Present-Day Capabilities of Numerical and Statistical Models for Atmospheric Extratropical Seasonal Simulation and Prediction

Jeffrey Anderson,* Huug van den Dool,+ Anthony Barnston,+
Wilbur Chen,+ William Stern,* and Jeffrey Ploshay*

ABSTRACT

A statistical model and extended ensemble integrations of two atmospheric general circulation models (GCMs) are used to simulate the extratropical atmospheric response to forcing by observed SSTs for the years 1980 through 1988. The simulations are compared to observations using the anomaly correlation and root-mean-square error of the 700-hPa height field over a region encompassing the extratropical North Pacific Ocean and most of North America. On average, the statistical model is found to produce considerably better simulations than either numerical model, even when simple statistical corrections are used to remove systematic errors from the numerical model simulations. In the mean, the simulation skill is low, but there are some individual seasons for which all three models produce simulations with good skill.

An approximate upper bound to the simulation skill that could be expected from a GCM ensemble, if the model's response to SST forcing is assumed to be perfect, is computed. This perfect model predictability allows one to make some rough extrapolations about the skill that could be expected if one could greatly improve the mean response of the GCMs without significantly impacting the variance of the ensemble. These perfect model predictability skills are better than the statistical model simulations during the summer, but for the winter, present-day statistical forecasts already have skill that is as high as the upper bound for the GCMs. Simultaneous improvements to the GCM mean response and reduction in the GCM ensemble variance would be required for these GCMs to do significantly better than the statistical model in winter. This does not preclude the possibility that, as is presently the case, a statistical blend of GCM and statistical predictions could produce a simulation better than either alone.

Because of the primitive state of coupled ocean–atmosphere GCMs, the vast majority of seasonal predictions currently produced by GCMs are performed using a two-tiered approach in which SSTs are first predicted and then used to force an atmospheric model; this motivates the examination of the simulation problem. However, it is straightforward to use the statistical model to produce true forecasts by changing its predictors from simultaneous to precursor SSTs. An examination of the decrease in skill of the statistical model when changed from simulation to prediction mode is extrapolated to draw conclusions about the skill to be expected from good coupled GCM predictions.

1. Introduction

The 1997–98 ENSO warm event has led to a resurgence of interest in the problem of seasonal and interannual prediction of the climate system. While many of the effects in the tropical atmosphere that are associated with anomalous tropical SSTs are relatively deterministic (Anderson and Stern 1996), the natural internal atmospheric variability of the extratropical atmosphere can serve to obscure the impact of tropical SST forcing (Kumar and Hoerling 1998). Despite this, there is strong evidence that the statistics of the extratropical climate still depend on the tropical SSTs, especially when these SSTs are strongly anomalous (Graham et al. 1987a,b; Barnston 1994).

A variety of methods for predicting the extratropical response to SST forcing have been developed. These methods include numerical models (Kumar et al. 1996), statistical models (Barnston and Smith

*Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey.
+National Centers for Environmental Prediction/Climate Prediction Center, Camp Springs, Maryland.
Corresponding author address: Dr. Jeffrey L. Anderson, Geophysical Fluid Dynamics Laboratory, Princeton University, P.O. Box 308, Princeton, NJ 08542.
E-mail: jla@gfdl.gov
In final form 29 March 1999.

1996), and a large assortment of hybrid statistical–dynamical models (Graham and Barnett 1995; Sarda et al. 1996). Statistical models are dependent upon the quality and quantity of historical observations of the ocean and atmosphere but are generally far less expensive to develop and run than are numerical models. Fully coupled ocean–atmosphere GCMs have become one popular tool for doing seasonal prediction/simulation, but these models can be exceptionally costly to develop, to integrate, and to validate.

In this study, an attempt is made to assess the capabilities of one statistical model and two numerical models to simulate the extratropical atmosphere given the SSTs observed during the simulation period. This is a simpler problem than the complete forecast problem in which only SSTs observed prior to the beginning of the forecast are provided to the forecast models. However, the present state of development of GCM prediction models is primitive enough that it can be useful to explore the simulation problem first. With some caveats (Wittenberg and Anderson 1998), the model skill for the simulation problem can be regarded as an upper bound for the skill that could be obtained in the full forecast problem.

The GCMs discussed here were developed independently at the Geophysical Fluid Dynamics Laboratory (GFDL) and the National Centers for Environmental Prediction (NCEP) and have been used for a variety of simulation and prediction experiments. For both of these models, an ensemble of integrations forced by observed SSTs is used to simulate the atmosphere. The statistical model is a canonical correlation analysis (CCA) that recently has seen widespread operational use in a forecast mode. Here, the CCA is also used in a simulation mode to allow a comparison that is as fair as possible. The three models are compared for seasonal mean simulations of a single field, 700-hPa height, over a single extratropical region, one associated with the Pacific–North American (PNA) pattern, using identical verification metrics.

The use of ensemble simulations with the numerical models may facilitate the computation of a priori estimates of simulation skill. The correlation between a measure of the ensemble spread and the simulation skill is examined to see the extent to which one can identify skillful simulations a priori.

The ensembles also allow the computation of a rough upper bound on model skill assuming that the variance of the ensemble responses is approximately correct, even if the mean of the response is not perfect. Comparing this perfect model predictability (Chen and van den Dool 1997) to the actual model simulation skill and to the skill of the CCA allows one to make some rough extrapolations about the skill that could be expected if one could greatly improve the mean response of the models without significantly impacting the variance of the ensemble. To exceed this level of skill, the variance of the ensemble response would have to be reduced in concert with a decrease in the mean error.

Section 2 presents an overview of the three simulation models and the verification metrics used for comparison. Sections 3 and 4 present comparisons of the numerical and statistical model simulations using the anomaly correlation. Section 5 explores the perfect model predictability and compares this bound for the numerical models to the simulation skill of the CCA. Section 6 briefly examines sensitivity to verification metric by using the rms error while section 7 presents discussion and conclusions.

2. Models and evaluation techniques

This study compares the capabilities of two numerical models and a statistical model to simulate the extratropical seasonal mean circulation given the global SSTs for a 10-yr simulation period. The two numerical models are recent-generation GCMs from NCEP and GFDL. Both models have known systematic errors and parameterization deficiencies. The statistical model is an application of the canonical correlation analysis approach of Barnston et al. (1994). This section gives details of the two GCMs and the CCA and concludes with a discussion of the metrics chosen to compare the quality of extratropical seasonal simulations.

a. GFDL experimental prediction model

The first GCM is the GFDL experimental prediction global spectral model, an 18-level spectral model truncated at T42 described in Gordon and Stern (1974, 1982). This version of the model uses "bucket" hydrology (Manabe 1969); orographic gravity wave drag (Stern and Pierrehumbert 1988); large-scale condensation and moist convective adjustment (both using a condensation criterion of 100%); shallow convection (Tiedtke 1988); cloud prediction (Gordon 1992); 12-h averaged seasonally varying radiative transfer; stability dependent vertical eddy fluxes of heat, momentum, and moisture throughout the surface layer and the free atmosphere (Sirutis and Miyakoda 1990); and ∇^4 hori-

zontal diffusion. Surface temperatures over land and sea ice are determined by solving a surface heat balance equation (Gordon and Stern 1982).

