Current Status of ENSO Forecast Skill

B. P. Kirtman and J. Shukla

Center for Ocean-Land-Atmosphere Studies

M. Balmaseda, N. Graham, C. Penland, Y. Xue and S. Zebiak

A Report to the Climate Variability and Predictability (CLIVAR) Numerical Experimentation Group (NEG)

October 2000

Abstract

The purpose of this comparison project is to assess the current state-of-the-art in predicting tropical Pacific sea surface temperature anomalies (SSTA). In order to make this assessment, retrospective forecasts of NINO3 (150^{o}-90^{o}W, 5^{o}S-5^{o}N) SSTA made by various research groups have been compared. Six dynamical (of various degrees of sophistication) prediction systems and one statistical prediction systems are considered here. The retrospective forecasts have been compared in terms of their correlation and root mean square error with respect to observations. Hit rate and false alarm rates are also compared. Remarkably, a forecast developed as a consensus of at least three separate prediction systems is arguably more skillful than any of the individual prediction systems. Comparisons have also been made to determine how well the models forecast the various phases of ENSO. Both the dynamical and statistical models produce useful forecast of the peak phase of the extreme warm and cold events up to two seasons in advance. However, none of the models adequately capture the detailed life cycle of the ENSO events nor are the models particularly good at predicting the timing of the onset of El Niño events. The period of retrospective forecasting is too short to adequately distinguish among the various models in terms of the correlation coefficient and the root mean square error.

A number of different unsuccessful attempts to measure the uncertainty in the forecast are described. These techniques try to relate the consistency of forecasts initialized one month apart to the error of the forecast. This approach fails because the forecasts initialized one month apart can be very consistent, but also can have large errors. Better ensembling techniques need to be developed to

accurately measure the uncertainty in the forecasts.

1. Introduction

An understanding of seasonal to interannual climate variability and its consequences in some regions of the globe has been one of the major accomplishments of the Tropical Ocean-Global Atmosphere (TOGA) program and has constituted one of the most significant achievements of climate research. This breakthrough results from the recognition that the Earth's climate is a coupled system involving the global atmosphere, the world oceans, and land surface processes, and, most importantly, the interactions among these components of the climate system. The most well known of these interactions is El Niño and the Southern Oscillation (ENSO), and it offers the greatest potential for seasonal to interannual climate predictions for societal benefit.

The TOGA program encouraged the development of several dynamical and statistical ENSO prediction systems. The forecast skill of these prediction systems in retrospective forecasting during the 1980s was described in Barnston et al. (1994). While Barnston et al. (1994) acknowledged that comprehensive dynamical models have the greatest potential for producing ENSO forecasts of societal benefit, they found that the statistically based forecasts performed about as well as the dynamically based forecasts. More recently, Barnston et al. (1999) and Landsea and Knaff (2000) looked at how well the 1997-98 ENSO was predicted. The problem with these two later studies is that they focused on one event, and it is unreasonable to expect the performance in one event to be indicative of the overall forecast skill.

Following the TOGA program has been the Climate Variability and Predictability (CLIVAR) program. One of the main objectives of CLIVAR is to extend the range and accuracy of seasonal to interannual climate prediction through the development of global coupled predictive models (WMO, 1995). In other words, one of the goals of CLIVAR is to realize the potential of comprehensive global coupled models in predicting seasonal to interannual climate variations. This is the same potential that Barnston et al. (1994) noted, but they argued had not been achieved. The main focus of this paper is to describe the current status of the ENSO prediction models based on a relatively large number of cases. Objective tests to distinguish the behavior of the various models is presented.

In order to begin to address the above objective, The CLIVAR-Numerical Experimentation Group (NEG) has recommended that an ENSO prediction comparison study be made. In this comparison project, various statistical and dynamical hindcasts of the NINO3 index (area averaged SSTA 150^o-90^oW, 5^oS-5^oN) produced by research groups and operational centers have been collected. In this report, these hindcasts are compared to each other and to the observational data in a uniform manner. Most of these groups continue to issue their forecasts in real time through the *Experimental Long Lead Forecast Bulletin* (http://www.iges.org/ellfb), and their performance during the 1997-98 El Niño and the 1998 La Niña onset is described in Barnston et al. (1999) and Landsea and Knaff (2000).

The basic findings of this comparison project are:

(i) both statistical and dynamical models produce useful tropical SSTA forecasts for the peak

phase of ENSO up to two seasons in advance.

(ii) A consensus forecast (i.e. an ensemble across prediction systems) is remarkably skillful, whereas an ensemble of realizations of a single prediction system improves the skill only marginally.

