Long-Lead Seasonal Forecasts—Where Do We Stand?

Anthony G. Barnston,* Huug M. van den Dool,* Stephen E. Zebiak,+ Tim P. Barnett,# Ming Ji,@ David R. Rodenhuis,* Mark A. Cane,+ Ants Leetmaa,@ Nicholas E. Graham,# Chester R. Ropelewski,* Vernon E. Kousky,* Edward A. O'Lenic,* and Robert E. Livezey*

Abstract

The National Weather Service intends to begin routinely issuing long-lead forecasts of 3-month mean U. S. temperature and precipitation by the beginning of 1995. The ability to produce useful forecasts for certain seasons and regions at projection times of up to 1 yr is attributed to advances in data observing and processing, computer capability, and physical understanding—particularly, for tropical ocean–atmosphere phenomena. Because much of the skill of the forecasts comes from anomalies of tropical SST related to ENSO, we highlight here long-lead forecasts of the tropical Pacific SST itself, which have higher skill than the U.S forecasts that are made largely on their basis.

The performance of five ENSO prediction systems is examined: Two are dynamical [the Cane–Zebiak simple coupled model of Lamont-Doherty Earth Observatory and the nonsimple coupled model of the National Centersfor Environmental Prediction (NCEP)]; one is a hybrid coupled model (the Scripps Institution for Oceanography–Max Planck Institute for Meteorology system with a full ocean general circulation model and a statistical atmosphere); and two are statistical (canonical correlation analysis and constructed analogs, used at the Climate Prediction Center of NCEP). With increasing physical understanding, dynamically based forecasts have the potential to become more skillful than purely statistical ones. Currently, however, the two approaches deliver roughly equally skillful forecasts, and the simplest model performs about as well as the more comprehensive models. At a lead time of 6 months (defined here as the time between the end of the latest observed period and the beginning of the predictand period), the SST forecasts have an overall correlation skill in the 0.60s for 1982–93, which easily outperforms persistence and is regarded as useful. Skill for extratropical surface climate is this high only in limited regions for certain seasons. Both types of forecasts are not much better than local higher-order autoregressive controls. However, continual progress is being made in understanding relations among global oceanic and atmospheric climate-scale anomaly fields.

It is important that more real-time forecasts be made before we rush to judgement. Performance in the real-time setting is the ultimate test of the utility of a long-lead forecast. The National Weather Service's plan to implement new operational long-lead seasonal forecast products demonstrates its effectiveness in identifying and transferring "cutting edge" technologies from theory to applications. This could not have been accomplished without close ties with, and the active cooperation of, the academic and research communities.

*Climate Prediction Center NCEP/NWS/NOAA
+Lamont-Doherty Earth Observatory, Columbia University
#Scripps Institution of Oceanography, University of California
@Coupled Model Project NCEP/NWS/NOAA
Corresponding author address: Anthony Barnston, Climate Prediction Center, W/NP51, World Weather Building, Room 604, 5200 Auth Rd., Camp Springs, MD 20746.
E-mail: wd51ab@sgi45.wwb.noaa.gov
In final form 5 October 1994.

1. Introduction

The Climate Prediction Center (CPC) [formerly the Climate Analysis Center (CAC)] of the National Centers for Environmental Prediction (NCEP) of the National Weather Service is preparing to issue operationally long-lead forecasts of 3-month mean temperature and precipitation for the United States. Experimental forecasts of this nature have been produced and issued since late 1992. The forecasts will be made for periods beginning a half month after the forecast time, progressing by monthly increments to one year into the future. This ambitious plan follows years of forecasting seasonal anomalies with zero lead, in which the period being forecast began at the time of the forecast. (A zero-lead forecast for a 3-month period has often been called a one-season lead forecast, in contrast to the stricter definition used here.) In the last half decade, however, advancement along several scientific fronts has opened doors to more sophisticated forecast possibilities.

Zero-lead forecasts for monthly or seasonal means of U.S. surface climate have been made routinely by several groups for more than a decade. Examples are those issued by the CPC or its predecessor (Namias 1964; Gilman 1985; Wagner 1989), Scripps Institution of Oceanography, and others. The CPC forecasts have been based in part on specific empirical tools such as the North Pacific SST effects described by Davis (1978), the climate state vector analog system (Barnett and Preisendorfer 1978; Livezey and Barnston 1988), and the effects of the El Niño/Southern Oscillation (ENSO) (Ropelewski and Halpert 1986).

Beginning in the 1980s, a few attempts to make forecasts with greater-than-zero lead time for extratropical surface climate have been made. All have been experimental and usually partly subjective, such as the U.S. temperature and precipitation forecasts of A. Douglas (see Preisendorfer and Mobley 1984) and U.S. temperature forecasts of Wagner (Wagner and Livezey 1984). An objective experimental long-lead forecast system using canonical correlation analysis (CCA), based on the work of Barnett and Preisendorfer (1987), was further developed and implemented in 1991 (Barnston 1994). Forecasts of U.S. surface temperature using CCA have been used as one of several inputs for CPC's zero-lead operational seasonal forecasts and have appeared in the CPC's *Experimental Long-Lead Forecast Bulletin* since fall 1992. Dynamically based forecast models have also been developed and are currently being used for long-lead extratropical prediction as the second of a two-stage process, the first being the prediction of tropical sea surface temperature (SST) anomalies. Examples are the hybrid and two-tiered coupled model projects, joint efforts of the Scripps Institution of Oceanography and the Max Planck Institute for Meteorology (Graham and Barnett 1994; Barnett et al. 1994; Bengtsson et al. 1993), and the full coupled model at the NCEP (formerly the National Meteorological Center) (Ji et al. 1994a,b).

The recent ability to produce forecasts useful at projection times longer than previously thought possible is attributed to rapid advancement in data observing and assimilation systems, computer capability, and understanding of the importance of tropical boundary conditions for the evolution of the ENSO phenomenon. This last achievement has been associated with new insight into tropical ocean–atmosphere interactions and their representation in coupled models. While some extratropical processes probably develop independently of the Tropics (e.g., blocking in northern ocean basins), much of the skill of the forecasts for the extratropics comes from anomalies of ENSO-related tropical SST (e.g., Barnett et al. 1994; Graham and Barnett 1994; Barnston 1994). The tropical Pacific SST anomalies themselves have been targets of empirical as well as dynamical forecast systems. Success in ENSO forecasting is essential to extratropical forecasting in the Pacific–North American region, because the tropical SST anomalies are an important cause of midlatitude upper atmospheric and associated surface climate anomaly patterns (Horel and Wallace 1981; Barnett 1981; Ropelewski and Halpert 1986).

A dependence of extratropical climate on tropical phenomena is also found for Atlantic tropical storm activity (Shapiro 1982; Gray et al. 1993; Elsner and Schmertmann 1993). Tropical SST strongly affects tropical continental climate, such as seasonal rainfall in northeastern Brazil (Ward and Folland 1991; Hastenrath and Greischar 1993; Graham 1994). In view of the above, we highlight in this paper forecasts of the ENSO-related tropical Pacific SST as the primary example of long-lead forecasting, with secondary emphasis on the emerging technology for extratropical prediction.

The recent ability to produce forecasts useful at projection times longer than previously thought possible is attributed to rapid advancement in data observing and assimilation systems, computer capability, and understanding of the importance of tropical boundary conditions for the evolution of the ENSO phenomenon.

In its own right, the ENSO phenomenon has become the focal point of many worldwide concerns, because of its large effects on the climate and the economy in various regions of the world (Ropelewski and Halpert 1986, 1987). Diagnosis and prediction of ENSO are thus of considerable interest to the general population. From the second half of the 1980s (Barnett et al. 1988) through the present, progress in understanding and simulating the physics and dynamics of ENSO has accelerated. Dynamical approaches have included 1) simple ocean models coupled with statistical atmospheres (Inoue and O'Brien 1984; Graham et al. 1992), 2) ocean general circulation models (GCMs) coupled with statistical atmospheres (Neelin 1990; Latif and Flugel 1991; Barnett et al. 1993), 3) a simple coupled model for both ocean and atmosphere (Cane et al. 1986), 4) an ocean GCM coupled to a simple atmospheric model (Latif et al. 1993a,b), and 5) the most complex case of an ocean GCM coupled with an atmospheric GCM (e.g., Ji et al. 1994b).

Physical ocean models coupled with statistical atmospheric models (e.g., the hybrid coupled model for the Tropics discussed in Barnett et al. 1993) are reasonably effective because the atmospheric response to the tropical oceanic boundary conditions can be derived from historical data with moderate success. Simple coupled ocean–atmosphere models apply the laws of physics in both media, using carefully chosen, abbreviated versions of the full equations of atmospheric and oceanic motion and interaction. While

their simplicity may cause them to neglect factors critical to their forecasts, it also eliminates problems related to erroneously simulated details that occur in full GCMs.

Purely statistical ENSO prediction models have also been developed. These include CCA (Graham et al. 1987a,b; Barnston and Ropelewski 1992), principal oscillation patterns (POPs) (Xu and von Storch 1990) or the related inverse modeling (Penland and Magorian 1993), the singular spectrum analysis–maximum entropy method (Keppenne and Ghil 1992), and others. Statistical models, which are usually much less costly to run than dynamical models, serve the purpose of setting a baseline skill level to which the skills of dynamical models can be compared. The skill of the dynamical models must exceed this baseline in order to justify their additional effort and cost.

Physical models vary considerably not only in their basic assumptions and equations but also in their physical domain. The Lamont simple coupled model (Zebiak and Cane 1987) uses a tropical Pacific domain, which precludes tropical–extratropical interactions in the simulations. The Scripps Institution for Oceanography (hereafter Scripps)–Max Planck Institute for Meteorology (MPI) hybrid coupled model (Barnett et al. 1993), using a complex ocean model and a statistical atmospheric model, also covers the Pacific from 30°N to 30°S. The NCEP coupled model uses full GCMs in both ocean and atmosphere and covers the midlatitudes as well as the Tropics. In the cases of all three models the equatorial SST anomaly in the central and eastern Pacific is based on several factors, a common one of which is the amount of heat stored in the top 100–200 m of ocean in the western and central tropical Pacific.