An eight-member ensemble forced by observed Atmospheric Modeling Intercomparison Project (AMIP) SSTs (Gates 1992) was integrated for 10 yr from 1 January 1979 through 31 December 1988. The initial conditions for the ensembles were taken from analyses for 12 December 1978 through 21 January 1979, sampled every 5 days. Each of these analyses was then used as an initial condition as if it were the analysis for 1 January 1979. A more detailed description of the ensemble integrations can be found in Stern and Miyakoda (1995). The first 12 months of the ensemble integration were discarded in an attempt to eliminate direct effects of the initial conditions including effects of soil moisture spinup. Because this model experiences a slow loss of mass, the *global* mean of the 700-hPa height field was removed from each seasonal mean simulation and from the observations before evaluating skill.

b. NCEP Medium-Range Forecast Model with climate resolution

The NCEP GCM used here is a modified version of the global short-range forecast model described in Kanamitsu et al. (1991). The model has a T40 spectral truncation with 18 vertical levels. The version of the NCEP model used here is close to the MRF9 model discussed in the appendix of Kumar et al. (1996). The altered parameterizations include a modified Kuo convective scheme (Ji et al. 1994) and a modified version of the Slingo and Slingo (1991) cloudiness parameterization. These differences from the GFDL model, along with a variety of differences in other parameterization schemes, make the two GCMs quite distinctive.

A 13-member ensemble was integrated for 45 yr from 1 January 1950 with the members differing only in the details of the initial conditions. Global SSTs were specified from NCEP's SST analyses. For consistency with the available GFDL model runs, only the years 1980 through 1988 are evaluated in this study. This set of ensemble integrations has been discussed in a number of reports including Chen and van den Dool (1997) and Kumar et al. (1996).

c. Canonical correlation analysis

CCA is a multivariate regression that linearly relates selected historically observed predictor field patterns to observed predictand field patterns. In the particular model used here, the predictor consists of 3-month means of near-global SST, and the predictand is 700-hPa geopotential heights at 133 grid points covering the PNA region during the same 3-month period. This is the same prediction design used in the CCA global climate specification experiments of Barnston and Smith (1996), and includes the preorthogonalization step used in Barnett and Preisendorfer (1987) and subsequently implemented by the National Weather Service as described in Barnston (1994). That is, rather than inputting raw predictor and predictand data into the CCA, a truncated set of empirical orthogonal functions of the predictors and the predictands are analyzed by the CCA, and the prediction is expanded back to geographical space as a final step. The SST and height datasets used for CCA model development span the relatively long 1950–96 sampling period, but the simulations examined here span just the 1980–88 period for which both numerical models have been integrated. In the 1950–96 training period, historical relationships between the SST anomaly and the 700-hPa anomaly patterns are modeled by the CCA. In making a prediction for a given 3-month predictand period, the current year's SSTs are projected onto these generic relationships, where each relationship corresponds to one of the several CCA modes (e.g., one representing ENSO, one representing a decadal trend, etc.) whose sum is used to approximate the climate state. The SSTs input to the CCA procedure are described in Barnston (1994) and have not been improved with the EOF-reconstruction method that has been applied to the NCEP SST analyses used to force the NCEP GCM. For applications like the CCA, using this older SST data does not seem to have a significant impact on simulations.

While the nature of the predicted 700-hPa patterns is determined by the projection process as described above, the amplitude of the patterns is governed by both the strength of the current predictor patterns as well as the reliability, or estimated a priori skill, of the historical relationships. Given equal strengths of the predictor patterns, lower reliability results in greater damping of the predicted pattern amplitude—in similar fashion to any linear regression—such that squared prediction errors are minimized over the training period. In this study, the expected skill of the predictions is estimated using cross validation, where each year is withheld in turn from the training data and used as the prediction target, and the climatological means and variances are defined on the basis of only the training data. A temporal correlation or rms error, computed using the resulting predictions and the corresponding

observations, would then be an estimate of the skill. In this study, the climatological parameters are defined on the basis of the periods covered by the numerical models rather than on the CCA's 1950–96 period without one target year. In producing the CCA simulations, however, the year being predicted is nonetheless omitted from the training sample to prevent it from contributing to the generic relationships and giving the CCA an advantage not available in real-time forecasting. This CCA procedure is somewhat different from the operational CCA forecast procedure at the Climate Prediction Center (CPC), which uses four 3-month seasons of SSTs as predictors and also uses 700-hPa height as a predictor.

It is possible to argue that, when applied in simulations, this procedure gives the CCA an advantage over the numerical models. SSTs in midlatitudes, at least in some regions, are to some extent directly forced by the midlatitude atmospheric flow (Lau 1997). Since the CCA simply associates two fields without assuming that one forces the other, it is possible for the CCA to get information about the midlatitude atmospheric flow that produced the *predictor* SSTs in midlatitudes. The GCM experiments, on the other hand, are predicated on the notion that the SSTs force the atmosphere. The midlatitude flow in the GCMs in these regions is a combination of response to remote (primarily tropical) SST forcing, stochastic midlatitude dynamics, and perhaps some local SST forcing on longer timescales. Some information from SSTs in regions where the atmosphere partly forces the ocean that could be used by the CCA is therefore unavailable to the numerical models. This potential advantage for CCA would disappear in true forecasts since the observed SSTs contemporaneous with the atmospheric fields being predicted would not be available. This issue is discussed further in section 7.

d. Measures of simulation skill

Simulations of seasonal (three month) means of 700-hPa height for the PNA region (20°–80°N, 180°–60°W) are evaluated in this study. The 700-hPa height was used traditionally in the production of seasonal forecasts at NCEP's CPC and gives a reasonable measure of the midtropospheric circulation (the use of other levels in the troposphere had no significant quantitative impact on the results). The PNA region is chosen in an attempt to focus on the signal available from external SST forcing, the tropical Pacific SST in particular. This is consistent with previous studies like that of Kumar et al. (1996), who studied similar questions for two versions of the NCEP GCM. Similar regions have also been used in a number of other studies that attempt to evaluate the capabilities of prediction models for seasonal extratropical forecasts (Chen and van den Dool 1997; Shukla 1998). Simulations have been evaluated for 12 3-month means per year (January–March, JFM; February–April, FMA; etc.). All measures of skill for the numerical models in sections 3, 4, and 6 are evaluated for the ensemble mean.