(iii) The periods of retrospective forecasting are too short in terms of distinguishing between the skill scores of the various prediction systems.

(iv) Models predict the sign of extreme events well, but too often predict warm or cold events when the observations call for normal conditions.

(v) Consistency among forecasts initialized one month apart is not a good *a priori* measure of forecast skill.

The remainder of this report is outlined as follows. Section 2 briefly describes the retrospective hindcasts contributed to this report and the observational data used for evaluation the performance of the models. The skill of the hindcast is described in Section 3 in terms of correlation coefficient, root mean square error, hit rates and false alarm rates . Attempts at using these hindcasts to develop a forecast of the forecast skill are discussed in Section 4. Some concluding remarks are made in Section 5. Throughout the remainder of this report, "hindcast", "forecast" and "prediction" are used synonymously. Strictly speaking, all the results presented apply only to retrospective forecasting.

2. The Prediction Systems

There are a number of different ENSO forecasting strategies including the purely statistical techniques, a combination of dynamic and statistical models, and purely dynamic models. Within these three categories, the models have varying degrees of sophistication and various initialization strategies. This section first describes the statistical prediction systems that are used in the comparison study. The dynamic models are then briefly described in order of increasing complexity, and finally the observational data is discussed. Table 1 summarizes all the prediction systems used in this study.

a. The Statistical Models

One statistical model was submitted for this comparison project. The Linear Inverse Modeling (LIM) prediction system (Penland and Magorian, 1993). The LIM prediction is made by applying a statistically obtained Green's function to an observed initial SSTA. While similar to principal oscillation pattern (POP) analysis (Xu and Von Storch, 1990), a relatively large number of oscillation patterns (modes) are used rather than just one or two. In the LIM prediction system, 20 EOFs of the SSTA are used and a three-month running mean has been applied to the SSTA. For this comparison there are 323 forecasts, each of 18 months duration initialized each month of each year from February 1965 to November 1991. In training the model, a "jackknife" procedure is used so that the entire record is separated into five 5-year blocks. The data outside the 5-year blocks are used to train the model. It should be noted that the LIM prediction system is not specifically trained to predict NINO3. It is, however, designed to give the best pattern of Indo-Pacific SSTA.

b. The Dynamic Models

There are six dynamic models used in this comparison study. Two forecast systems were submitted by the Lamont-Doherty Earth Observatory. These prediction systems are referred to as LDEO1 and LDEO2, respectively. The LDEO1 system (Cane et al., 1986; Zebiak and Cane, 1987) was the first dynamic model to routinely predict ENSO-related SST fluctuations in the eastern tropical Pacific. It covers the tropical Pacific region, and predicts monthly anomalies using linear shallow water equations for both the atmosphere and ocean. However, the model uses more complicated nonlinear formulations for the atmospheric heating and the ocean mixed layer thermodynamics. The model is initialized using wind stress anomalies derived from FSU pseudo-stress. The LDEO2 system is identical to LDEO1 except for the initialization procedure (Chen et al., 1995). In LDEO2, a coupled initialization is employed in which observed wind stress anomalies (FSU) are assimilated into the coupled model. The assimilation procedure has latitude dependence so that very near the equator the winds are primarily determined by the coupled model and off the equator the wind stress anomalies are mostly computed from the observations.

The University of Oxford (UOX) coupled model consists of a statistical atmosphere coupled to a two active layer ocean model (Balmaseda et al., 1994). The ocean model is first forced by observed wind stress (FSU) during 1961-91. The output of this ocean-only simulation is then used to build the statistical atmosphere, which assumes that the wind stress anomalies are a linear function of the first six EOFs of the model SST and heat content anomalies. The NINO3 forecasts are then post-processed using the initial observed SSTA and the model predicted tendency so that at the initial time there is no error.

The Scripps Institution of Oceanography prediction system (SIO) consists of a comprehensive dynamic ocean model coupled to a statistical atmosphere (Barnett et al., 1993). The ocean model (Latif, 1987), developed at the Max Planck Institut fur Meteorologie, is a primitive equations model for the tropical Pacific Ocean. It has 13 levels in the vertical, 10 being within the top 300 m. The statistical atmospheric model uses a canonical correlation/regression analysis-like procedure to predict wind stress anomalies from SSTA. The coupled model is initialized by first forcing the statistical atmosphere with observed SSTA, and then using the resulting wind stress to force the ocean model. With this procedure, the observed SSTA is implicitly used to initialize the coupled model.