While the estimate of predictive skill produced from most of the models in the above-described categories are roughly comparable, specific differences can be found. Such differences may occur in the skill's seasonality and geographical distribution as well as its decay with forecast lead time. It should also be noted that overall skill score similarity does not imply forecast similarity.

Several models have been subjected to the challenge of producing a succession of real-time forecasts. In this paper we examine the forecast skills of five currently or potentially operational ENSO forecasting systems: 1) the Lamont-Doherty Earth Observatory (hereafter Lamont) simple coupled model (Zebiak and Cane 1987), 2) the Scripps–MPI hybrid coupled model (Barnett et al. 1993), 3) NCEP's comprehensive coupled model (Ji et al. 1994a,b), 4) the CPC's statistical CCA model (Barnston and Ropelewski 1992), and 5) the CPC's empirically constructed analog model (Van den Dool 1994). The purpose of this presentation is to assess the effectiveness of current routine ENSO forecasting and to lay out prospects for future achievement not only for ENSO forecasts but for their ultimate application: forecasts of extratropical climate and of tropical climate in regions distant from the ENSO action areas.

In section 2 the five forecast models are briefly described. Section 3 examines the performance of the models, and section 4 announces consequent decisions of the National Weather Service regarding operational issuance of long-lead forecasts. A summarizing discussion and concluding remarks are given in section 5.

2. Major characteristics of five ENSO forecast models

In this section we briefly highlight the features of the two dynamical forecast models (Lamont and NCEP coupled models), the hybrid coupled model of Scripps–MPI, and two empirical forecast models of NCEP (CCA and constructed analog). We also note here that the SST data used for all the models discussed here, as well as for forecast verification, come from combinations of COADS (Coupled Ocean–Atmospheric Data Set) (Slutz et al. 1985) and NCEP (Reynolds 1980), the latter being used by all models for 1980 and later. Forecasts were made and verified for area average SST over discrete regions specified below. Observed area average SST may differ slightly for the same time and region from one model to another, because the periods over which climatological means are based are not identical.

a. The Lamont model

The simple coupled dynamical model developed at Lamont-Doherty Earth Observatory (Cane et al. 1986; Cane and Zebiak 1987; Zebiak and Cane 1987) is well known as the first physical model dedicated to routine diagnosis and prediction of ENSO fluctuations for the benefit of the scientific community and other users. It covers the tropical Pacific region only and predicts specified monthly departures from climatology (i.e., it is an anomaly model). It uses linear shallow water dynamics for both the ocean and atmosphere, but includes more complicated nonlinear forms for atmospheric heating and ocean mixed layer thermodynamics. The model is not initialized with analyzed SST data; only wind stress anomalies (derived from The Florida State University analyses) enter into the initialization. The model was constructed to simulate ENSO over a 12-month period (Zebiak 1984; Cane et al. 1986). In hindcast (i.e., retrospective forecast) mode, the model has simulated the variability in the tropical

Pacific SST starting in the early 1970s. Since fall 1985, the model has remained unchanged and has continuously produced forecasts in a completely independent setting. Much of its success is due to favorable reproduction of the heat storage mechanism in the subsurface western and central equatorial Pacific Ocean (Wyrtki 1985), attributable to ocean wave dynamics. For the period from 1970 to the early 1990s, statistically significant predictive skill is found up to 12–16 months lead (Cane 1992). Real-time forecasts for Niño 3 (bounded by 5°N–5°S, 90°–150°W) using the Lamont model have been issued in the CPC's operational *Climate Diagnostics Bulletin* since summer 1989, and for the entire tropical Pacific Basin in CPC's semioperational *Experimental Long-lead Forecast Bulletin* and the *Climate Diagnostics Bulletin* since fall 1993. Prior to these times, the Lamont model forecasts were distributed to interested worldwide users and were posted on the ENSO.INFO electronic bulletin board file on the Internet.

b. The Scripps–MPI hybrid coupled model

A hybrid coupled model of the tropical ocean–atmosphere system has been developed jointly at Scripps Institution of Oceanography and the Max Planck Institute for Meteorology (Barnett et al. 1993). The ocean model, created at MPI for the tropical strip (Latif 1987), is a fully nonlinear GCM bounded by 30°N and 30°S latitude and by Asia and South America. It has 13 vertical levels, 10 of which are within the top 300 m. The seasonal cycle is governed by a Newtonian heat flux and observed wind stress (Goldenberg and O'Brien 1981). The vertical mixing scheme is dependent upon the Richardson number (Pacanowski and Philander 1981). The atmospheric model is statistical, deriving the wind stress forcing for the ocean GCM using the GCM's SST. This is done with a CCA-like regression model, using historical observed data fields of anomalous SST and the corresponding wind stress. The ocean GCM provides the SST anomaly to the atmospheric model that, in turn, produces the wind stress anomaly that subsequently forces the ocean, producing an updated SST field. The coupling process includes a Model Output Statistics (MOS)-like statistical correction of the SST fields produced by the ocean GCM. The hybrid coupled model is initialized with wind stress fields derived from observed SST data; thus, it is indirectly "spun up" with SST information. Considering the entire 1965–93 period, the model has demonstrated statistically significant predictive skill for up to 12–18 months, with best performance for the central equatorial Pacific and for winter forecasts. The model was constructed using data from the 1965–85 period, leaving 1986–93 for independent forecast testing. Real-time forecasts by the Scripps–MPI model for one or more portions of the equatorial Pacific have appeared in the *Experimental Long-Lead Forecast Bulletin* since spring 1994. Since 1991 such forecasts have occasionally been issued on the ENSO.INFO file on the Internet.

c. The NCEP coupled model

A comprehensive coupled ocean–atmosphere GCM has been developed at the NCEP (formerly the National Meteorological Center) over the last several years (Ji et al. 1994a,b). The Pacific basin ocean model was created originally at the Geophysical Fluid Dynamics Laboratory (GFDL) by Bryan (1969) and Cox (1984) and subsequently improved by Philander (1987). It covers a domain of 45°S–55°N, 120°E–70°W. Its zonal resolution is 1.5°, and its meridional resolution is 0.33° within 10° of the equator. Between 10° and 20° away from the equator the meridional resolution decreases gradually to 1°. There are 28 vertical levels, most of which are concentrated in the upper ocean. The atmospheric model is a T40 version of the NCEP Medium Range Forecast (MRF) model with 18 vertical levels. The convective parameterizations of the MRF have been tuned for more realistic tropical air–sea interactions and convection. Exchange of surface momentum and heat fluxes and SST at the air–sea interface occur at 5-model day intervals, representing the timescale of ocean response to surface wind stress. The ocean thermal field, including SST and subsurface temperature, is initialized using an ocean data assimilation system (Ji et al. 1994c). The model was developed and tuned for 6-month lead forecasts using data from the cold and warm ENSO episodes of 1988/89 and 1991/92, respectively. Routine real-time forecasts of SST anomalies in the tropical Pacific basin began during 1993; these have appeared in the *Climate Diagnostics Bulletin* and the *Experimental Long-Lead Forecast Bulletin.*

d. The NCEP CCA model

CCA is a statistical technique that models linear relationships between fields of the predictors and the SST predictands, using a specific variation of EOF analysis (Barnett and Preisendorfer 1987; Barnston and Ropelewski 1992). In the version used at the CPC that was developed with some initial guidance from Scripps Institution of Oceanography, the predictor fields consist of global sea level pressure and the tropical Pacific SST itself for several periods prior to the target time; the target time SST is the predictand. Specifically, four consecutive 3-month predictor periods are followed by a lead time (a data "skip") and then a single 3-month predicted period. CCA essentially performs a multivariate linear regression, in which patterns in the predictand are related to preceding

patterns in the four time-staggered predictor fields. Systematic evolution of the predictor fields over time relating to a data-defined predictand pattern is thus identified. The relationships governing the optimal prediction are defined over the period of record and then applied to the future year being forecast. The CCA model predicts a set of regions spanning across the tropical Pacific and Indian Oceans. Real-time CCA forecasts for the SST in a particular Pacific region centered approximately 40% of the distance between Niño 3 and Niño 4 (i.e., 120°–170°W, to be called Niño 3.4) have appeared in the *Climate Diagnostics Bulletin* since early 1990 and in the *Experimental Long-Lead Forecast Bulletin* since its inception in fall 1992. In the late 1980s a few forecasts using an earlier version of CCA were posted by Scripps Institution of Oceanography in the file ENSO.INFO on the Internet.

e. The NCEP constructed analog model

Recently, Van den Dool (1994) developed a method of constructing a better analog than any that occurs naturally, using an optimal linear combination of the SST in all available years to model the base state to be matched more precisely. The predictor field (through which analog matches are sought) consists of near-global SST over four consecutive 3-month predictor periods, followed by a lead time (a "skipped" period), and then a single 3-month predicted period. Skill experiments have demonstrated that a single constructed analog leads to higher skill than classical composites of natural analogs/antianalogs (as used in Livezey and Barnston 1988; Barnston and Livezey 1989). Thus, the production of a better analog match appears to outweigh the loss of the nonlinearity that would be preserved in the climatic scenarios defined by natural analogs. Real-time forecasts for the SST in 120°–170°W (Niño 3.4) using constructed analogs have appeared in the *Experimental Long-Lead Forecast Bulletin* beginning in summer 1994.

3. The SST forecast skills

This section is not intended to compare the SST forecast skills among the five models but to answer the question, Where do we stand? in a collective sense. In doing this, skill will sometimes be examined on an individual model basis.

a. The absolute need for real-time forecasts

A purely objective, bias-free method of estimating the forecast skill of any given method in a truly independent (future) forecast setting does not exist. *There simply is no substitute for real-time forecasting.* For each of the five methods, we currently have only a small sample of years of unadulterated real-time forecasts. In view of the low-frequency nature of ENSO, we need at least 10 years before judgement can be passed on to forecasting ability. However, we cannot wait that long. Consequently, carefully derived estimates of skill based on hindcasts (in which the model "knows" data that occurred later than the time being forecast) or retroactive real-time forecasts (in which no data occurring later than the forecast time are made available to the model) are needed. We have assembled a dataset of the latter type for the 1982–93 period for the empirical CCA and constructed analog forecasts.