The spatial anomaly correlation

$$\mathrm{AC} = \frac{\sum_i A_i (F_i - C_i)(V_i - C_i)}{\left[\sum_i A_i (F_i - C_i)^2 \sum_i A_i (V_i - C_i)^2\right]^{1/2}} \quad (1)$$

is the primary measure of skill in this study. Here C_i, F_i, and V_i are the climate, simulated, and verification height values at grid point i, respectively, which is associated with a surface area A_i and the sums extend over all grid points in a given geographic region. The climate field, C_i, is an average of NCEP reanalysis data over the period 1979–95. Because of the mass loss problem mentioned above, C_i, F_i, and V_i are all departures of the 700-hPa field from the corresponding global mean.

Careful cross validation was used throughout the simulation evaluation process so that the year being evaluated was never used in the definition of the climatology (see the appendix for a brief list of other cautions that must be observed when using AC to evaluate the relative quality of simulations/forecasts). Measures similar to the AC have been used historically for evaluation of seasonal forecasts (Kumar 1996), although there are many deficiencies of this measure that are not frequently discussed in the literature (Deque and Royer 1992). The root-mean-square error

$$\mathrm{rms} = \left[\frac{\sum_i A_i (F_i - V_i)^2}{\sum_i A_i}\right]^{1/2}, \quad (2)$$

where the sum extends over all grid points in a given region, is also discussed briefly to give some feeling for the dependence of the results on the verification measure used to evaluate simulations. For both the AC

and rms, both model and observed data are the average of instantaneous 0000 and 1200 UTC fields. Extreme caution should be exercised when extrapolating the results presented here for AC and rms to other measures of simulation quality.

3. Comparison of PNA region anomaly correlations

Figure 1 shows a comparison of the anomaly correlation of the three models averaged over the 9 yr (8 yr for the NDJ and DJF seasons) as a function of the 3-month season. The CCA has considerably higher mean AC in fall, winter, and early spring. In the summer, the CCA AC is at its smallest and is roughly comparable to the AC of the GFDL model. The NCEP model has very small ACs compared to the other models, due partly to large systematic model errors, which will be addressed in the next section. The CCA is clearly superior overall with an average AC over all years and seasons of 0.39 compared to 0.24 and 0.07 for the GFDL and NCEP models, respectively. If one squares these ACs to get a measure of explained variance, the gap between the CCA and the numerical models grows even larger although even the CCA explained variance of 15% is small.

Figure 2 displays the time series of ACs for each of the models over the 106-month period from DJF 1980 through OND 1988. There is a great deal of noise and scatter in the plot; however, some conclusions can be made. First, the CCA appears to have higher ACs throughout most of the simulation period. Second, each of the three models has times during which its simulations are quite different from the observations, for instance, NCEP in fall 1983, GFDL in spring 1981, and CCA in summer 1985. However, the CCA generally has fewer simulations with large negative ACs. Third, there are times when all three models simultaneously have relatively high ACs. The most obvious is during the winter of 1982–83, but there are less notable periods late in the winters of 1981/82 and 1985/86. There are no periods of high ACs among all three models during seasons other than winter or early spring. There are also a more limited number of cases in which all three models have low AC, in particular in the late fall of 1985. Unless the three models share a common systematic error of some sort, periods when all three have ACs less than 0 should be regarded as the result of chance; note that the failure of all three models to incorporate the impacts of a volcanic event

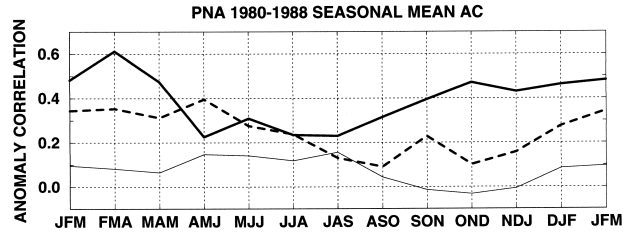

FIG. 1. PNA region anomaly correlation as a function of 3-month season averaged over years 1980–88 for the CCA (thick), NCEP model (thin), and GFDL model (thick dashed) simulations.

found in the real atmosphere would be one type of common systematic error here. These negative AC episodes are a good rough significance test for the positive AC events. The positive AC events must be considerably greater in magnitude and/or in frequency than the negative AC events in order to be judged to be more than the result of random chance.

Occasions when a model has unusually high skill have been discussed a great deal in the context of the model response to extreme ENSO events like 1982–83 (or the most recent 1997–98 event). It is important to emphasize that it is only meaningful to discuss the skill of exceptionally successful forecasts if one can make an a priori identification of when such good forecasts will occur (see section 5c). If one cannot make such a priori identifications, then discussions of the mean skill or of the complete historical distribution of skill are more relevant.

It is likely that newer GCMs in current use or under development at both NCEP and GFDL are better in many respects than the older GCMs investigated here, although it is unlikely that the newer GCMs have improved enough to have significant qualitative impact on the results of this or later sections. This claim will be investigated as long integrations and seasonal forecasts from newer model versions become available. Results from model comparison projects like AMIP (Gates 1992) also indicate that it is unlikely that other GCMs currently in use are sufficiently superior

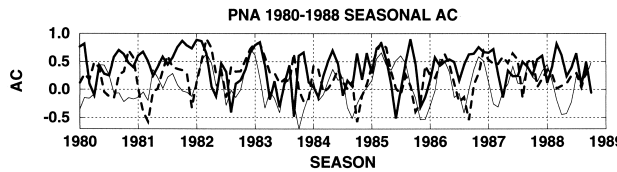

FIG. 2. PNA region anomaly correlation as a function of 3-month season for the CCA (thick), NCEP model (thin), and GFDL model (thick dashed) simulations; seasons run from JFM 1980 through OND 1988.

