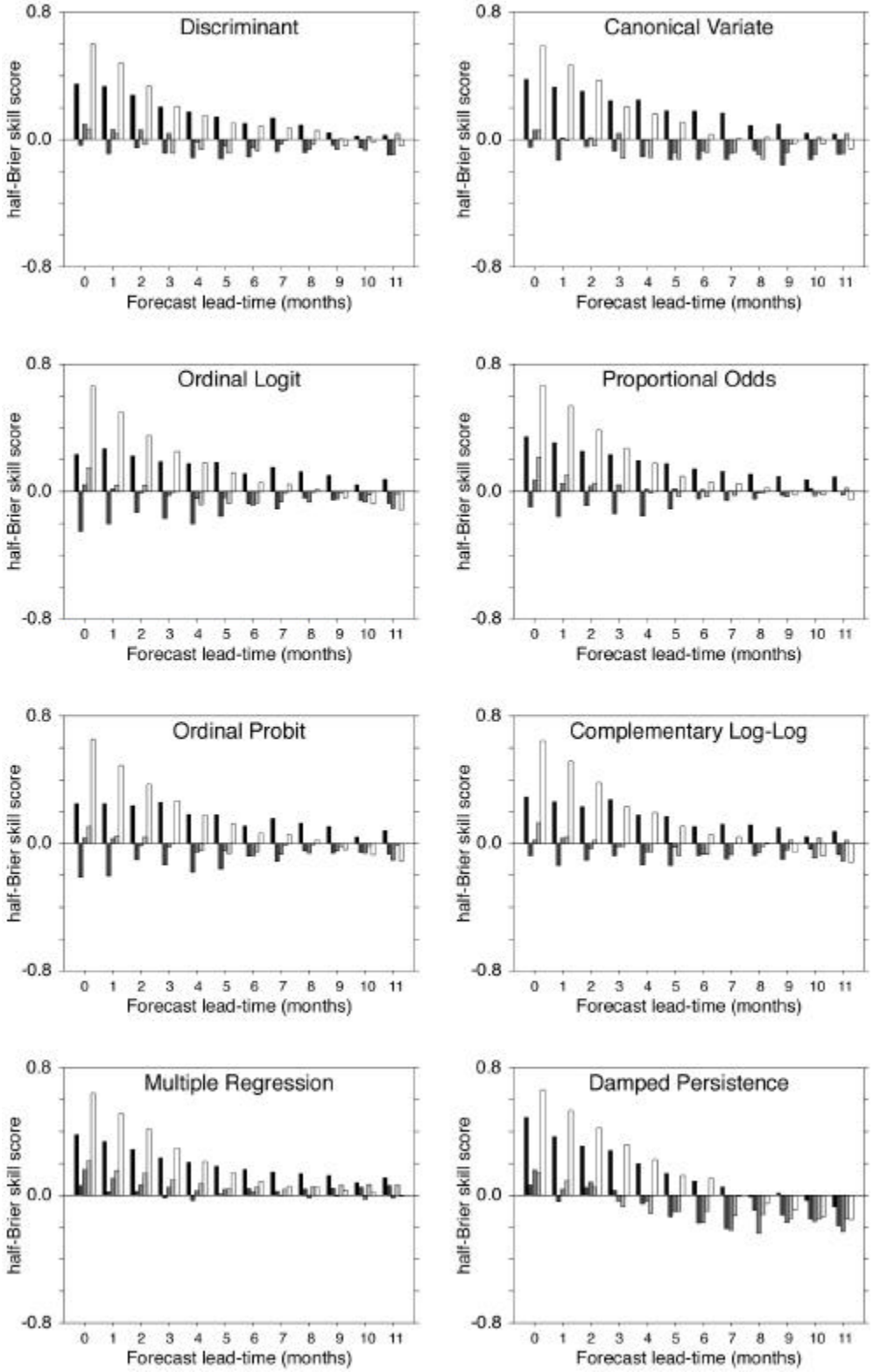

Fig. 3. Brier skill scores for retroactive forecasts at increasing lead times of La Niña (black bars), cool (dark gray bars), normal (gray bars), warm (light gray bars), and El Niño (white bars) conditions for the 20-yr period Jan 1981–Dec 2000.