There are techniques that attempt to simulate independent forecasts for statistical forecast methods such as linear regression or CCA. For such methods, cross validation (Michaelsen 1987) is thought to give approximately representative results as it withholds each year in turn from the model's developmental sample and makes a forecast for it. However, skill inflation can still occur when there are interannual autocorrelations in the data history, as in an ENSO-related SST or sea level pressure time series. This problem can be largely overcome by withholding groups of consecutive years for each set of forecast trials. Skill deflation can also occur when true skill is low, due to a degeneracy inherent in cross validation (Barnston and Van den Dool 1993). The success of the model building exercise is also jeopardized in nonstationary regimes in which predictive rules identified over a relatively long period begin failing after a later point in time. Cross validation is applicable to analog methods, because the year for which analogs are sought is excluded as an analog candidate and the climatology in terms of which all years' data are expressed can be recomputed with that year withheld (Van den Dool 1987). Cross validation as described here is impractical in assessing independent period forecast skill for dynamical models, because iterative retuning with each year held out would be far too cumbersome a task. The usual practice is to develop the model using data for part of the available period and to test independent forecast skill on the remaining part.

b. Postprocessing of model output and independent nature of recent forecasts

The current version of Lamont model forecasts is based on the model development period of 1970–85 for postprocessing purposes—that is, for determination of systematic biases. Model biases have been removed for all forecasts, separately by forecast target season and lead time as determined from the development period. Forecasts have also been adjusted for differences in variance with respect to the observations in the same manner. Because forecasts

TABLE 1. Five ENSO forecast model characteristics and skill.

Authors	Zebiak and Cane (1987)	Barnett et al. (1993)	Ji et al. (1994b)
Model	Physical: simple coupled	Hybrid: coupled nonsimple ocean GCM, statistical atmosphere	Physical: coupled nonsimple GCMs
Details of model	Six-member ensemble	No ensembles	Four-member ensemble
Lead time	6.5 months	6 months	6 months
Time from forecast start to center of predicted period	8 months	7.5 months	7.5 months
Predicted SST region (All 5°N–5°S)	Niño 3 90°–150°W (eastern Pacific)	140°–180°W (central Pacific)	Niño 3.4 120°–170°W (east-central Pacific)
Period of record	1970–93	1966–93	1984–93
Proportion of evaluation period containing independent forecasts	8/12 = 0.67	8/12 = 0.67	6/10 = 0.60
Skill (1982–93)	Corr 0.62; rmse 0.95; Design[c] mixed; std dev=1.08	Corr 0.65[a]; rmse 0.97; Design mixed; std dev=1.10	Corr 0.69[b]; rmse 0.83[b]; Design mixed; std dev=1.00

[a]Standard correlation for Scripps–MPI model is 0.69 (see appendix).
[b]Skill for 1984–93.
[c]See footnote c in continuation of Table 1 on next page.

have been made using the same original version of the model since late 1985, the 1986–93 period can be regarded as fully independent and as real time.

The development period for the Scripps–MPI hybrid coupled model is 1965–85. Although there was no need for bias correction (in part because a MOS correction scheme was applied during coupling), a temporal phase postprocessor was applied to the forecasts as a function of their lead time. Forecasts for the 1986–93 period can be considered independent.

Skills for the NCEP model are based on the relatively short data record of mid-1982 to 1993, and the even shorter target period record of 1984–1993. Model biases are subtracted from the forecasts specific to the model starting month, based on hindcasts over the data record. Model development was carried out for the 1988–89 and 1991–92 periods, leaving 1984–87, 1990, and 1993 for independent forecasting. Thus, within the 1984–93 evaluation period the proportion of years available for independent forecasts is slightly lower than that for the Lamont and Scripps–MPI models during 1982–93.

For the statistical CCA model each year is withheld in turn and the model developed over the 1956–93 period but without any explicit influence from the withheld year that is the forecast target. Forecast anomalies, damped toward climatology to minimize mean squared error, are reinflated such that their variance equals that of the observations. For a second, perhaps more realistic, skill estimate of CCA, truly real-time forecasts are simulated by omitting all data occurring after the time of the forecast. Referred to as retroactive real-time forecasting, this is carried out for the 1982–93 period. It is about as close as one can come to a real test (as for the Lamont and Scripps–MPI models for 1986–93 and the NCEP model for its separated intervals of independent years), one difference being that in "real" real time there are hard to avoid problems such as unavailable or erroneous predictor data.

The constructed analog forecasts were produced in a cross-validation design in similar fashion to the CCA forecasts. The scheme holds out the year being forecast in the sense that it cannot be used to construct the analog. Additionally, the climatology, in terms of which all years' data are expressed, is repeatedly recalculated excluding the base year. Because the variance of the forecasts is realistic, inflation of forecast anomalies is not necessary. Forecasts were made both using

TABLE 1 (continued).

Authors	Barnston and Ropelewski (1992)	Van den Dool (1994)
Model	Statistical: CCA	Empirical: constructed analog
Details of model	Four consecutive 3-month predictor periods	Four consecutive 3-month predictor periods
Lead time	6 months	6 months
Time from forecast start to center of predicted period	7.5 months	7.5 months
Predicted SST region (All 5°N–5°S)	Niño 3.4 120°–170°W (east-central Pacific)	Niño 3.4 120°–170°W (east-central Pacific)
Period of record	1956–93	1956–93
Proportion of evaluation period containing independent forecasts	12/12 = 1.00	12/12 = 1.00
Skill (1982–93)	Corr 0.66, rmse 0.89, Design[c] "real"; std dev is 1.11	Corr 0.65, rmse 0.89, Design "real"; std dev is 1.11

[c]Design variations: "real" is retroactive real time, and mixed is mixture of hindcasts and independent forecasts (IF); proportion of IF years to total years available in 1982–93 is shown in row above.

the entire 1956–93 period (the cross-validation version) and in retroactive real-time mode, as was done for the CCA forecasts.

c. Overall skill

Table 1 provides basic information about the five forecast models, followed by skill scores for the 12-year period of 1982–93. A few explanatory notes about the scores are in order: The scores are based on forecasts for all times of the year—that is, 12 running 3-month target periods per year. Note that the lead times and the forecast regions are similar but not identical among the models. Both correlation skill and root-mean-square error (rmse) skill are presented. The forecasts and observations have been standardized over the period of record on which the forecast model is based (shown in Table 1). The rmse scores are computed with respect to these standardized values and thus have no physical units. Because SST variability during the 1982–93 period has been greater than that during the models' longer term base periods, the rmse scores are higher than would be expected if standardization had been done with respect to the 1982–93 period. The standard deviation of the observed SST that is already standardized over the longer term is shown in Table 1 underneath the rmse score; these exceed unity except for the NCEP coupled model that has no longer term.

To more effectively describe the skill of forecasts over shorter periods than the basic one (e.g., beginning after 1970 for the Lamont coupled model), the subperiod means are not removed in the computation of the correlation skill. More detail and the rationale of this version of the correlation are provided in the brief appendix.

The verification scores at the bottom of Table 1 indicate correlations in the 0.60s and rmse's in the 0.80s to 0.90s. These scores succinctly express the current state of the art in ENSO prediction: moderate forecasting ability at the two-season-lead time, with skills considered useful by most standards. This level of skill is comparable to forecasts of extratropical 500-mb height based on numerical weather prediction at 5–6 days' lead.

The scores also show similar overall forecast skill among the five dissimilar methods, indicating that none of them is decisively better or poorer than the others. In fact, if 19 *independent* realizations of forecast skill are assumed over the 12-year period, the 95% confidence interval of the true correlation skill

TABLE 2. Persistence skill for two SST target regions (1970–93). The mean SST for the 3-month period to be persisted (prior to $t = 0$) is used as the forecast for future periods whose center ranges from 0 to 12 months later than the center of the persisted period.

Time between centers of 3-month periods (months)	**Lead time** (months)	**Niño 3** (5°N–5°S, 120°–150°W)	**Niño 3.4** (5°N–5°S, 120°–170°W)
0	−3	1.00	1.00
3	0	0.81	0.82
6	3	0.52	0.54
9	6	0.21	0.24
12	9	−0.04	0.01

about a sample estimate of 0.65 is 0.28 to 0.85–an enormous range. With a longer period of record, as exists for some of the models, this range would narrow considerably. An assumption of 19 independent samples in the 1982–93 period is roughly approximated by the decorrelation (autocorrelation e-folding) time of observed SST of 7.5 months, but this may be too liberal in view of weaker autocorrelations (negative, then positive) at lags of over 1 year or may be too strict in view of the superposed higher-frequency variations. In any event, skill differences of, say, 0.58 versus 0.71 do not approach being statistically significantly different from one another.

Table 2 shows the mean skill of persistence forecasts for two of the predictand regions over the 1970–93 period. It is clear that all methods outperform persistence forecasts, whose skill is in the low 0.20s at 6 months' lead and reaches zero at 9 months' lead (a 1-year temporal offset). The skills obtained from the model forecasts (>0.60) are roughly equal to persistence skills at 1–2 months lead.