The Center for Ocean-Land-Atmosphere Studies (COLA) anomaly coupled model (Kirtman et al., 1997) includes a global atmospheric general circulation model (AGCM; DeWitt, 1996) coupled to a

Pacific basin version of the Geophysical Fluid Dynamics Laboratory (GFDL) ocean model (Pacanowski et al., 1993). The AGCM is a global spectral model triangularly truncated at total wave number 30 with 18 unevenly spaced levels in the vertical. The ocean simulates only the tropical Pacific basin with 1.5^{o} longitude by 0.5^{o} latitude resolution in the deep tropics and 20 levels in the vertical. An empirical procedure using the winds at the top of the boundary layer is used to define the surface stress (Huang and Shukla, 1997). The ocean is initialized using an iterative procedure designed to reduce simulated SSTA errors in the eastern Pacific (Kirtman and Schneider, 1996). The component models are anomaly coupled in the sense that only the predicted anomalies of wind stress and SST are exchanged at the air-sea interface.

The National Centers for Environmental Prediction (NCEP) coupled model also includes a global AGCM coupled to the GFDL ocean model (Ji et al., 1994a, 1994b, 1996). The AGCM is a modified version of the NCEP medium range forecast (MRF) model. The AGCM is a global spectral model triangularly truncated at total wave number 42 with 18 vertical levels. The AGCM convective parameterization was tuned to produce a more realistic tropical wind stress simulation and there is an empirical correction applied to the model wind stress at the air-sea interface. The ocean thermal field is initialized with an ocean data assimilation system (Ji et al., 1995). Similar to the COLA prediction system, the NCEP prediction system employs the anomaly coupling procedure, but only applied to the wind stress exchanged at the air-sea interface.

c. The Observed Data

The predictions collected for this project are only for NINO3 area averaged SSTA. Similarly, these prediction are compared to "observed" SSTA in the NINO3 region. There are a number of different "observed" data sets available that were examined for verification. However, the differences in the "observed" SSTA are small compared to the forecast errors and have little impact on the results presented here. For the period of January 1963-October 1981, the global sea ice and sea surface temperature (GISST2.2a; Parker et al., 1995) is used, and for the period of November 1981-1995, the so-called OI-SST is used (Reynolds, 1988).

3. Prediction

This section compares the forecast skill of the various NINO3 predictions. First a consensus forecast was defined, and all the forecasts were compared in a uniform format. Traditional measures of skill such as correlation coefficient and root mean square error (rmse) were also considered. These traditional measures were examined as a function of season. Error estimates of these skill scores are provided and the skill scores are compared for identical cases. Hit rate versus false alarm rate plots are also provided, which indicate how well the models forecast the various phases of ENSO.

a. Forecast Evolution

One of the conclusions of this comparison study is that a consensus forecast (i.e., an ensemble across prediction systems) provides a remarkably skillful prediction. This is in contrast to an ensemble forecast based on multiple realizations with an individual prediction system which seems to improve the forecast skill of a particular forecast only marginally. The consensus forecast was made by taking the average across all the forecast systems. We required at least three different model predictions at a particular lead time to form a consensus forecast. No attempt was made to "optimize" the consensus in terms of weighting the forecasts.

Figure 1a shows the zero-lead forecast (i.e. initial condition) for all the prediction systems, the consensus forecast and the observations. The symbols correspond to the individual forecasts plotted each January, April, July and October. The solid curve corresponds to the observations and the dashed curve is the consensus initial conditions. The NCEP zero-lead forecast is not shown because it is the same as what was defined as the observations and is not used in calculating the consensus. The UOX and LIM directly use their own observational data to define their initial condition, and it is expected that their zero-lead forecast should agree quite well with what has been used to define the observations in this study. The UOX, LIM, COLA, LDEO1, LDEO2, and SIO prediction systems are all used to calculate the consensus initial condition when the forecasts are available.

For the most part, all the individual zero-lead forecasts agree fairly well with the observations and, as expected, the LIM and UOX correspond to the observations. However, there are periods of time where the COLA, LDEO1, LDEO2 and SIO initial conditions have relatively large errors. For example, during 1968-69, the SIO initial conditions were considerable cooler than the observations. The COLA, zero-lead forecasts under estimate the amplitude of the 1982-83 warm episode. The LDEO1 initial conditions have the poorest correspondence with the observations and are particularly poor during the relatively cold periods of 1970-71, 1973-74, 1984-85 and 1988-89. The LDEO1 forecasts and, to a lesser degree, the LDEO2 forecasts tend to over estimate the initial condition during warm events.