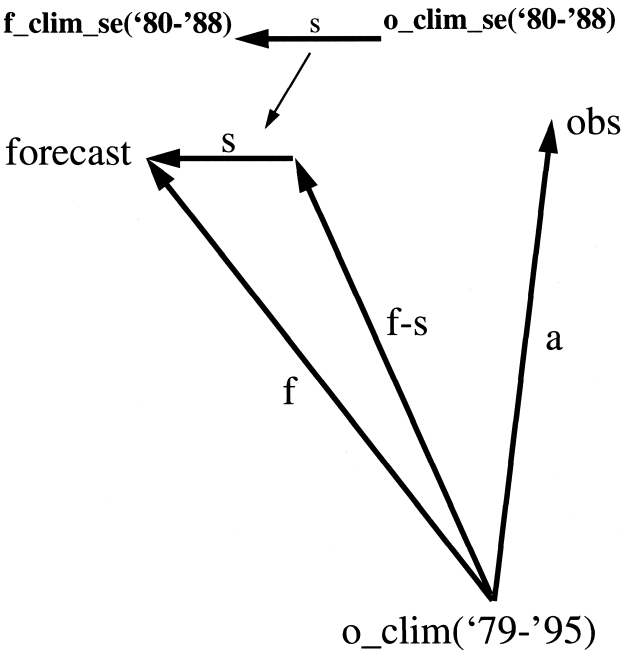

FIG. 3. Schematic diagramming the computation of AC and a systematic error correction. The AC defined in (1) is the cosine of the angle between the simulated anomaly vector, **f,** and the observed anomaly vector, **a,** in the area-weighted verification phase space. The anomalies are from an observed climatology, o_clim, the 1979–95 mean of the NCEP reanalysis. The systematic error correction, **s,** is the difference between the model climatology, f_clim_se, and an observed climatology, o_clim_se, both defined for the period 1980–88. The systematic error corrected AC is the cosine of the angle between **f** – **s** and **a.**

to the GCMs examined here to qualitatively impact the results of the comparison to the statistical model. However, continued improvements to GCMs can be expected to improve their performance compared to the CCA results.

It is appropriate to comment on the error bounds that are associated with the plots in Figs. 1 and 2 and the figures in later sections. The AC computations in this study are for the PNA region, which covers a substantial fraction of the Northern Hemisphere. Following van den Dool and Chervin (1986), the number of statistical degrees of freedom for this region can be estimated to be substantially greater than 10; to be conservative, $N = 10$ degrees of freedom are assumed here. For uncorrelated fields, the correlation of a pair of random samples has a normal distribution with mean 0 and standard deviation $s = 1/(N-2)^{1/2}$. For single-season correlations like those in Fig. 2, the expected error is less than $1/(8)^{1/2}$. In Fig. 1, where $T = 9$ or 10 independent seasonal correlations have been averaged, the expected error is reduced by an additional factor of $T^{1/2}$ resulting in expected errors that are bounded above by 0.12. If the fields are drawn from a population that has nonzero expected correlation, the details of this computation are more complex but the results are essentially unchanged. Differences among the skills for the CCA and the GFDL and NCEP GCMs are thus well beyond expected statistical uncertainty.

4. Impacts of systematic error correction

The previous section evaluated the ACs of the numerical models with no a posteriori correction of the models' systematic errors. It could be argued that this is the only fair way to compare the capabilities of numerical models and statistical models. Nevertheless, it has been traditional to verify numerical model output after the application of some simple statistical corrections to account for the most obvious systematic errors of the models (Chen and van den Dool 1997); often, verification studies in the literature do not even discuss the details of their systematic error corrections (Kumar et al. 1996). In a sense, the application of a posteriori systematic error corrections in the validation of numerical models results in a hybrid statistical–dynamical simulation system. Although this could be interpreted as giving an unfair advantage to numerical models in a comparison with purely statistical models, this section will proceed to apply simple systematic error correction to the numerical model results.

A variety of ways to apply systematic error correction when computing ACs has been used (Miyakoda et al. 1986; Deque and Royer 1992). The AC defined in (1) can be interpreted as the cosine of the angle between the simulated anomaly and the observed anomaly vectors in the area-weighted verification phase space (this is $< \mathbf{f}, \mathbf{a} >$ in Fig. 3, where $< \mathbf{x}, \mathbf{y} >$ represents the cosine of the angle between the vectors **x** and **y**). The anomalies are from an observed climatology that is the 1979–95 mean of the NCEP reanalysis. The systematic error correction traditionally applied by the CPC at NCEP is used here, changing AC to $< \mathbf{f} - \mathbf{s}, \mathbf{a} >$ in Fig. 3. The systematic error correction, *s,* is simply the difference between the model climatology, f_clim_se, and an observed climatology, o_clim_se, both defined over a systematic error correction period that may be different from both the simulation period and the period used to define o_clim. This systematic error correction has been applied to

the forecasts from both numerical models with the systematic error correction period 1980–88 used to compute f_clim_se and o_clim_se. In all cases, systematic error corrections were applied in a cross-validated framework so that no information about the particular year being validated is used in the systematic error correction for that year.

Figure 4 shows a comparison of the seasonal mean ACs for the NCEP and GFDL models with systematic error correction and the unmodified CCA. The overall mean values of the ACs have been increased slightly to 0.27 for the GFDL and significantly to 0.26 for the NCEP models, values still considerably smaller than the CCA's 0.39. The ACs for the two numerical models are quite similar, both in the mean and as a function of season, with the systematic error correction applied. During the spring and early summer, the numerical model ACs are roughly equivalent to those for the CCA. However, the CCA ACs continue to be considerably larger than those for the numerical models during the fall and especially the winter. A traditional simple systematic error correction is insufficient to make the numerical model results competitive with the CCA. Possibly, the application of more sophisticated statistical corrections to the numerical model results could lead to additional increases in AC, but it is not clear how much information would be coming from the numerical models if this were done. A systematic error correction has not been applied to the CCA. The CCA systematic error is zero over its full training period by definition, but it could be nonzero over the 1980–88 period being used in this study.

The values of ACs are notoriously sensitive to seemingly small details of the verification procedure, especially when the magnitude of the AC is low (< 0.5). Details like the definition of the observed climate period, the exact region of the verification, the period over which systematic error corrections are computed, etc., can have significant impacts on the ACs obtained. Sensitivity to these and many other parameters of the verification were also found in this study, which points out that it is absolutely essential to use a consistent verification algorithm when comparing simulations or forecasts produced by different modeling systems (see the appendix). In this study, this was accomplished by verifying all three simulations with the same software; attempts to coordinate separate consistent verifications at GFDL and NCEP were difficult and, in the end, judged to be impractical. Experiences like this suggest that some effort to provide standardized verification software could be extremely useful in facilitating research that requires comparison of the abilities of different forecast methods.

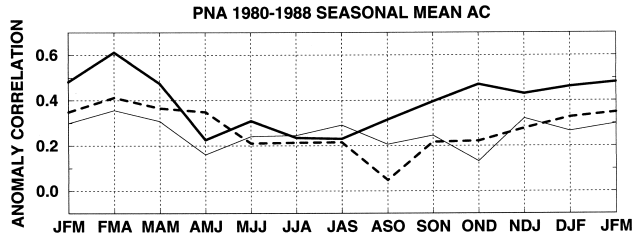

FIG. 4. PNA region anomaly correlation as a function of 3-month season averaged over years 1980–88 for the CCA (thick), systematic error corrected NCEP model (thin), and systematic error corrected GFDL model (thick dashed) simulations.