In the next three subsections, specific aspects of the behavior of the five models are examined in some detail. Readers whose interests are more general may wish to skip to section 4.

d. Temporal variation of skill

Table 3 shows skills for each of the five methods, as in the bottom row of Table 1, for several periods ending in December 1993. CCA and constructed analog skills using cross validation and using retroactive real-time forecasts are presented. It is evident that the skill of each method has noticeable period-dependent fluctuations and that scores have averaged higher in the more recent periods. The early 1990s, roughly the period of published real-time forecasts, have been an exception (Fig. 1). Figure 1 provides a closer look at the correlation (panel a) and rmse (panel b) skill fluctuations in the form of a 24-month moving average of skill for each model over the 1982–93 period. Note that the rmse is affected in part by the variance of the observed SST (already standardized, but only over the model's longer

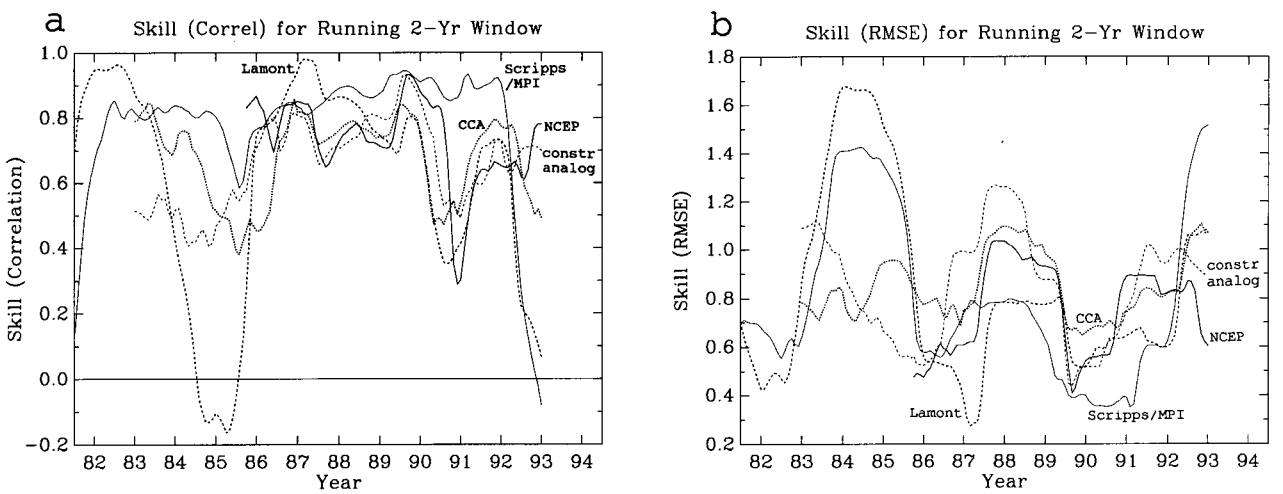

FIG. 1. Two-year moving average of skill of five forecast models for July 1981–December 1993, with skill measured using (a) correlation and (b) rmse. The thicker solid curve shows skill for NCEP coupled model forecasts, the thinner solid curve for the Scripps–MPI hybrid coupled model, the thicker dashed curve for the Lamont coupled model, the thinner dashed curve for the constructed analog model and the dotted curve for the CCA model. On the abscissa, which indicates the center of the 2-year period, tick marks indicate January. For the two-season lead time shown, NCEP coupled model forecasts are available from October 1984 (with moving average thus starting October 1985), retroactive real-time-constructed analog and CCA moving average starting January 1983, and the two other dynamical models starting earlier than July 1981.

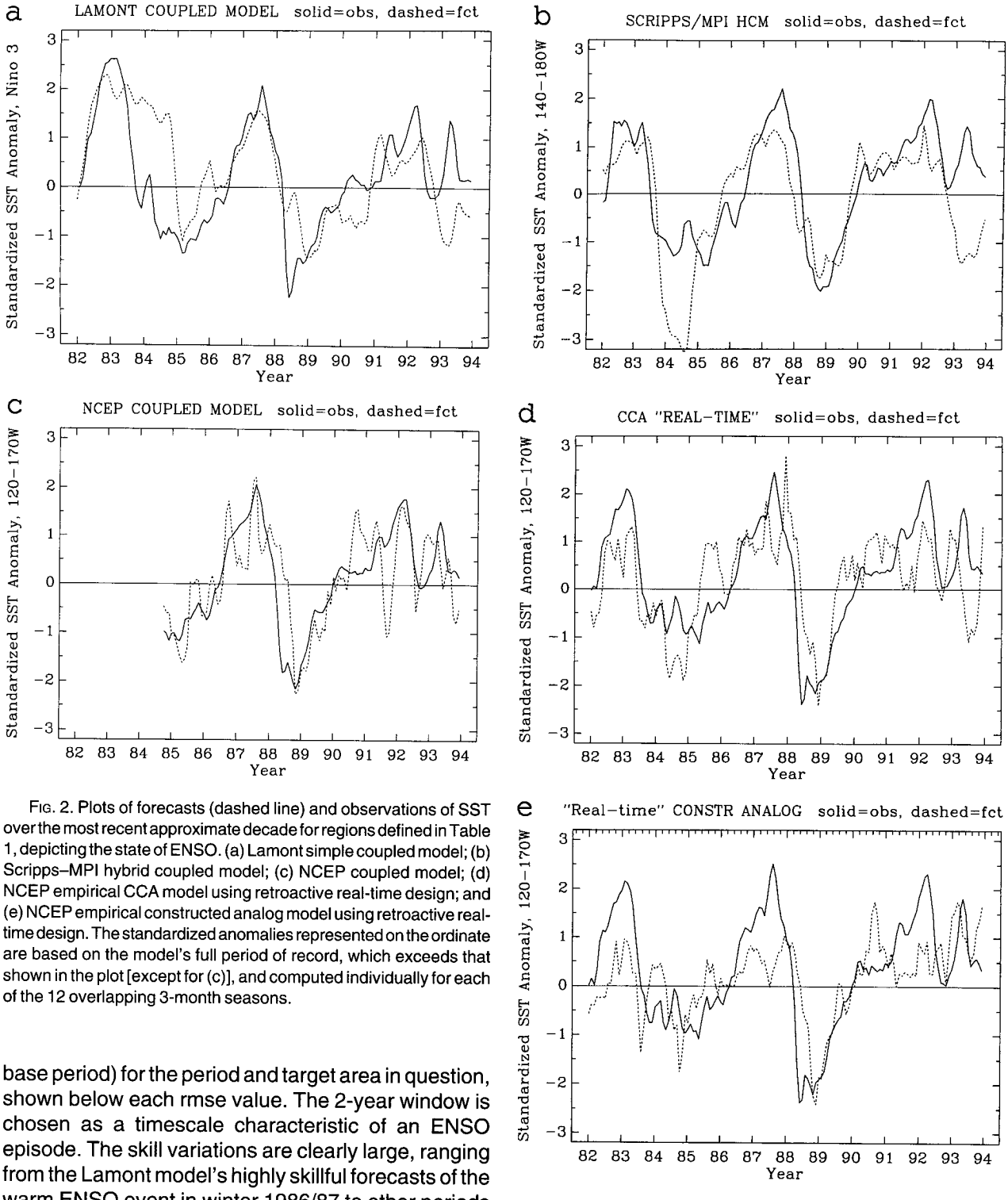

FIG. 2. Plots of forecasts (dashed line) and observations of SST over the most recent approximate decade for regions defined in Table 1, depicting the state of ENSO. (a) Lamont simple coupled model; (b) Scripps–MPI hybrid coupled model; (c) NCEP coupled model; (d) NCEP empirical CCA model using retroactive real-time design; and (e) NCEP empirical constructed analog model using retroactive real-time design. The standardized anomalies represented on the ordinate are based on the model's full period of record, which exceeds that shown in the plot [except for (c)], and computed individually for each of the 12 overlapping 3-month seasons.

base period) for the period and target area in question, shown below each rmse value. The 2-year window is chosen as a timescale characteristic of an ENSO episode. The skill variations are clearly large, ranging from the Lamont model's highly skillful forecasts of the warm ENSO event in winter 1986/87 to other periods of poor skill on the parts of several of the models.

The two measures of skill shown in Fig. 1 allow for further characterization of the error. For example, during 1984 the Scripps–MPI model has a relatively poor (high) rmse skill (Fig. 1b), but its correlation skill is not below average. This can occur during periods when forecasts and observations have high amplitude with sizable differences between them (for a high rmse) despite being on the same side of the overall mean (for a moderately high correlation).

Figure 2 contains forecast versus observation plots for the 1982–93 period (except 1984–93 for the NCEP coupled model) for the five forecast models. Only the

TABLE 3. Skill for five ENSO forecast models over several periods. Both the cross-validated and retroactive real-time versions of CCA and constructed analog are shown. The standard deviations of the observed SSTs in the target area (which already were standardized with respect to the model's longer base period)are shown beneath the rmse scores in parentheses in case normalization is desired. The rmse scores are in standardized units with respect to the longer base periods.

Correlation

Model	Beginning of period ending Dec 1993. Jan 1965	Jan 1975	Jan 1980	Jan 1985
Lamont coupled	—	0.58	0.58	0.64
Scripps–MPI hybrid coupled	0.51[b]	0.49	.64[a]	.65[a]
NCEP coupled	—	—	—	0.72
NCEP statistical CCA, cross-validated	0.54	0.43	0.59	0.68
NCEP statistical CCA, retroactive real time	—	—	—	0.64
NCEP empirical, constructed analog, cross validated	0.61	0.50	0.64	0.66
NCEP empirical, constructed analog, retroactive real time	—	—	—	0.69

[a]Standard correlations for Scripps–MPI model for periods beginning in January 1980 and January 1985 are 0.66 and 0.67, respectively (see appendix).
[b]1966.

Root-mean-square error (rmse)(standard deviation of observed SST in parentheses)

Model	Beginning of period ending Dec 1993. Jan 1965	Jan 1975	Jan 1980	Jan 1985
Lamont coupled	—	0.90 (0.97)	0.96 (1.02)	0.75 (0.96)
Scripps–MPI hybrid coupled	0.99[b] (1.00)	1.05 (1.01)	0.93 (1.03)	0.88 (1.11)
NCEP coupled	—	—	—	0.75 (1.00)
NCEP statistical CCA, cross validated	0.99 (1.04)	1.06 (1.03)	0.94 (1.05)	0.86 (1.13)
NCEP statistical CCA, retroactive real time	—	—	—	0.91 (1.13)
NCEP empirical, constructed analog, cross validated	0.85 (1.04)	0.92 (1.03)	0.85 (1.05)	0.89 (1.13)
NCEP empirical, constructed analog retroactive real time	—	—	—	0.86 (1.13)

[b]1966.

retroactive real-time versions of the CCA and constructed analog models are shown. Skills for these recent years (or for 1985–93; see Table 3) are higher than those that include earlier years. This may be related to the generally greater intensity of the recent period ENSO fluctuations (especially the warm followed by cold sequence in 1986–89), providing a strong low-frequency signal. It may also reflect improvements in the data quality (including some "cleaning up" done long after initial real-time forecasts were made, as reanalysis is a perpetual process).