Figures 1b, 1c, 1d and 1e show the 3, 6, 9 and 12-month lead forecasts for all the prediction systems, respectively. These forecasts correspond to the same initial conditions shown in Fig. 1a. At short lead times (3 months; Fig. 1b), the forecasts generally capture the major warm and cold episodes. There are also periods when there is a relatively large spread among the various predictions (e.g., 1984-85 cold period); nevertheless, the consensus forecast performs quite well. For longer lead times (6-9 months; Fig. 1c and 1d), there is a substantial increase in the disagreement among the forecasts, and it appears that the consensus forecast is superior to any individual prediction system. At long lead times (12 months; Fig. 1e), it is difficult to detect any useful information in the individual forecasts. The consensus forecast captures some aspects of the strong warm and cold events.

Figures 1a-e focus attention on how well the details of the observed SSTA are captured by the various prediction systems, providing a somewhat pessimistic view of the current state-of-the-art in ENSO forecasting. The results look somewhat encouraging for large ENSO events only. Figure 2 shows two-season lead ensemble predictions and observations for the major warm and cold events only. The horizontal axis corresponds to the December-January-February (DJF) mean of the indicated years. The vertical axis indicates the ensemble mean prediction for forecasts initialized during the preceding June through August. Thus, Fig. 2 shows the predicted DJF mean for forecasts with lead times of 5 to 9 months. While the amplitude of the anomaly is not particularly well predicted, the sign of the anomaly is captured in almost all the cases. There is also some indication that the models perform better during warm events.

b. Skill Scores

Figure 3 shows the correlation coefficient and the rmse of all the prediction systems and for the consensus forecast, respectively. These skill scores are calculated over all available forecasts for each prediction system. The number of forecasts in each case is noted on the figure and in Table 1. There are 336 cases that make up the consensus forecasts. It should be noted that these skill scores are not evaluated over the same time intervals or the same number of cases. This issue is discussed in greater detail below.

For lead times up to 12 months, the correlation coefficient for NCEP, COLA, SIO, LDEO2, UOX and LIM cluster fairly close together. The LDEO1 correlation coefficient appears to be an outlier; however, a test of statistical significance reveals little difference at lead times of six months. For short lead times (0-3 months) the correlation coefficient for almost all the models is greater than 0.8 explaining approximately 65% of the SSTA variability. After about three months, the correlation coefficient decays rapidly. The LDEO1, LDEO2 and consensus forecasts have the slowest decay, whereas the statistical model has the strongest decay. Beyond twelve months, there is a fairly large spread in the correlation coefficient with the LDEO2 and consensus forecasts having the largest values and the LDEO1 having the smallest values. Considering all lead times, the consensus forecast arguably has the largest correlation coefficient.

Similar to the correlation coefficient, the rmse of the different forecasts systems cluster close together. The notable exceptions are the LDEO1 and to a lesser degree the LDEO2 prediction systems. For the short lead time, the rmse for most of the models is about 0.4^{o}C which is about half the observed NINO3 SSTA standard deviation. As the lead time increases, the rmse growth with most of the models appears to saturate at about the observed standard deviation. The LDEO1 system has a relatively large rmse which saturates considerably higher than the observed standard deviation. The UOX model has relatively small initial rmse (to be expected given the statistical corrections made) which grows rapidly to saturate at a relatively large value. Also, the consensus forecast, which had the largest correlation coefficient, has the smallest rmse.

As mentioned above, the problem with Fig. 3 is that skill scores are calculated over different periods and include different sample sizes. Table 2 and table 3 address this problem in different ways. First, in Table 2, we show the correlation coefficient along with the 99% confidence interval at lead times of 6, 9 and 12 months, respectively. The confidence intervals are calculated using Fisher's z-transformation and the error function (or incomplete gamma function)(1). See Press et al. (1992) for details. The confidence interval removes some of the ambiguity associated with the different sample sizes. The confidence intervals also demonstrate that there is no statistical difference among the correlation coefficients at 6, 9 and 12 months. For example, at a lead time of six months, the LDEO1 system has the smallest correlation (0.591), but adding the confidence interval gives a correlation of 0.701. The NCEP model, on the other hand, has the highest correlation (0.752) and subtracting the confidence interval gives a correlation of 0.662, indicating that it is not statistically different from the LDEO1 correlation at the 99% confidence level.