5. Use of ensemble distributions

The numerical model results used here consist of ensembles of simulations. In addition to evaluating the AC of the ensemble, one can also attempt to extract useful information from the distribution of the ensemble. First, the ensembles are used to compute an approximate upper bound on the simulation ACs one could expect if the numerical models being used were able to exactly reproduce the behavior of the atmosphere given an initial condition and bounding SSTs. Second, the linear correlation of the ensemble spread with the simulation skill is examined to see if the skill of the simulations can be predicted a priori.

a. Perfect model predictability

Following Chen and van den Dool (1997), one can use an ensemble of simulations to place an approximate upper bound on the expected value of the AC by making a perfect model assumption; they refer to this as ensemble mean predictability, here it is referred to as perfect model predictability. Suppose for a moment that the numerical model is perfect in that, if given appropriate initial conditions (for the atmosphere, land surface, etc.) and observed SSTs, it can exactly reproduce the response of the real atmosphere. If this were the case, each member of the ensembles of long simulations used here can be regarded as being equally likely to represent the response of the atmosphere to the observed SST forcing. One can compute seasonal mean ACs as done in sections 3 and 4 by replacing the time series of observations with a single member of the ensemble and continuing to treat the remaining ensemble members as simulations. The climatology for the AC calculation is defined as the 1980–88 sea-

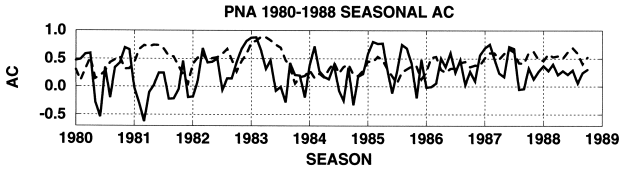

FIG. 5. Systematic error corrected GFDL model anomaly correlation (solid) and anomaly correlation of perfect model predictability for GFDL model (dashed) for 3-month season from JFM 1980 through OND 1988.

sonal mean of the "observed" ensemble member. The perfect model predictability AC and rms can then be evaluated in exactly the same way as for the original experiments. A more robust estimate of the perfect model skill can be obtained by computing the skill with each individual ensemble member being treated as the observations in turn and averaging the resulting values of AC and rms. Note that these perfect model skill bounds are for an ensemble size one smaller than that used in the results of the previous section, but this difference should have a relatively small quantitative impact (Murphy 1989). It is unreasonable to expect an imperfect model to produce simulations with higher ACs than those produced in this perfect model context, so this perfect model predictability AC can be taken as an approximate upper bound on the AC that one can expect to attain with a given GCM.

A systematic error correction can also be applied in this perfect model context as in section 4. By definition, the perfect model assumption implies that there is no systematic error in the model. Since careful cross validation is used, the systematic error correction procedure's only effect is to adjust the ensemble mean for a particular year toward the mean response of the single observed ensemble member for all other years. The result is that "systematic error correction" actually reduces the mean ACs for all seasons in the perfect model case; this was discussed in detail in Barnston and van den Dool (1993). For this reason, all *predictability* results discussed here will be without the application of the systematic error correction.

Figure 5 shows the complete time series of the AC for the GFDL ensemble mean simulations (with systematic error correction) and the perfect model predictability. The mean value of the perfect model predictability is 0.44, considerably higher than the 0.27 for the actual simulation ACs. The perfect model predictability is highest during spring 1981 and during early 1983 at which time the predictability AC becomes as large as 0.9 (this predictability maximum lags the maximum of prediction AC by a few months). This demonstrates that the impact of the SSTs on the model at this time is unusually strong, restricting the ensemble response to a very limited portion of the climatological distribution.

Not surprisingly, the model simulation AC (solid line in Fig. 5) is generally significantly less than the perfect model predictability values. There are a few times when the simulation AC is higher than the perfect model predictability values, most notably during two periods in 1985. A priori, one cannot hope for the model to produce ACs higher than the predictability values, so events like this have to be regarded as a result of chance. In other words, in this particular instance, the real atmosphere's response to SST forcing happened to look more like the ensemble mean than a randomly chosen member of the ensemble.

b. Comparison of perfect model ACs and CCA

Figure 6 presents a comparison of the perfect model predictability ACs from the GFDL and NCEP ensembles (cf. Fig. 13 of Chen and van den Dool, which shows NCEP results for a longer period) and the simulation ACs from the CCA. The GFDL and NCEP predictability means are roughly equal at 0.44 and 0.42, respectively; the models also have similar seasonality (Fig. 6), which gives some credibility to this predictability estimate. The CCA simulation ACs are slightly lower in the mean, which is encouraging in that it suggests there is potential for improvements in the models' mean response, which could lead to GCMs that are better than CCA. At least in a perfect model context, these GCMs are capable of producing simulations that are better than those provided by a statistical model. However, Fig. 6 reveals that during the autumn and winter, the CCA ACs are somewhat higher than the predictability ACs from the two numerical models. The CCA AC drops during the summer months, but this behavior is not seen in the NCEP

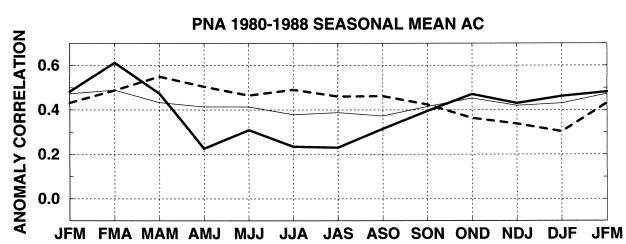

FIG. 6. PNA region anomaly correlation as a function of 3-month season averaged over years 1980–88 for the CCA (thick), NCEP perfect model predictability (thin), and GFDL perfect model predictability (thick dashed).

perfect model predictability ACs, which are roughly constant throughout the seasonal cycle, or in the GFDL model, which has a wintertime minimum in predictability AC.

On the surface, this suggests that it may be difficult to improve these numerical models sufficiently to compete with statistical models in the extratropical winter. However, it is also possible that model errors lead these numerical models to respond less strongly to imposed SST forcing than does the real atmosphere. In this case, improvements to the models could lead to higher values for the predictability AC, as well as to higher values for the simulation ACs. The summer results are also encouraging for numerical modelers. If the perfect model response is reasonable, it is possible that improvements to the models' mean response could result in simulations with ACs higher than those provided by CCA. Still, Fig. 6 should act as a sobering result for numerical modelers since it suggests that the CCA is doing very well indeed, even compared to one measure, independently confirmed in two numerical models, of the expected upper bound of numerical model skill.