Figure 2 shows that all five models predicted the strongest ENSO episodes reasonably well (especially for 1986/87, and somewhat for 1982/83 and 1988/89), and generally performed less well for the weaker fluctuations (e.g., 1991/92, spring 1993) and neutral periods (e.g., 1983/84 and 1990/91). Each model did well for some events and not as well for others, the most and least successful sets varying from model to model. There is a suggestion that the dynamical models have greater variations in performance than the empirical models. For example, the 1986/87 warm event appears to have been forecast most accurately by the Lamont model, and the 1988/89 event by the Scripps–MPI model. However, these same models made the least accurate forecasts for the unusual spring 1993 warm event. It is more difficult to find events that were forecast best or worst by one of the

TABLE 4. Intercorrelations among the five models' SST forecasts.

All seasons

Lamont	Scripps–MPI	NCEP coupled	CCA	Constr. Analog	
1.00	0.22	0.50	0.16	0.08	Lamont
0.22	1.00	0.55	0.66	0.36	Scripps–MPI
0.50	0.55	1.00	0.64	0.65	NCEP coupled
0.16	0.66	0.64	1.00	0.70	CCA
0.08	0.36	0.65	0.70	1.00	Constructed analog

Correlations among observations in the three target regions:
Niño 3.4 vs. Niño 3: 0.95; Niño 3.4 vs. 140°–180°W: 0.94; Niño 3 vs. 140°–180°W: 0.81.

Northern cold season only (December–March)

Lamont	Scripps–MPI	NCEP coupled	CCA	Constr. analog	
1.00	0.19	0.56	0.34	0.38	Lamont
0.19	1.00	0.50	0.49	0.24	Scripps–MPI
0.56	0.50	1.00	0.86	0.84	NCEP coupled
0.34	0.49	0.86	1.00	0.86	CCA
0.38	0.24	0.84	0.86	1.00	Constructed analog

Correlations among observations in the three target regions:
Niño 3.4 vs. Niño 3: 0.94; Niño 3.4 vs. 140°–180°W: 0.95; Niño 3 vs. 140°–180°W: 0.84.

empirical models, although brief periods may be detected in Fig. 1.

Figures 1 and 2 demonstrate that while a two-season-lead forecast skill since the early 1980s is useful (with overall correlation skill >0.60 and rmse below 1), there are cases of marginal and occasionally fully inadequate model performance. Differences in overall skill among the methods are fairly small. Besides possible differences in forecast ability, these skill differences may be due to the differing periods upon which model tuning was based (resulting in slightly different degrees of independence over the 1982–93 verification period), the slightly different verification periods (at least for the NCEP coupled model), and especially the differing target regions. Regarding the last item, it is found that the skill of the NCEP coupled model decreases somewhat when Niño 3 rather than Niño 3.4 is used as the target region, eliminating the small difference between its skill and that of the Lamont model (Tables 1 and 3) whose forecasts shown here are for Niño 3. It is also interesting that retroactive real-time CCA and constructed analog forecasts do not generally fare more poorly than their cross-validation counterparts. This may be peculiar to the specific period studied here, or could be a result of a negative bias that can occur in cross validation when true skills are not high (Barnston and Van den Dool 1993).

Table 4 shows correlations among the forecasts of the five models over the 1982–93 period (except starting in October 1984 for correlations with the NCEP coupled model) for all seasons (Table 4a) and for only the December through March northern winter period (Table 4b). Generally greater agreement is found for winter, when the forecast skill is highest. Because the Scripps–MPI and the Lamont models forecast regions different from those of the other three models, the correlations of their forecasts with those of the other models are naturally expected to be lower. (Note the correlations among the observations in Table 4.) The NCEP coupled model's forecast correlations also may not be fully comparable to the others due to its different period of record. Aside from these factors, there are no pairs or groups of models that yield highly similar forecasts. While some of the correlations are fairly high (e.g., CCA and constructed analog, which are both linear models that use related predictor data), the correlations generally reflect the considerable differences in the methods through which the SST forecasts are derived. The Lamont model

forecasts correlate least with the others largely because of its unique prolongation of the 1982/83 El Niño into 1984.

The forecast records of the Lamont model extend back to 1971, those of the Scripps–MPI model to 1966, and those of CCA and constructed analog (using cross validation only) to 1956. Table 3 shows that the overall skills of these models were lower than for the more recent periods. Plots of the forecasts versus observations for the longer term (as in Fig. 2, not shown) indicate slightly more frequent, lower-amplitude ENSO fluctuations in the 1950s and 1960s than those of the recent 15–20 yr. While some of the early major fluctuations were forecast correctly by two or more of the models, there were also cases of poor skill lasting 1–2 yr for two or more models. Data quality considerations, as well as the lower signal-to-noise ratios, may have contributed to the lower early period skill. This is easy to understand for CCA, in which, using cross validation, the entire 1956–93 period is used to develop predictive "rules."

e. Seasonality of skill

Our skill examination has concentrated on a single lead time averaged over all times of the year. The skill's seasonality and lead-time dependence are also of interest. While a detailed look at these is not intended here, we can show some basic findings. Table 5 shows mean correlation skill for 1982–93 (except 1984–93 for the NCEP coupled model) for the five forecast methods for 6-month lead (6.5-month lead for Lamont model) partitioned by season, where 3-month mean periods centered on December– February define northern winter, etc. Seasonality in ENSO forecast skill was first noted by Hasselmann and Barnett (1981), whose findings were extended and updated in subsequent studies. To varying degrees, the skills of all of the methods examined here clearly show seasonal variation. For the NCEP coupled model and especially for CCA and the constructed analog, forecasts made in the fall (or winter) for the following summer (or fall) have lower skill than those made at other times. This is related to the so-called spring barrier, which is difficult to traverse in a forecast (Ji et al. 1994b; Latif et al. 1994). Forecast skills of the Lamont and Scripps–MPI models appear to have somewhat less seasonal dependence. A similar analysis for the four models capable of forecasting back to 1971 (not shown) produces generally similar seasonal dependencies, although the Lamont and Scripps–MPI models show somewhat more variation than in Table 5. The decay of skill as a function of lead is seasonally dependent for all five models, where the most rapid decay occurs as spring begins being traversed in the lead period (see Xue et al. 1994). The decay rates of both the Lamont and Scripps–MPI model skill have smaller seasonal dependencies largely because they retain more skill through the spring barrier. While the seasonally averaged rate of skill decay does not vary substantially among the five models, *slightly* more skill appears to be retained from 9–12-month leads for the dynamical than for the statistical models.

f. Interpretive considerations

In assessing the performance of the five models it is important to account for differences in their forecasting situations. Perhaps more important than the slightly longer lead time used in the Lamont model is that the target region used here (Niño 3) has been found harder to predict, by most dynamical or empirical models that predict several regions, than the Niño 3.4 region, which in turn has been found slightly harder to predict than areas centered still closer to the dateline. While the Lamont model forecasts the SST field throughout the tropical Pacific, forecasts of Niño 3 are issued because that region has been thought (correctly or incorrectly) to best represent ENSO, and because the Lamont model's highest forecast skill is found in the eastern half of the Pacific rather than close to the dateline (Miller et al. 1993). The Scripps–MPI model also forecasts a complete tropical Pacific SST field, but highlights the region roughly between Niño 3 and the dateline that it predicts most successfully.

It is noted that only the NCEP coupled model availed itself of subsurface oceanic predictor data, and of course, only the three physical models used subsurface physics in the formation of their forecasts. The two empirical models did not explicitly use the subsurface in any way. The skills of any of the five models could likely be increased with the addition of some of the currently lacking features. For example, in a recent experiment 20°C isotherm depth data were added to CCA's set of predictors for the period of June 1982 through 1993, with values for 1961–82 reconstructed using analyzed wind stress. Resulting cross-validation skills were slightly increased, especially at longer leads for late spring and summer forecasts; however, CCA's performance remains at the same general level as that of the other models. Similarly, the NCEP coupled model recently underwent a refinement of its flux climatology and the installation of a MOS correction for the stress anomalies produced by the atmospheric model. Skills are improved, particularly in the eastern portion of the central tropical Pacific (i.e., Niño 3). An improved Scripps–MPI hybrid coupled model (to be called HCM-2), with higher resolution and global tropical oceanic coverage, is near completion, with expectations of higher forecast skill.

Aside from the details of skill seasonality and lead sensitivity, it appears that in general at this time the

TABLE 5. Seasonality of skill of five ENSO forecast models, based on 1982–93 (except 1984–93 for NCEP coupled model). The standard deviation of observed SST that was already standardized with respect to the models' longer base periods is indicated. The RMSE scores are in standardized units with respect to the models' longer base periods.