While Table 2 resolves some of the ambiguity associated with the sample size, the different time periods may also affect the skill scores (e.g., Kirtman and Schopf, 1998). Table 3 shows the 6, 9 and 12-month correlation coefficient and 99% confidence interval calculated for only a subset of 96 cases. These 96 cases corresponding to a forecast initialized each month of each year 1982, 1983, 1984, 1986, 1987, 1988, 1989, 1991. For each forecast system, the correlation coefficient and confidence interval is computed over the same number of cases. Examination of Table 3 also indicates that there is no statistical difference among the correlation coefficients at the 99% confidence level.

The seasonality of forecast skill has been the subject of active research. Figure 4 (page 1 and page 2) clearly shows that all the models have similar seasonality in correlation coefficient. The top-left to bottom-right tilt in the 0.6 correlation coefficient isopleth indicates that the skill of the forecasts initialized in the later part of the year (September through December) drops below 0.6 earlier than forecasts initialized during boreal spring. On the other hand, the forecasts initialized in the latter part of the year tend to have larger correlation coefficients at short lead times (0-3 months). There are, of course, exceptions. For example, the NCEP model appears to be most skillful for forecasts initialized during May to June. Several of the forecasts also appear to have a "return of skill." The LIM forecasts, for example, initialized during August-September lose skill (i.e. correlation drops below 0.5) during the following boreal spring and regain skill during the boreal summer. This return of skill can also be detected to a lesser degree in all the other prediction systems. This return of skill appears to be smallest for the LDEO2 system, and, remarkably, the consensus forecasts has the strongest return of skill signature.

c. Hit Rate versus False Alarm Rate

Fig. 2 indicates that there is reason to believe that the models can accurately predict the phase of ENSO. With the following hit rate versus false alarm rate plots, this issue is explored in more detail. An additional question to be addressed is how often do the prediction models forecast erroneous warm or cold events. For instance, what percentage of the time do the various models predict a significant warming when the observations indicate near normal conditions? Alternatively, when the observations indicate significant warming, what percentage of the time do the models forecast near normal conditions? In order to address this type of question, hit rate and false alarm rates for the various prediction systems were calculated.

The procedure for calculating hit rates and false alarm rates is described in Stanski et al. (1989) and is briefly summarized here. First, the forecast and the observations are normalized by their respective standard deviations. Next contingency tables for each forecast system are constructed. For example, the NCEP contingency table for a lead time of six months is

		Forecast	**Category**		
Observations	Warm	Normal	Cold	Total	Hit Rate
Warm	43	8	0	51	0.84

Normal	27	32	10	69	0.46
Cold	4	28	31	63	0.49
Total	74	68	41		
False Alarm Rate	0.23	0.32	0.33		

The hit rate for warm events is the total number of correctly forecasted warm events (43) divided by the number of times it was observed warm (51), which gives a hit rate of 0.84 for the NCEP prediction system. Similarly, the false alarm rate for warm events is the percent of the forecast warm events given the a warm event did not occur (31/132), which gives a false alarm rate of 0.23. The hit rate and false alarm rates for the normal and cold categories are computed in a similar fashion. The warm, normal and cold categories are defined so that they are equally probable.

The hit rate and the false alarm rate should be considered together. This is because the hit rate can be easily increased by "crying wolf," but at the expense of more false alarms. The false alarm rate can be decreased by under forecasting the number of extreme events, but this will also reduce the hit rate. The perfect prediction system would have a hit rate of 1.0 and a false alarm rate of 0.0, and, as in the example above, when the hit rate is larger than the false alarm rate, we refer to the predictions as being skillful.

The hit rate versus the false alarm rate for each prediction system (including CONS) for all three terciles is shown in Figs. 5a-c, 6a-c and 7a-c for lead times of 3, 6 and 9 months, respectively. Uniformly, all the models do well at forecasting warm events 3 months (Fig. 5a) in advance - the hit rates are significantly larger than the false alarm rates. The coupled initialization in LDEO2 reduces the false alarm rate over the LDEO1 prediction system. For this lead time, the LIM result is closest to the upper left corner, although the UOX and CONS forecasts score quite well. The NCEP system has the best hit rate, but at the expense of having a false alarm rate that is in the middle of the pack.

In terms of forecasting the normal tercile (Fig. 5b), the models lie closer to the diagonal and are, therefore, less skillful. Overall there is a slight increase in the false alarm rate compared to the warm tercile, but most of the loss of skill is due to reduced hit rates. The hit rates and the false alarm rates for the cold tercile (Fig. 5c) are somewhat lower than the warm tercile, and, with this metric, almost all the models are skillful at predicting cold events with three months lead.