Of course, there is no evidence presented here to suggest that the numerical models are constrained to the same degree as the real atmosphere by SST forcing. The models might have too much or too little ensemble spread compared to the appropriate "real world" uncertainty associated with specified SSTs, causing our perfect model predictability estimates to be too low or too high, respectively. This question could be addressed, given a sufficiently long and accurate observational record, by using methods such as those in Anderson (1996).

c. Predicting model skill

The most common use of information about ensemble spread has been to attempt a priori predictions of the skill of forecasts (simulations) (Barker 1991). In general, the quality of the skill predictions is evaluated using the linear correlation between the spread and skill.

The perfect model ACs discussed in the previous section can be regarded as a natural measure of the degree of ensemble spread for the purposes of predicting the simulation AC of the ensemble mean. There are other ways that one could choose to measure the ensemble spread (Barkmeijer et al. 1993), but these are not explored further here. Table 1 shows the linear correlations of the perfect model AC with both the raw simulation ACs and the systematic error corrected ACs for the NCEP and GFDL models. These correlations are generally very small, with the largest being 0.35 for the systematic–error-corrected NCEP model (a correlation of 0.39 would be significant at the 95% level if one allows for two degrees of freedom per year). Despite these poor correlations, it is possible that there is some a priori information about simulation AC available in cases of particularly strong signal. For instance, the start of 1983 has the largest values of perfect model AC for both the GFDL and NCEP models (although these maxima occur a few months after the highest simulation AC), indicating that the models are very confident about the response of the PNA region to the SST forcing during this period. Likewise, the model simulation ACs are by far the highest at approximately the same time indicating that the ensemble mean of the PNA simulations are close to what was observed. The 1997–98 ENSO event should offer an additional realization of the atmospheric response to extremely strong SST forcing. Future experiments should be able to better address the possibility that GCMs can predict high skill in such cases.

TABLE 1. Correlations between seasonal mean ensemble perfect model anomaly correlation and ensemble anomaly correlation for the GFDL and NCEP models.

	GFDL	NCEP
No systematic error correction	0.28	0.09
Systematic error corrected	0.11	0.35

There are also cases where the perfect model AC is relatively high while the model simulation ACs are low, for instance, much of 1981 or summer of 1983 in the GFDL model (Fig. 5). These cases do not necessarily indicate that the perfect model predictability ACs are poor predictors of the model simulation ACs since the former can be viewed as an upper bound on the *expected value* of the simulation AC. It would be possible to have most members of the ensemble simulation relatively close to the ensemble mean with a few outliers. In such cases, the perfect model framework suggests that most of the time one would expect simulations with high ACs, but occasionally one might by chance get a simulation with markedly lower ACs. This is not the case for the GFDL model in the two instances noted above since all members of the en-

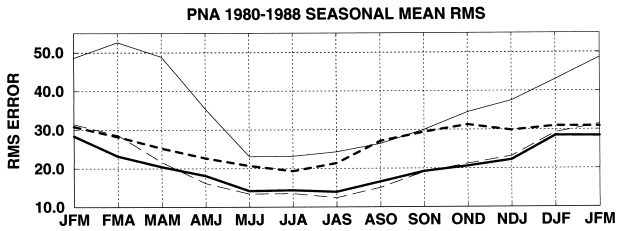

FIG. 7. PNA region rms (m) error as a function of 3-month season averaged over years 1980–88 for the CCA (thick), NCEP model (thin), GFDL model (thick dashed), and climatology (thin dashed). No systematic error corrections have been performed for the two numerical models.

semble have ACs roughly equal to that of the ensemble mean.

The amplitude of the simulation field from the CCA could also be used to produce a priori predictions of simulation AC. The simulation amplitude is computed as the square root of the sum of squares of the standardized anomalies over all grid points of the predicted 700-hPa field. While the year-round correlation is only weakly positive (near 0.1), it is better during the cold half of the year when skills are highest. The correlation is best for the first few seasons of 1983; however, there are other cold season strong signal cases in which the relationship does not hold well.

6. Rms results

To give some indication of the sensitivity of the results of the previous sections to the choice of the verification measure, a brief summary of results for the rms error is presented. Figure 7 shows the seasonal mean rms for the CCA, the NCEP and GFDL model ensemble means without systematic error correction, and for climatology. In the overall mean, the CCA has the smallest rms error (19.9 m) although it is only negligibly better than the rms of a climatological forecast (20.3 m). The GFDL model has larger rms error than the CCA for all seasons and an overall mean of 26.4 m while the NCEP model has much greater mean rms error (35.6 m), most of it resulting from extremely large seasonal mean rms error in the late winter and spring. In the rms error framework, CCA continues to be clearly superior to the numerical models.

Figure 8 displays the rms error for the CCA (same as Fig. 7 but with axes stretched) and for the GFDL and NCEP models with a systematic error correction identical to that performed for AC in section 4. The numerical models now have greatly reduced and nearly identical rms errors (overall means are 21.06 m for NCEP and 21.36 m for GFDL) that are only slightly larger than those for the CCA.

If one computes the perfect model predictability of section 5 in terms of rms, the GFDL model has an overall mean of 21.0 m and the NCEP model has 23.8 m. It is interesting to note that the perfect model predictability rms for the NCEP model is larger than the systematic error corrected rms in the mean. This is an indication that the spread of the models is too large. If the spread of the ensemble were consistent with the perfect model hypothesis in section 5, the perfect model predictability rms error would be a firm lower bound on the rms.

For the rms error measure, it is very easy to get mean rms errors that are roughly equivalent to those produced by the climatological forecast, and most likely very difficult to get rms errors significantly less than the climatological rms. This is consistent with the known relationship between the AC and the rms error for undamped forecasts in which an AC of approximately 0.5 is required in order to outperform climatology in terms of rms error (Roads 1986; Barnston 1992). The CCA model applied here is damped, but with respect to the period 1951–96, so the correspondence between an AC of 0.5 and the rms error of climatology may not hold exactly. If damped for the appropriate period, the rms error of the CCA should be better than that of climatology for any AC greater than zero.

7. Discussion

The abilities of the dynamical GFDL and NCEP GCMs and NCEP's empirical CCA to *simulate* the extratropical tropospheric seasonal flow over the PNA region (see also Barnett et al. 1997) when provided with observed global SSTs have been compared for the period 1980–88. The CCA simulations were found to have higher skill overall when using either the anomaly correlation or the rms error as the verification metric. Even when simple statistical corrections were applied to reduce the systematic errors of the GCM simulations, the simulation skill of the CCA was still superior. These results are consistent with those from Kumar et al. (1996), who compared PNA region simulations from two different versions of the NCEP GCM using the AC with some variant of NCEP's systematic error correction. They also briefly compared these simulations to a simple statistical model for the

winter season but felt they had insufficient information to make conclusions about the relative merit of the numerical and statistical models.