Model		Target season			
		Spring	Summer	Fall	Winter
Lamont coupled	CORR	0.62	0.59	0.60	0.68
Jan 1982–Dec 1993	RMSE	0.99	1.01	0.94	0.84
	Obs std dev	1.12	1.13	1.02	1.04
Scripps–MPI	CORR	0.61[a]	0.61[a]	0.69[a]	0.72[a]
hybrid coupled	RMSE	1.08	1.06	0.87	0.83
Jan 1982–Dec 1993	Obs std dev	1.19	1.10	1.03	1.06
NCEP coupled	CORR	0.76	0.65	0.59	0.89
Oct 1984–Dec 1993	RMSE	0.70	0.81	0.90	0.48
	Obs std dev	1.00	1.00	1.00	1.00
NCEP statistical CCA,	CORR	0.75	0.45	0.64	0.79
retroactive real time	RMSE	0.82	1.09	0.86	0.74
Jan 1982–Dec 1993	Obs std dev	1.16	1.12	1.05	1.06
NCEP empirical constructed	CORR	0.72	0.44	0.67	0.75
analog, retroactive real time	RMSE	0.89	1.06	0.83	0.75
Jan 1982–Dec1992/93	Obs std dev	1.16	1.12	1.05	1.06

[a]Standard correlation for the Scripps–MPI model is 0.67, 0.66, 0.73, and 0.74 for spring, summer, fall, and winter, respectively (see the appendix).

physical models examined here do not yet significantly outperform the two empirically based models, or even a local second-order autoregressive process. In fact, such an autoregressive model was developed specifically as a control for the CCA model discussed here and was found to be a tough competitor in predictive skill (section 7 of Barnston and Ropelewski 1992). Thus, the empirical and physical models have captured the essentials of interannual ENSO variability but have not yet greatly exceeded autoregressive results. There is still room for improvement. Whether a *potential* for improvement exists depends on the inherent predictability of the ocean–atmosphere system, which we are not able to assess at this point.

4. Practical implications for tropical and extratropical forecasting

a. Plans for new NWS forecast products

The skills obtained by the five different forecast models at a two-season lead, while not much better than those of second-order autoregressive models, are far superior to those of more commonly used controls such as chance or persistence. They also are high enough to be considered useful (i.e., correlation is at least 0.5 or 0.6). As a result, the two dynamical, one hybrid, and two empirical forecasts for tropical Pacific SST are being issued at lead times of up to 1 year on a quarterly basis in the *Experimental Long-Lead Forecast Bulletin* and three of them are being issued on a monthly basis in the *Climate Diagnostics Bulletin*.

Starting in January 1995, forecasts for the state of the ENSO, as indicated by some of the methods discussed above, will be issued routinely by the National Weather Service's Climate Prediction Center. In addition, "consolidated" forecasts for *United States temperature and precipitation* will begin being issued. The long-lead forecasts progressing out to one year for both ENSO and U. S. surface climate will be issued monthly as the *Climate Outlook*. The current *Monthly and Seasonal Weather Outlook,* that has contained zero-lead 30- and 90-day forecasts since the 1970s, will no longer be issued. The tools contributing to the final, single U.S. surface climate forecasts will be CCA, the NCEP coupled model, and the optimal climate normals approach (Huang et al. 1994). The

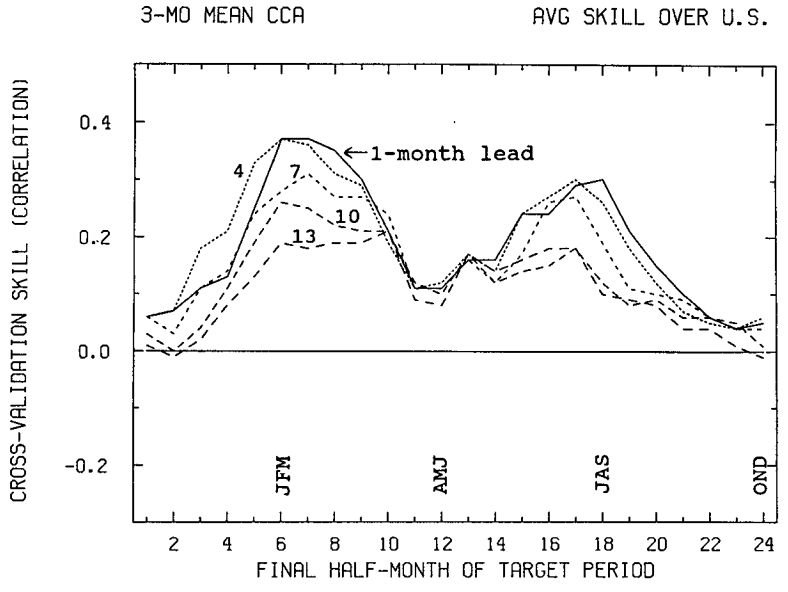

FIG. 3. The annual cycle of cross-validation correlation skill of CCA forecasts of 3-month mean temperatures averaged over 59 stations across the continental United States. Season 1 is the 3-month period ending in mid-January, season is the period ending at the end of January, etc. Shown are lead times of 1 month (solid curve), 4 months (shortest dashed curve), 7 months (second shortest dashed curve), 10 months, and 13 months (longest dashes). Seasons progress by half-month increments.

latter is an empirical technique in which the average condition over the previous k yr for the time of year being forecast is persisted, where k varies spatially and seasonally in order to maximize local predictive skill. The constructed analog method is now in an experimental stage (Van den Dool 1994) and may become a contributing tool in the future. Tools from outside of NCEP (e.g., the Scripps–MPI hybrid coupled model used with an atmospheric GCM; Barnett 1994) also may eventually be invited to contribute.

b. Extratropical forecasts

Based on experimentation with CCA-based forecasts (Barnston 1994), the most skillful predictors for U.S. surface climate are the fields of global SST and Northern Hemisphere 700-mb height at several consecutive prior 3-month periods. Figure 3 shows the correlation skill for the 1956–93 period, averaged over 59 fairly equally distributed U.S. stations, as a function of season for several lead times (see Fig. 3 caption). It is noteworthy that there is a modest level of skill at certain times of the year—late winter and late summer—that decreases only slowly with increasing lead time. Other empirical forecast methods have shown maxima at these times of year as well, such as analog forecasting (Barnston and Livezey 1989; Livezey 1990) and persistence (Van den Dool 1983, Van den Dool et al. 1986). Figure 4 shows the geographical distribution of 1956–93 CCA skill over the U.S. for the January–March season at 6 months' lead—that is, for forecasts made at the end of the previous June. Portions of the Northeast have skill comparable to that for forecasts of the tropical Pacific SST at the same lead time, with lower but statistically significant skill in other areas. Part but not all of the skill in winter has its origin in ENSO, as shown by Ropelewski and Halpert (1986), and also clearly identified in the CCA-produced diagnostics (not shown). In fact, the skill is higher when only warm or cold ENSO year forecasts are considered. The one-season-lead skill in winter (not shown) has by now a classic ENSO spatial pattern, peaking along the Gulf of Mexico and southeastern states, the northern tier from the Great Lakes to the northern Rockies, and somewhat along the West Coast. The 6-month-lead forecast (Fig. 4) shows a less dominant ENSO pattern because observed predictor data through the prior June leaves uncertainty regarding the ENSO condition to be expected 7–8 months later. The SST in other parts of the globe may have a larger share of influence, possibly on timescales other than that of ENSO, when the tropical Pacific SST influence is uncertain.

If the winter tropical Pacific SST itself were moderately well predicted at a two-or-more-season lead time using a dynamical ocean model and supplied to a statistical atmospheric model such as CCA, greater skill in forecasting seasonal mean U.S. surface climate than that demonstrated above using CCA alone might be expected. The potential benefit of a two-

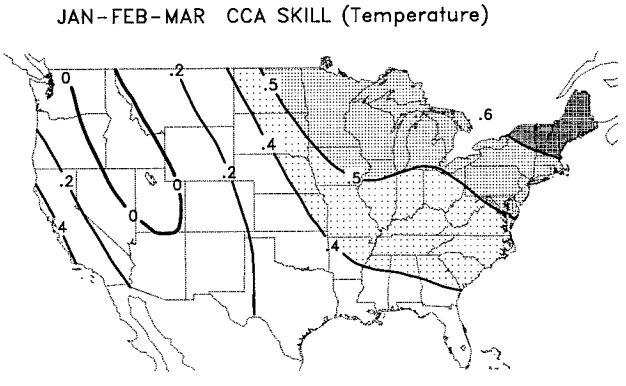

FIG. 4. Geographical distribution of cross-validated CCA correlation skill across the United States for forecasts of January–March at 6-month lead (i.e., using data through prior June) for 38 yr (1956–93). Mean skill is 0.32 (Heidke skill 0.18).

tiered forecast process, proposed by Bengtsson et al. (1993) and Barnett et al. (1994), was demonstrated for forecasts of extratropical Northern Hemisphere winter 700-mb height using the Scripps–MPI hybrid coupled model with a variant of CCA (Graham and Barnett 1994). These forecasts are competitively skillful in certain regions and/or seasons, especially during warm and cold ENSO episodes. A similar, if not higher, level of skill was also obtained using the dynamically predicted tropical SST to force a state-of-the-art atmospheric GCM developed at MPI (the ECHAM3 model; Roeckner et al. 1992). Such a dual dynamical system is the Scripps–MPI two-tiered coupled model (Barnett et al. 1994), one of whose recent successes is demonstrated by the geographical distribution of skill for forecasts of January–February precipitation anomalies in the United States at two and one-third-season lead for strong ENSO years (Fig. 5). Similar experimentation has been performed with the NCEP coupled model, with comparably successful results during strong ENSO episodes. Consequently, the two-tiered version of the NCEP coupled model is one of the several tools now being operationally implemented at NCEP.