As expected, with increased lead time there is a general decrease in the hit rates and an increase in the false alarm rates. However, overall most of the models are skillful at predicting the warm and cold events at lead times of 6 (Fig. 6a,c) and 9 (Figs. 7a,c) months. In contrast, the models are only marginally skillful at predicting normal conditions for these longer leads.

4. Prediction of Prediction Skill

One of the key outstanding questions in ENSO forecasting is how to estimate the reliability of the forecast. In order to address this question, two approaches were taken. First, an ensemble forecasting technique was adopted. The members of the ensemble were defined to be forecasts initialized one month apart. Hopefully, taking the ensemble mean reduces the amplitude of the internal variability and calculating the ensemble spread may provide an estimate of the uncertainty of the forecast. The second approach was to an EOF analysis of the forecast evolution to define signal to noise ratios and to relate these ratios to the forecast error. Unfortunately, both of these techniques fail to give a useable *a priori* estimate of the uncertainty of the forecast. Nevertheless, this failure does clearly demonstrate that using the spread of forecasts initialized one month apart is not adequate to determine the uncertainty in the forecast.

Figure 8 (page 1 and page 2) shows the scatter diagrams of the ensemble mean forecast error and the ensemble spread at a lead time of six months for each prediction system. The ensemble size is three, although a number of different ensemble sizes have been tried with no qualitative change in the results. These scatter diagrams indicate that there is little or no association between the accuracy of the ensemble mean forecast and the ensemble spread when the ensemble consists of forecasts initialized one month apart. These data have been examined in a number of different ways including the absolute value of the ensemble mean forecast error as well as stratifying the data by the phases of ENSO and by month of the year. In all such calculations, there is little or no association between the ensemble spread and the ensemble mean forecast error.

An indication of why there is little association between forecast spread and the forecast error can be seen in Fig. 9 which shows an example from two of the prediction systems (COLA and LDEO1); however, there are similar examples to varying degrees with all the other prediction systems. Figure 9 shows the evolution of the observed NINO3 SSTA and the forecasts for the COLA system initialized during January-May 1975 and the LDEO1 system initialized during January-May 1983. In both cases the forecasts are quite consistent, but also quite wrong.

Finally, an attempt was made to estimate a signal to noise ratio and relate this ratio to the forecast error. To define the signal, the forecast evolution was filtered using an EOF analysis, where the two leading EOFs were considered to be the signal. The noise was defined to be the mean squared amplitude of the remaining EOFs. We then considered a number of different ways of calculating the forecast error, but we were unable to find an association between the error and the signal to noise ratio.

5. Summary and Concluding Remarks

The current state-of-the-art in ENSO prediction was evaluated based on retrospective forecasts with six dynamical and two statistical prediction systems. A number of typical forecast "skill" measures were calculated. Based on the correlation coefficient and the rmse, it is concluded that all the models produce useful forecasts for lead times up to 6 months. Forecast for the peak phase of ENSO two seasons in advance accurately capture the sign of the event. However, there is a large degree of spread among the models at six-month lead times indicating that many of the details of the SSTA still cannot be accurately predicted. An interesting aspect of the skill comparison was the fact that a consensus forecast was arguably more skillful than any of the particular prediction systems. Based on these skill comparisons, it is not possible to determine which prediction systems are more skillful in a statistically significant way.

Hit rates and false alarm rates were calculated to assess how well the models predict the various phases of ENSO. For all lead times (up to 9 months), the models are skillful (i.e., the hit rate exceeds the false alarm rate) at predicting warm events. The models uniformly have difficulty predicting near normal conditions. Cold events are less skillfully predicted than warm events, nevertheless almost all the models produce skillful forecasts at lead time of 9 months using this particular measure.

One of the goals of CLIVAR is to accelerate the development of sophisticated dynamic models for climate prediction and the intent of this work was to assess the status of ENSO prediction using dynamic models. The statistical models provide a useful benchmark for evaluating the dynamical models. At the end of TOGA, it was hoped that the dynamical models would have improved considerably. Based on this comparison project, it cannot be concluded that the dynamical models are clearly superior to the statistical models. In fact, it is not possible to distinguish between the statistical models and the dynamic models in terms of their skill scores.

Finally, a number of calculations were made in an attempt to develop a forecast of the forecast skill, but with no success. While part of the problem comes from the limited amount of data, it is clear that the spread of forecasts initialized one month apart does not provide a good measure of forecast uncertainty. The problem may not be with the idea of using ensembles to determine forecast uncertainty, but rather with the actual techniques used to develop the ensemble members. More research into ensembling techniques as well as signal to noise ratios is required. Retrospective forecasting for a much larger number of past cases is required to test the model performance and to develop techniques to quantify the reliability of the forecasts.