Although the CCA is more skillful than the GCMs, the skill of the CCA itself is relatively low despite periods (early 1983) with skill substantially higher than the mean. It is possible that even low levels of skill could be useful on seasonal timescales, but it is also important to note that skills for true predictions will almost certainly be lower than those for the simulations in the mean.

One could hope to increase the utility of these low skill simulations by making an a priori identification of cases that are expected to have unusually good (or poor) skill. The ensemble spread from the GCM simulations is one candidate for a predictor of the simulation skill. Although the linear correlation of a measure of spread and the anomaly correlation was small, there is still some hope that times of exceptionally small spread are concurrent with times of high simulation anomaly correlation. This might allow the a priori identification (i.e., before validating the simulation skill) of simulations that are expected to be more useful than normal. In the models examined here, all cases of high simulation anomaly correlation were associated with relatively small spread, but not all cases of small spread were associated with high simulation anomaly correlation.

Information about the GCM ensembles was used to calculate the perfect model predictability, a rough estimate of an upper bound on the simulation skill that could be expected in a perfect model context. These upper bounds on expected skill were found to be higher than the simulation skill of the GCMs and the summertime CCA skill; however, the GCM upper bounds on skill were somewhat lower than CCA skill during the winter.

All the results presented so far are simulations in which the models are given the SSTs from the verification period. Because of the primitive state of coupled GCMs, it is difficult to determine exactly how much of a reduction in skill can be expected when one switches from simulations for which contemporaneous SSTs are provided to forecasts in which only SSTs preceding the verification period are provided [Livezey et al. (1996); Barnston et al. (1994) compared the capabilities of many methods for predicting tropical Pacific SSTs on seasonal timescales]. However, this question can be answered for the CCA. Figure 9 compares the ACs for CCA simulations and CCA forecasts in which the predictors are SSTs from the

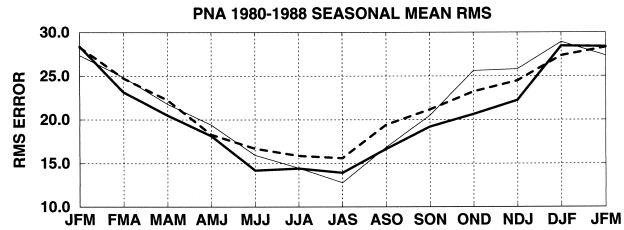

FIG. 8. PNA region rms error (m) as a function of 3-month season averaged over years 1980–88 for the CCA (thick), systematic error corrected NCEP model (thin), and systematic error corrected GFDL model (thick dashed).

3-month period preceding the 3-month forecast period. There is a reduction in AC for all seasons with winter ACs dropping approximately 5% while summer ACs become essentially zero. It seems reasonable to expect qualitatively similar reductions in AC with numerical models when the switch is made from simulation to forecast. As discussed in section 2, it is possible to argue that the CCA simulation is given an advantage over the numerical models through the use of observed values of the midlatitude SSTs. Unless the midlatitude oceans are strongly forced by some deterministic component of the midlatitude flow, this advantage is probably slight. If the CCA is gaining any advantage from the use of observed midlatitude SSTs, the reduction in numerical model skill from simulation to forecast might be somewhat less than for the CCA. In the perfect model comparison, however, CCA is penalized since all imperfections in the observational data lower the ACs for the CCA but have no impact on the perfect model ACs. Reanalysis, in the broadest sense, can only improve CCA or other statistical control forecasts.

Despite some pessimistic conclusions, there are still a number of reasons to hope that significantly higher skills than those found here could be obtained for extratropical seasonal forecasts. The experiments

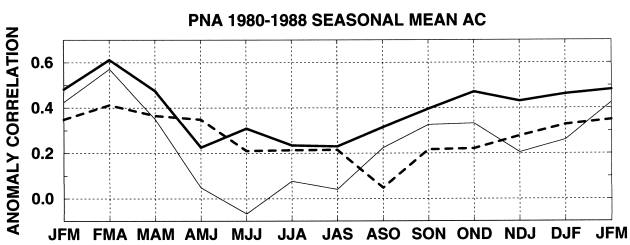

FIG. 9. PNA region anomaly correlation as a function of 3-month season averaged over years 1980–88 for the CCA simulation (thick), CCA forecast (thin), and systematic error corrected GFDL model simulation (thick dashed).

here ignore observed atmospheric initial conditions, and there is some evidence that these can be important even for seasonal timescale predictions. Information from the observed state of low-frequency atmospheric phenomena, such as the stratospheric quasi-biennial oscillation (QBO) is one potential mechanism by which atmospheric initial conditions could impact seasonal forecasts. The models used here also ignore information about the observed land surface conditions and, in the case of the GCMs, use relatively simple and unrealistic land surface parameterizations. Since the land surface may be associated with processes with timescales that are at least as long as a season, information about the state of the land surface could serve to increase extratropical forecast skill (Huang et al. 1996). There is also a slight possibility that SST-forced atmosphere-only GCM integrations may provide unrepresentative simulations, and that the use of a good fully coupled model could produce forecasts with skills higher than those produced in the observed SST simulations (Wittenberg and Anderson 1998).

There is another long-term advantage to using numerical models rather than statistical models for prediction. It is possible that statistical models will continue to be competitive with or even superior to numerical models for the seasonal prediction problem. However, it is difficult to gain a complete physical understanding of the climate system through statistical models, while it is likely that the development of increasingly realistic numerical models may lead to an increased physical understanding, if not better forecast skill.

Acknowledgments. The authors would like to thank the NCEP Coupled Model Project for providing the NCEP GCM integrations, and the Experimental Prediction Group at GFDL for providing the GFDL GCM integrations. J. Lanzante, A. Broccoli, J. Mahlman, and two anonymous reviewers provided many helpful suggestions that have been incorporated in this report.

Appendix: Checklist for anomaly correlation

The anomaly correlation is widely used in verifying atmospheric forecasts, but it is an extremely unstable statistic that can be quantitatively influenced by a number of factors. In computing ACs for the different models in this paper, great care was taken to ensure that all of the following areas were the same for all models. Even relatively slight inconsistencies in any of these can lead to large differences in the computed ACs.

The climatology:

1) Definition of climatology, specifically 2–7
2) Number of years used in estimate of climatology
3) Observing times (0000, 0600, 1200, 1800 UTC) or daily mean
4) Harmonic smoothing of climatology: how many harmonics to retain
5) Time period over which climatology is defined
6) Time interpolation of climatology if required
7) Year to be verified left out of climatology (cross-validation mode)

Technical aspects:

1) Definition of areal extent and situation of grid points
2) Areal mean in or out
3) Zonal mean in or out; zonal mean over a sector or full globe
4) Truncated or otherwise smoothed field, or full model resolution
5) Weighting with respect to latitude (equal-area grid issue)
6) Model's systematic error corrected or not (see 7 under climatology section above)
7) Standardized data or not
8) Summing over time, carrying each term, or averaging ACs

The model and the data:

1) If the model loses mass, what has been done about it (related to 2 under technical aspects section above)
2) Is it a single model forecast or an ensemble average (latter is more smooth)
3) Forecast data at 0000 UTC only, daily mean, 0000+1200 UTC averaged; use climatology accordingly
4) Need to interpolate the climatology to the forecast grid or vice versa

References

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

——, and W. F. Stern, 1996: Evaluating the potential predictive utility of ensemble forecasts. *J. Climate,* **9,** 260–269.