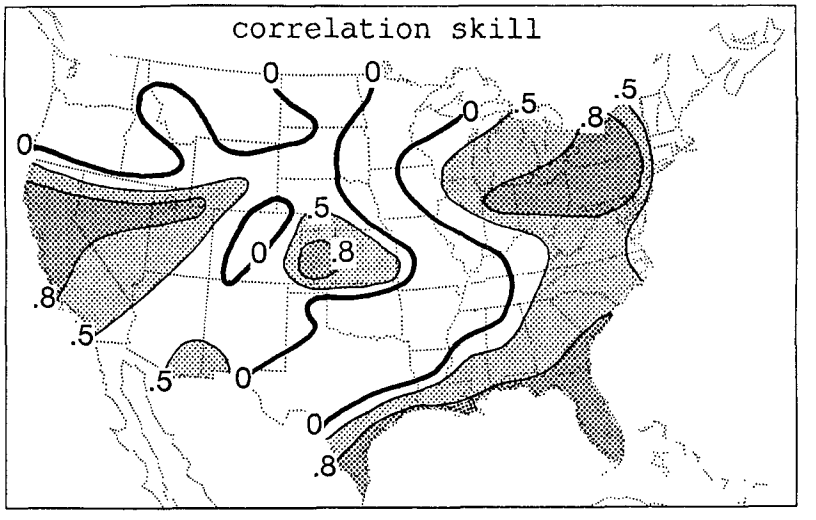

FIG. 5. Geographical distribution of skill of the Scripps–MPI hybrid coupled model forecasts for January–February total U.S. precipitation at two and one-third season lead (i.e., using data through prior May) for seven strong ENSO events. Skill is expressed as a correlation between model forecasts and observations.

c. Utility

The dissemination of measurably skillful forecasts with much longer lead times by NCEP–CPC should be useful to a wide variety of users, particularly for industrial or commercial applications and government planning. They will be less well suited for the general public, partly because they are expected to verify more successfully over an integrated time period (e.g., an entire 5-month "winter," or several consecutive winters) than for shorter embedded periods that interest private citizens most frequently. However, these forecasts will be limited in that 1) ENSO forecasts for late spring through fall at 6 months' lead are relatively unskillful because of the spring barrier mentioned above, and 2) occasional skill "droughts" lasting as long as 1–2 yr may occur despite the moderately skillful expected average performance. In the forecasts to be issued by CPC beginning in mid-December 1994, the expected forecast skill will be expressed using uncertainty indicators (e.g., probabilities, or the associated "error bars") for U.S. surface climate as well as ENSO. For U.S. temperature and precipitation, for example, this will convey a nearly complete lack of forecast skill in late spring (April–June) and late fall at many locations. In late winter (January–March), while temperature forecasts at short to moderate lead times are relatively skillful in the ENSO-influenced regions (the Gulf, the Southeast, and the north-central to northwestern tier), other regions such as the central Rockies, the Great Basin, and central plains have negligible expected skill. As implied by potential predictability studies (Shea and Madden 1990), time-averaged winter weather as a response to persistent remote boundary condition anomalies may be lacking in these regions whose climate is governed more exclusively by synoptic weather events.

While the United States temperature and precipitation forecasts will, as before, be expressed probabilistically to express uncertainty, a new feature of the forecast format will be the distinction between regions having relatively confident forecasts for near-normal conditions and a complete lack of confidence, resulting in climatological probabilities. The former condition cannot be expressed in the existing CPC forecast format because the "near-normal" category has been assumed to lack any predictive skill (Gilman 1985; Van den Dool and Toth 1991). Using the above-described set of more advanced empirical and dynamical tools, we are now occasionally able to forecast near-normal conditions with measurable confidence.

5. Conclusions and discussion

The five models discussed here are some initial tools in what we expect will be an increasingly rich supply of oceanic and atmospheric climate forecast-

ing instruments. Each was chosen for a reason. The CCA and constructed analog methods establish a basis of skill from empirical information; the Lamont model, whose forecasts have become familiar to the science community, is thought to contain the essential physics; the Scripps–MPI hybrid coupled model is a blend of methods with its sophisticated ocean model and a "slave" statistical atmosphere; and finally, the comprehensive NCEP coupled model, with its initialization using a real-time ocean data assimilation system, holds great potential. A few other models currently able to produce real-time forecasts of equatorial Pacific SST are mentioned in section 1, and still others are becoming or are already capable of such forecasts.

A moderate level of mean skill has been demonstrated by all the methods examined here for forecasting tropical Pacific SST, with correlation skill averaging in the 0.60s at a lead time of two seasons. (Recall our strict definition of lead time—the time that is *completely* skipped.) Within the range of statistical uncertainty, the overall skills of all the methods discussed here are equal. However, the skill of any of the methods varies over time about its average, resulting in periods of very high as well as very low skill. With an expression of uncertainty to accompany the forecasts, users will have the best opportunity to understand and benefit from them. This will be especially valuable (and possibly frustrating) for users, many of whom strongly prefer a categorical forecast product despite the inherently probabilistic nature of forecasts of both ENSO and U.S. surface climate.

In the skill evaluations presented here, we attempted to use simulations of independent forecasts (cross validation, or retroactive real time) as equally as possible for the five models. A statistical model's automatic usage of future data to make forecasts for an intermediate time in a cross-validation setting cannot be mimicked in a physical model integration. The closest equivalent for a physical model is to forecast periods that were not used in any way in the model construction and tuning process. More than half of the skill evaluation period used here satisfied that condition. Another step toward a real-time simulation is to exclude the dynamical model's target period from the computation of the model's observed climatology statistics (as if it were a future period).

That skills comparable to those of the dynamical models described above are currently approximated with a completely statistical model at 6 months' lead suggests that the potential of dynamical models may not yet be fully realized. If the ocean–atmosphere system contains sufficient inherent predictability, the dynamical models should be able to outperform models that do not use the equations of physical oceanic and atmospheric motion or cannot accommodate nonlinearity.

The importance of quality control of real-time data, and good maintenance of the overall database, cannot be overemphasized. It is the skill of true real-time forecasts that will eventually be evaluated. This skill can measure up to estimates of independent forecast skill (e.g., retroactive real-time simulations) only if observational errors or real-time data assimilation failures occur very infrequently.

To make an adequate assessment, the skill of a forecast method should only be studied using a large set of forecasts made to simulate independent or real-time conditions. Often, too much is read into the outcome of a small set of forecasts, or even a single forecast. By absolute standards, the level of skill of two-season-lead forecasts is modest, and given the small number of degrees of freedom for an area the size of the United States, it should come as no surprise to obtain some very good as well as very poor forecasts. In long-range forecasting, it takes over 10 years to obtain a sample of forecasts sufficient to perform a meaningful verification study. ENSO episodes increase the time between independent climatic realizations, which may occur far less than once per season. Because we cannot wait that long to test a single method, it is imperative that methods be developed to more accurately estimate skill of future operational forecasts using hindcast verification. This issue has received plenty of attention for statistical methods but too little for forecasts made by dynamical methods, particularly when these methods are run repeatedly on the same datasets in slightly different configurations of the model.

Because we cannot wait that long to test a single method, it is imperative that methods be developed to more accurately estimate skill of future operational forecasts using hindcast verification.

In conclusion, over the last decade our achievement in forecasting ENSO fluctuations has been quite substantial, especially in view of how recently we had virtually nothing. We now have moderate capability at two-season lead, and more modest but statistically significant capability at leads of over 1 yr. An analogous evolution of skill in short-range forecasting using numerical weather prediction occurred in the 1960s and 1970s. Our ability to forecast extratropical climatic conditions, however, with the exception of specific

regions for certain times of the year, is considerably weaker. We need to continue our efforts in both areas, concentrating heavily on model design but also on dataset quality and on verification methods. While future achievements will probably occur gradually, there is every reason to expect that our long-term efforts will result in additional knowledge, and, it is hoped, higher forecast skill.

Appendix

In computing the correlation skill of forecasts over shorter periods than the basic one (e.g., beginning after 1966 for the Scripps–MPI hybrid coupled model), the subperiod means are not removed and are not used for computing the standard deviation terms. Rather, the basic period mean (which is zero in the standardized data used in the analyses here) is used. Such a coefficient is bounded by +/–1 as is the conventional correlation coefficient. This is done so that, for example, if in the subperiod the forecasts and observations showed small-amplitude out-of-phase variations but both were generally on the same side of the longer base period mean, a positive correlation would result, and we believe justifiably. The standard correlation coefficient would be negative in this situation. On the other hand, if there were in-phase variations but the forecasts and observations were on opposite sides of the mean (indicating a subperiod forecast bias, which can be viewed as a general miss), a negative correlation would result. The standard correlation coefficient would be positive in that case. The two versions of the correlation are identical when the subperiod and the basic period are equal, as for the NCEP coupled model.

The correlation scores of three of the other four models are slightly increased, overall, with use of the modified version, because their forecasts were successful in terms of the general anomaly sign over the 1982–93 period (both forecasts and observations averaged above the long-term mean), but less successful in the details of timing and magnitude. A few minor exceptions to this occur in the period- or season-specific skills (Tables 3 and 5) and are not indicated. However, a mild cold bias is generally found in the Scripps–MPI model forecasts (its forecasts for 1982–93 average slightly below its long-term forecast mean), but its timing of subperiod fluctuations is successful, causing the conventional correlation to be slightly higher than the modified version. Because this type of error is easier to correct than nonsystematic errors, we indicate in footnotes in the score tables the conventional correlation coefficient for the Scripps–MPI model when it is higher than the modified version.

References

Barnett, T.P., 1981: Statistical prediction of North American air temperatures from Pacific predictors. *Mon. Wea. Rev.*, **109**, 1021–1041.

——, and R.W. Preisendorfer, 1978: Multifield analog prediction of short-term climate fluctuations using a climate state vector. *J. Atmos. Sci.*, **35**, 1771–1787.

——, and ——, 1987: Origins and levels of monthly and seasonal forecast skill for North American surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev.*, **115**, 1825–1850.

——, N. Graham, M. Cane, S. Zebiak, S. Dolan, J. O'Brien, and D. Legler, 1988: On the prediction of the El Niño of 1986–87. *Science*, **241**, 192–196.

——, M. Latif, N. Graham, M. Flugel, S. Pazan, and W. White, 1993: ENSO and ENSO-related predictability. Part 1: Prediction of equatorial Pacific sea surface temperature with a hybrid coupled ocean–atmosphere model. *J. Climate*, **6**, 1545–1566.

——, L. Bengtsson, K. Arpe, M. Flugel, N. Graham, M. Latif, J. Ritchie, E. Roeckner, U. Schlese, U. Schulzweida, and M. Tyree, 1994: Forecasting global ENSO-related climate anomalies. *Tellus*, **46A**, 381–397.

Barnston, A. G., 1994: Linear statistical short-term climate predictive skill in the Northern Hemisphere. *J. Climate*, **7**, 1513–1564.

——, and R.E. Livezey, 1989: An operational multifield analog/antianalog prediction system for United States seasonal temperatures. Part II: Spring, summer, fall, and intermediate three-month period experiments. *J. Climate*, **2**, 513–541.