Based on these results, it is concluded that accurately predicting the strength and timing of ENSO events continues to be a critical challenge for both dynamical and statistical models of all levels of complexity. Improved models, data and initialization strategies are required to address the problem of predicting tropical eastern Pacific SSTA. Prediction of regional precipitation and circulation will not be possible without accurate predictions of SSTA.

References:

Balmaseda, M. A., Anderson, D. L. T., and M. K. Davey, 1994: ENSO prediction using a dynamical ocean model coupled to a statistical atmospheres. *Tellus*, **46A**, 497-511.

Barnett, T. P., M. Latif, N. Graham, M. Flugel, S. Pazan and W. White, 1993: ENSO and ENSO related predictability: Part I, prediction of equatorial Pacific sea surface temperatures with a hybrid coupled ocean-atmosphere model. *J. Climate*, **6,** 1545-1566.

Barnston, A. G., M. H. Glantz, and Y. He, 1999: Predictive skill of statistical and dynamic climate models in forecasts of SST during the 1997-98 El Niño episode and the 1998 La Niña onset. *Bull. Amer. Meteor. Soc.*, **80,** 217-243.

Barnston, A. G., H. M. van den Dool, S. E. Zebiak, T. P. Barnett, M. Ji, D. R.. Rodenhuis, M. A. Cane, A. Leetmaa, N. E. Graham, C. R. Ropelewski, V. E. Kousky, E. A. O'Lenic, and R. E. Livezey, 1994: Long-lead seasonal forecasts- where do we stand? *Bull. Amer. Meteor. Soc.*, 75, 2097-2114.

Cane, M. A., S. E. Zebiak, and S. C. Dolan, 1986: Experimental forecasts of El Niño. *Nature*, **321**, 827-832.

Chen, D., S. E. Zebiak, A. J. Busalacchi and M. A. Cane, 1995: An improved procedure for El Niño forecasting: Implications for predictability. *Science*, **269**, 1699-1702.

DeWitt, D. G., 1996: The effect of cumulus parameterization on the climate of an atmospheric general circulation model: Annual mean and interannual variability. COLA Tech. Rep. 27, 49pp.

Goldenberg, S. D., and J. J. O'Brien, 1981: Time space variability of tropical Pacific wind stress. *Mon. Wea. Rev.*, **109**, 1190-1207.

Huang, B., and J. Shukla, 1997: An examination of the AGCm simulated surface wind stress and low level winds over the tropical Pacific Ocean. *Mon. Wea. Rev.*, **125**, 985-998.

Ji, M., A. Kumar and A. Leetmaa, 1994a: A multi-season climate forecast system at the National Meteorological Center. *Bull. Amer. Meteor. Soc.*, **75,** 569-577.

Ji, M., A. Kumar and A. Leetmaa, 1994b: An experimental coupled forecast system at the National Meteorological Center: Some early results. *Tellus*, **46A**, 398-418.

Ji, M., A. Leetmaa and J. Derber, 1995: An ocean analysis system for seasonal to interannual climate studies. *Mon. Wea. Rev.*, **123**, 460-481.

Ji, M., A. Leetmaa, V. E. Kousky, 1996: Coupled model predictions of ENSO during the 1980s and 1990s at the National Centers for Environmental Prediction. *J. Climate*, **9,** 3105-3120.

Kirtman, B. P., and E. K. Schneider, 1996: Model based estimates of equatorial Pacific wind stress. *J. Climate,* **9**, 1077-1091.

Kirtman, B. P., J. Shukla, B. Huang, Z. Zhu and E. K. Schneider, 1997: Multiseasonal predictions with a coupled tropical ocean global atmosphere system. *Mon. Wea. Rev.*, **125**, 789-808.

Kirtman, B. P., and P. S. Schopf, 1998: Decadal variability in ENSO predictability and prediction. *J. Climate*, **11**, 2804-2822.

Landsea, C. W., and J. A. Knaff, 2000: How much skill was there in forecasting the very strong 1997-98 El Niño? *Bul. Amer. Meteor. Soc.,* **81**, 2107-2119.

Latif, M., 1987: Tropical ocean circulation experiments. *J. Phys. Oceanogr.*, **17**, 246-263.

Pacanowski, R. C., K. Dixon and A. Rosati, 1993: The GFDL modular ocean model users guide, version 1.0. GFDL Ocean Group Tech. Rep. No. 2, 77pp. [Available from GFDL/NOAA, Princeton University, Princeton NJ 08542.]