Barker, T. W., 1991: The relationship between spread and forecast error in extended-range forecast. *J. Climate,* **4,** 733–742.

Barkmeijer, J., P. Houtekamer, and X. Wang, 1993: Validation of a skill prediction method. *Tellus,* **45A,** 424–434.

Barnett, T. P., and R. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev.,* **115,** 1825–1850.

——, K. Arpe, L. Bengtsson, M. Ji, and A. Kumar, 1997: Potential predictability and AMIP implications of midlatitude climate variability in two general circulation models. *J. Climate,* **10,** 2321–2329.

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

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

——, and H. M. van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate,* **6,** 963–977.

——, and T. M. Smith, 1996: Specification and prediction of global surface temperature and precipitation from global SST using CCA. *J. Climate,* **9,** 2660–2697.

——, and Coauthors, 1994: Long-lead seasonal forecasts: Where do we stand? *Bull. Amer. Meteor. Soc.,* **75,** 2097–2114.

Chen, W. Y., and H. M. van den Dool, 1997: Atmospheric predictability of seasonal, annual, and decadal climate means and the role of the ENSO cycle: A model study. *J. Climate,* **10,** 1236–1254.

Deque, M., and J. F. Royer, 1992: The skill of extended-range extratropical winter dynamical forecasts. *J. Climate,* **5,** 1346–1356.

Gates, W., 1992: AMIP: The Atmospheric Model Intercomparison Project. *Bull. Amer. Meteor. Soc.,* **73,** 1962–1970.

Gordon, C. T., 1992: Comparison of 30-day integrations with and without cloud–radiation interaction. *Mon. Wea. Rev.,* **120,** 1244–1277.

——, and W. F. Stern, 1974: Spectral modelling at GFDL. Int. Symp. on Spectral Methods in Numerical Weather Prediction, Copenhagen, Denmark, WMO Rep. 7, 46–80.

——, and ——, 1982: A description of the GFDL global spectral model. *Mon. Wea. Rev.,* **110,** 625–644.

Graham, N. E., and T. P. Barnett, 1995: ENSO and ENSO-related predictability. Part II: Northern Hemisphere 700-mb height predictions based on a hybrid coupled ENSO model. *J. Climate,* **8,** 544–549.

——, J. Michaelsen, and T. P. Barnett, 1987a: An investigation of the El Nino–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 Nino–Southern Oscillation cycle with statistical models. 2. Model results. *J. Geophys. Res.,* **92,** 14 271–14 289.

Huang, J., H. M. van den Dool, and K. G. Georgakakos, 1996: Analysis of model-calculated soil moisture over the United States (1931–1993) and applications to long-range temperature forecasts. *J. Climate,* **9,** 1350–1362.

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

Kanamitsu, M., and Coauthors, 1991: Description of NMC global data assimilation and forecast system. *Wea. Forecasting,* **6,** 425–435.

Kumar, A., and M. Hoerling, 1998: Annual cycle of Pacific–North American predictability associated with different phases of ENSO. *J. Climate,* **11,** 3295–3308.

——, ——, M. Ji, A. Leetmaa, and P. Sardeshmukh, 1996: Assessing a GCMs suitability for making seasonal predictions. *J. Climate,* **9,** 115–129.

Lau, N.-C., 1997: Interactions between global SST anomalies and the midlatitude atmospheric circulation. *Bull. Amer. Meteor. Soc.,* **78,** 21–33.

Livezey, R. E., M. Masutani, and M. Ji, 1996: SST-forced seasonal simulation and prediction skill for versions of the NCEP/MRF model. *Bull. Amer. Meteor. Soc.,* **77,** 507–517.

Manabe, S., 1969: Climate and the ocean circulation. I. The atmospheric circulation and the hydrology of the earth's surface. *Mon. Wea. Rev.,* **97,** 739–774.

Miyakoda, K., J. Sirutis, and J. Ploshay, 1986: One month forecast experiments—without anomaly boundary forcings. *Mon. Wea. Rev.,* **114,** 2363–2401.

Murphy, J. M., 1989: The impact of ensemble forecasts on predictability. *Quart. J. Roy. Meteor. Soc.,* **114,** 463–493.

Roads, J. O., 1986: Forecasts of time averages with a numerical weather prediction model. *J. Atmos. Sci.,* **43,** 871–892.

Sarda, J., G. Plaut, C. Pires, and R. Vautard, 1996: Statistical and dynamical long-range atmospheric forecasts: Experimental comparison and hybridization. *Tellus,* **48A,** 518–537.

Shukla, J., 1998: Predictability in the midst of chaos: a scientific basis for climate forecasting. *Science,* **50,** 728–731.

Sirutis, J., and K. Miyakoda, 1990: Subgrid scale physics in 1-month forecasts. Part I: Experiment with four parameterization packages. *Mon. Wea. Rev.,* **118,** 1043–1064.

Slingo, A., and J. M. Slingo, 1991: Response of the National Center for Atmospheric Research Community Climate Model to improvements in the representation of clouds. *J. Geophys. Res.,* **96,** 15 341–15 357.

Stern, W., and R. Pierrehumbert, 1988: The impact of orographic gravity wave drag parameterization on extended range predictions with a GCM. Preprints, *Eighth Conf. on Numerical Weather Prediction,* Baltimore, MD, Amer. Meteor. Soc., 745–750.

——, and K. Miyakoda, 1995: The feasibility of seasonal forecasts inferred from multiple GCM simulations. *J. Climate,* **8,** 1071–1085.

Tiedtke, M., 1988: Parameterization of cumulus convection in large-scale models. *Physically-Based Modeling and Simulation of Climate and Climate Change,* M. Schlesinger, Ed., D. Reidel, 375–431.

van den Dool, H. M., and R. M. Chervin, 1986: A comparison of month-to-month persistence of anomalies in a general circulation model and in the earth's atmosphere. *J. Atmos. Sci.,* **43,** 1454–1466.

Wittenberg, A. T., and J. L. Anderson, 1998: Dynamical implications of forcing a model with a prescribed boundary. *Nonlinear Processes in Geophys.,* **5,** 167–179.