——, and C.F. Ropelewski, 1992: Prediction of ENSO episodes using canonical correlation analysis. *J. Climate*, **5**, 1316–1345.

——, and H.M. van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate*, **6**, 964–977.

Bengtsson et al., 1993: A two-tiered approach to long-range climate forecasting. *Science*, **261**, 1026–1029.

Bryan, K., 1969: A numerical method for the study of the World Ocean. *J. Comput. Phys.*, **4**, 347–376.

Cane, M.A., 1992: Tropical Pacific ENSO models: ENSO as a mode of the coupled system. *Climate System Modeling*, K.E. Trenberth, Ed., Cambridge University Press, 583-614.

——, and S.E. Zebiak, 1987: Prediction of El Niño events using a physical model. *Atmospheric and Oceanic Variability*, H. Cattle, Ed., Royal Meteorological Society, 153–181.

——, ——, and S.C. Dolan, 1986: Experimental forecasts of El Niño. *Nature*, **321**, 827–832.

Climate Diagnostics Bulletin: Near real-time analyses, ocean/atmosphere. V. E. Kousky, Ed.

Cox, M.D., 1984: A primitive, 3-dimensional model of the ocean. GFDL Ocean Group Tech. Rep. No. 1., Geophysical Fluid Dynamics Laboratory, 143 pp.

Davis, R.E., 1978: Predictability of sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.*, **8**, 233–246.

Elsner, J.B., and C.P. Schmertmann, 1993: Improving extended-range seasonal predictions of intense Atlantic hurricane activity. *Wea. Forecasting*, **8**, 345–351.

Experimental Long-Lead Forecast Bulletin, A. G. Barnston, Ed.

Gilman, D.L., 1985: Long-range forecasting: The present and the future. *Bull. Amer. Meteor. Soc.*, **66**, 159–164.

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

Graham, N.E., 1994: Experimental predictions of wet season precipitation in northeastern Brazil. *Proc. of the 18th Annual Climate Diagnostics Workshop*, Climate Analysis Center, NOAA, Boulder, CO, 378–381.

——, and T.P. Barnett, 1994: ENSO and ENSO-related predictability. Part 2: Northern Hemisphere 700-mb predictions based on a hybrid coupled ENSO model. *J. Climate*, **7**, in press.

——, J. Michaelsen, and T. P. Barnett, 1987a: An investigation of the El Niño–Southern Oscillation cycle with statistical models. 1. Predictor field characteristics. *J. Geophys. Res.*, **92**, 14 251–14 270.

——, ——, and ——, 1987b: An investigation of the El Niño–Southern Oscillation cycle with statistical models. 2. Model results. *J. Geophys. Res.*, **92**, 14 271–14 289.

——, T.P. Barnett, and M. Latif, 1992: Considerations of the predictability of ENSO with a low-order coupled model. *Proc. of the 16th Annual Climate Diagnostics Workshop*, Climate Analysis Center, NOAA, Los Angeles, CA, 323–329.

Gray, W. M., C. W. Landsea, P. Mielke, and K. Berry, 1993: Predicting Atlantic basin seasonal tropical cyclone activity by 1 August. *Wea. Forecasting*, **8**, 73–86.

Hasselmann, K., and T.P. Barnett, 1981: Techniques for linear prediction for systems with periodic statistics. *J. Atmos. Sci.*, **38**, 2275–2283.

Hastenrath, S., and L. Greischar, 1993: Further work on the prediction of northeast Brazil rainfall anomalies. *J. Climate*, **6**, 743–758.

Horel, J.D., and J.M. Wallace, 1981: Planetary-scale atmospheric phenomena associated with the Southern Oscillation. *Mon. Wea. Rev.*, **109**, 813–829.

Huang, J., H.M. van den Dool, and A.G. Barnston, 1994: Seasonal temperature prediction with three season lead using optimal climate normals. *J. Climate*, **7**, submitted.

Inoue, M., and J.J. O'Brien, 1984: A forecasting model for the onset of a major El Niño. *Mon Wea. Rev.*, **112**, 2326–2337.

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

——, ——, and ——, 1994b: An experimental coupled forecast system at the National Meteorological Center: Some early results. *Tellus*, **46A**, 398–418.

——, A. Leetmaa, and J. Derber, 1994c: An ocean analysis system for climate studies. *Mon. Wea. Rev.*, **122**, in press.

Keppenne, C.L., and M. Ghil, 1992: Adaptive filtering and prediction of the Southern Oscillation Index. *J. Geophys. Res.*, **97**, 20 449–20 454.

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

——, and M. Flugel, 1991: An investigation of short range climate predictability in the tropical Pacific. *J. Phys. Oceanogr.*, **96**, 2661–2673.

——, A. Sterl, E. Maier-Reimer, and M.M. Junge, 1993a: Climate variability in a coupled GCM. Part I: The tropical Pacific. *J. Climate*, **6**, 5–21.

——, ——, ——, and ——, 1993b: Structure and predictability of the El Niño/Southern Oscillation phenomenon in a coupled ocean–atmosphere general circulation model. *J. Climate*, **6**, 700–708.

——, T.P. Barnett, M.A. Cane, M. Flugel, and N.E. Graham, 1994: A review of ENSO prediction studies. *J. Climate*, **7**, submitted.

Livezey, R.E., 1990: Variability of skill of long-range forecasts and implications for their use and value. *Bull. Amer. Meteor. Soc.*, **71**, 300–309.

——, and A.G. Barnston, 1988: An operational multifield analog/antianalog prediction system for United States seasonal temperatures. 1. System design and winter experiments. *J. Geophys. Res.*, **93**, 10 953–10 974.

Miller, A.J., T.P. Barnett, and N. E. Graham, 1993: A comparison of some tropical ocean models: Hindcast skill and El Niño evolution. *J. Phys. Oceanogr.*, **23**, 1567–1591.

Namias, J., 1964: A 5-year experiment in the preparation of seasonal outlooks. *Mon. Wea. Rev.*, **92**, 449–464.

Neelin, J.D., 1990: A hybrid coupled general circulation model for El Niño studies. *J. Atmos. Sci.*, **47**, 674–693.

Pacanowski, R.C., and S.G.H. Philander, 1981: Parameterization of vertical mixing in numerical models of tropical oceans. *J. Phys. Oceanogr.*, **11**, 1443–1451.

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

Philander, S.G.H., W.J. Hurlin, and A.D. Seigel, 1987: A model of the seasonal cycle in the tropical Pacific Ocean. *J. Phys. Oceanogr.*, **17**, 1986–2002.

Preisendorfer, R.W., and C.D. Mobley, 1984: Climate forecast verifications, United States mainland, 1974–83. *Mon. Wea. Rev.*, **112**, 809–825.

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

Roeckner, E., K. Arpe, L. Bengtsson, S. Brinkop, L. Dumenil, M. Esch, E. Kirk, F. Lunkeit, M. Ponaier, B. Rockel, R. Sausen, U. Schlese, S. Schubert, and M. Windelband, 1992: Simulation of the present-day climate with the ECHAM model: Impact of model physics and resolution. Max Planck Institut fur Meteorologie, Rep. No. 93, 172 pp.

Ropelewski, C.F., and M.S. Halpert, 1986: North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). *Mon. Wea. Rev.*, **114**, 2352–2362.

——, and ——, 1987: Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation (ENSO). *Mon. Wea. Rev.*, **115**, 1606–1626.

Shapiro, L., 1982: Hurricane Climatic Fluctuations. Part II: Relation to large-scale circulation. *Mon. Wea. Rev.*, **110**, 1014–1023.

Shea, D.J., and R.A. Madden, 1990: Potential for long-range prediction of monthly mean surface temperatures over North America. *J. Climate*, **2**, 1444–1451.

Slutz, R., S. J. Lubler, J. D. Hiscox, S. D. Woodruff, R. J. Jenne, D. H. Joseph, P. M. Steurer, and J. D. Elius, 1985: Comprehensive Ocean Atmosphere Data Set. National Oceanic and Atmospheric Administration, Boulder, CO, 268 pp. [Available from Climate Research Program, ERL, R/E/AR6, 325 Broadway, Boulder, CO 80303.]

Van den Dool, H.M., 1983: A possible explanation of the observed persistence of monthly mean circulation anomalies. *Mon. Wea. Rev.*, **111**, 539–544.

——, 1987: A bias in skill in forecasts based on analogs and antilogues. *J. Climate Appl. Meteor.*, **26**, 1278–1281.

——, 1994: Searching for analogues, how long must we wait? *Tellus*, **46A**, 314–324.

——, and Z. Toth, 1991: Why do forecasts for "near normal" often fail? *Wea. Forecasting*, **6**, 76–85.

——, W.H. Klein, and J.E. Walsh, 1986: The geographical distribution and seasonality of persistence in monthly mean air temperatures over the United States. *Mon. Wea. Rev.*, **114**, 546–560.

Wagner, A.J., 1989: Medium- and long-range forecasting. *Wea. Forecasting*, **4**, 413–426.

——, and R.E. Livezey, 1984: Applications of Monte Carlo methods to estimate the skill significance of a small sample of experimental seasonal forecasts. *Proc. of the 8th Annual Climate Diagnostics Workshop*, Climate Analysis Center, NOAA, Downsview, Ontario, Canada, 387–393.

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

Wyrtki, K., 1985: Water displacements in the Pacific and the genesis of El Niño cycles. *J. Geophys. Res.*, **90**, C4, 7129–7132.

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

Xue, Y., M.A. Cane, S.E. Zebiak, and M.B. Blumenthal, 1994: On the prediction of ENSO: A study with a low order Markov model. *Tellus*, **46A**, 512–528.

Zebiak, S.E., 1984: Tropical atmosphere-ocean interaction and the El Niño/Southern Oscillation phenomenon. Ph.D. thesis, Massachusetts Institute of Technology, 261 pp.

——, and M.A. Cane, 1987: A model El Niño–Southern Oscillation. *Mon. Wea. Rev.*, **115**, 2262–2278.