Parker, D. E., M. Jackson, and E. B. Horton, 1995: The GISST sea surface temperature and sea ice climatology. Hadley Center for Climate Research Tech. Note 63, 49 pp. [Available from the Hadley Centre Meteorological Office, London Road, Bracknell, Berkshire RG12 2SY, United Kingdom]

Penland, C., and T. Magorian, 1993: Prediction of Niño-3 sea surface temperatures using linear inverse modeling. *J. Climate*, **6**, 1067-1076.

Press, W. H., S. A. Teukolsky, W. E. Vetterling and B. P. Flannery, 1992: *Numerical Recipes in FORTRAN: The art of scientific computing, second edition*. Cambridge University Press, 963 pp.

Reynolds, R. W., 1988: A real time global sea surface temperature analysis. *J. Climate*,**1**, 75-86.

Stanski, H. R., L. J. Wilson, W. R. Burrows, 1989: Survey of common verification methods in meteorology. World Weather Watch Tech. Report #8. (WMO/TD, #358), WMO, Geneva, 144 pp.

WMO, 1995: CLIVAR, a study of climate variability and predictability: science plan. WCRP-89 (WMO/TD No. 690). WMO., Geneva.

Xu, J. S., and H. von Storch, 1990: Predicting the stat of the southern oscillation using principal oscillation pattern analysis. *J. Climate*, **3**, 1316-1329.

Zebiak, S. E., and M. A. Cane, 1987: A model of El Niño and the southern oscillation. *Mon. Wea. Rev.*, **115**, 2262-2278.

Figure Captions:

Figure 1(a). Time series of observed (solid cyan curve) NINO3 SSTA, the consensus (dashed black curve) forecast and the forecasts (symbols) from the various prediction systems. The forecasts are for a lead time of zero months (i.e., the forecast initial condition).

Figure 1(b). Time series of observed (solid cyan curve) NINO3 SSTA, the consensus (dashed black curve) forecast and the forecasts (symbols) from the various prediction systems. The forecasts are for a lead time of three months.

Figure 1(c). Time series of observed (solid cyan curve) NINO3 SSTA, the consensus (dashed black curve) forecast and the forecasts (symbols) from the various prediction systems. The forecasts are for a lead time of six months.

Figure 1(d). Time series of observed (solid cyan curve) NINO3 SSTA, the consensus (dashed black curve) forecast and the forecasts (symbols) from the various prediction systems. The forecasts are for a lead time of nine months.

Figure 1(e). Time series of observed (solid cyan curve) NINO3 SSTA, the consensus (dashed black curve) forecast and the forecasts (symbols) from the various prediction systems. The forecasts are for a lead time of twelve months.

Figure 2. Two season lead forecasts of extreme warm and cold events. The forecasts are initialized during June-August and are verifying the following December-February so that the forecasts are for lead times of 5 to 9 months. The specific years of the warm and cold events are noted along the x-axis for the verification time.

Figure 3. Skill scores (anomaly correlation and root mean square error) for all the prediction systems. The dashed black curve denotes the skill of consensus and the other colors correspond to the various prediction systems noted in left margin. The number of cases over which the correlation coefficient and root mean square error is calculated is also noted in the left margin.

Figure 4. (Page 1 and page 2) Correlation coefficient of the various prediction systems as a function of initial month (y-axis) and lead time (x-axis). The contour interval is 0.1 and the shaded values are greater than 0.5. The same cases used in Fig. 3 are used in Fig. 4.

Figure 5. Hit rate versus false alarm rate (see text for details) at a lead time of 3 months. (a) Warm events, (b) near normal conditions and (c) cold events.

Figure 6. Hit rate versus false alarm rate (see text for details) at a lead time of 6 months. (a) Warm events, (b) near normal conditions and (c) cold events.

Figure 7. Hit rate versus false alarm rate (see text for details) at a lead time of 9 months. (a) Warm events, (b) near normal conditions and (c) cold events.

Figure 8. (Page 1 and page 2) Scatter plots of the ensemble mean error versus ensemble spread. The ensemble size is three and is made up of forecasts initialized one month apart.

Figure 9. Time series of the COLA forecasts initialized in January-May 1975 and the LDEO1 forecasts initialized in January-May 1983. The forecasts are given with the dashed curves and the observation are given with the solid curves. Lead time zero corresponds to January 1975 in the case of COLA and January 1983 in the case of LDEO1.

1. The number of degrees of freedom (dof) used is the same as the number of forecasts cases which probably gives an over-estimate of the dof and gives smaller confidence intervals.