Boreal winter predictions with the GEOS-2 GCM: The role of boundary forcing and initial conditions

By YEHUI CHANG, SIEGFRIED D. SCHUBERT* and MAX J. SUAREZ

Goddard Laboratory for Atmospheres, USA

(Received 1 March 1999; revised 28 September 1999)

SUMMARY

Ensembles of atmospheric General Circulation Model (GCM) seasonal forecasts and long-term simulations are analysed to assess the controlling influences of boundary forcing and memory of the initial conditions. Both the forecasts and simulations are carried out with version 2 of the Goddard Earth Observing System (GEOS-2) GCM forced with observed sea surface temperatures (SSTs). While much of the focus is on the seasonal time-scale (January–March; 1981–95) and the Pacific North American (PNA) region, we also present results for other regions, shorter time-scales, and other known modes of variability in the northern hemisphere extratropics.

Forecasts of indices of some of the key large-scale modes of variability show that there is considerable variability in skill between different regions of the northern hemisphere. The eastern North Atlantic region has the poorest long-lead forecast skill, showing no skill beyond about 10 days. Skilful seasonal forecasts are primarily confined to the wave-like El Niño Southern Oscillation (ENSO) response emanating from the tropical Pacific. In the northern hemisphere, this is similar to the well-known PNA pattern. Memory of the initial conditions is the major factor leading to skilful extratropical forecasts of lead time less than one month, while boundary forcing is the dominant factor at the seasonal time-scale. Boundary forcing contributes to skilful forecasts at sub-seasonal time-scales only over the PNA region.

The GEOS-2 GCM produces average signal-to-noise ratios which are less than 1.0 everywhere in the extratropics, except for the subtropical Pacific where they approach 1.5. An assessment of the sampling distribution of the forecasts suggests the model’s ENSO response is very likely too weak. These results show some sensitivity to the uncertainties in the estimates of the SST forcing fields. In the North Pacific region, the sensitivity to SST forcing manifests itself primarily as changes in the variability of the PNA response, underscoring the need for an ensemble approach to the seasonal-prediction problem.

KEYWORDS: Ensemble predictions ENSO response Seasonal forecasts Signal-to-noise ratio

1. INTRODUCTION

The ability to produce skilful dynamical predictions in middle latitudes beyond the limits imposed by the approximately 2-day average doubling time for small errors (e.g. Lorenz 1982) has been an elusive goal of the extended-range weather and short-term climate prediction communities. For example, recent experience at an operational centre shows that the average skill of numerical weather forecasts is insignificant beyond 15 days (Van den Dool 1994). The physical basis for long-range (monthly and longer) forecasts with atmospheric general circulation models (AGCMs) rests primarily on the strength of the response of the atmosphere to the slowly varying components of the earth system: the so-called ‘boundary forcings’ including sea surface temperature (SST), soil moisture, sea ice and snow (e.g. Shukla 1984).

A number of early studies suggested that, in addition to the influence of the lower-boundary forcing, memory of certain slowly varying components of the atmospheric flow can on occasion contribute to skilful monthly-mean forecasts (e.g. Shukla 1981; Miyakoda et al. 1986; Palmer 1988, Park and Schubert 1997). This is tied to the notion of a link between predictability and the time-scale of atmospheric phenomena (e.g. Van den Dool and Saha 1990; Schubert et al. 1992), suggesting that improvements in the simulation of low-frequency phenomena such as blocking, planetary-scale waves, the Madden–Julian Oscillation (MJO), and the zonal index, can lead to periods of extended-range predictability. Extratropical connections with the MJO have, for example, been demonstrated in a number of studies (e.g. Weickmann et al. 1985; Higgins and Schubert

* Corresponding author: National Aeronautics and Space Administration, Goddard Space Flight Center, Greenbelt, Maryland 20771, USA.

2293
1996; Park et al. 1995). There is, however, a general consensus that boundary forcings are the primary source of atmospheric predictability at seasonal time-scales. In particular, the relatively strong impact of SST anomalies on middle-latitude variability on seasonal time-scales (versus a weak impact on monthly time-scales) has been demonstrated by Palmer (1987).

The early studies of monthly and seasonal atmospheric predictability were limited by inadequate models, the lack of high quality global boundary conditions, and limited computing capabilities which generally allowed only case-studies and/or regional impact studies. The pioneering work of Lau (1985) lead to a better understanding of the key role of SST anomalies on the interannual variability of the tropics as well as on the Pacific North American (PNA) region. This work set the stage for a host of multi-year simulation experiments, with various AGCMs, carried out at various research institutions, and recently co-ordinated as part of the Atmospheric Model Intercomparison Project (AMIP; Gates 1992). These experiments have done much to improve our understanding of the variability produced by SST anomalies (especially in the tropics) and the limitations imposed by model deficiencies. However, Lau’s and other experiments with multiple realizations of ‘AMIP’ integrations with the same model suggest a single long-term integration (as in AMIP) is insufficient to estimate the impact of boundary forcing on middle-latitude variability (e.g. Barnett et al. 1997). This is due to the sensitivity of the evolution of the atmosphere to initial conditions, and to the limited constraints imposed on the extratropical atmosphere by the slowly varying boundary forcing such as SST, sea ice, and soil moisture anomalies (e.g. Tribbia and Baumhefner 1988; Barnett et al. 1994; Brankovic et al. 1994; Kumar and Hoerling 1995; Stern and Miyakoda 1995).

The sensitivity to initial conditions leads naturally to one of the fundamental problems in seasonal forecasting: the determination of the signal-to-noise ratio. Here the ‘noise’ refers to the uncertainty in forecasts coming from our uncertainty in the initial conditions. This differs from the initial-condition uncertainty in short- and medium-range forecasts, in that at the seasonal time-scale these errors are near their saturation values. In fact, one of the goals of the Dynamical Seasonal Prediction (DSP; Shukla et al., personal communication) project is to re-examine the relative contributions of the memory of the initial conditions and boundary forcing in seasonal forecasts, to determine whether any skill exists in seasonal forecasts resulting from memory of the initial conditions. Thus, the DSP project seasonal integrations are natural companions to ensemble AMIP integrations in which, by design, the distribution functions of the noise are saturated.

Until very recently, operational long-range (monthly and longer) predictions have relied primarily on empirical techniques. Such prediction methods rely on statistical relationships established from an analysis of past observations. Numerous observational studies have, for example, linked short-term climate variations to the El Niño Southern Oscillation (ENSO). Ropelewski and Halpert (1987) and Halpert and Ropelewski (1992) used station observations to show that precipitation and surface temperature anomalies associated with ENSO occur throughout the tropics and over substantial regions of the extratropics including regions of North and South America, Australia, India and Africa. Further evidence exists for a biennial component to the Southern Oscillation (SO), such that the year following a particular SO event develops anomalies that tend to be of opposite sign (Kiladis and Diaz 1989; see also Lau and Sheu 1988).

The identification of such a remote response to ENSO is only part of a much larger effort to identify recurring low-frequency modes of atmospheric variability. For example, Wallace and Gutzler (1981) summarize some of the basic patterns, which
include the PNA pattern, the western Pacific pattern, the eastern, western and North Atlantic patterns, and the Eurasian pattern. An outstanding research problem is to determine the nature of these modes of variation, and to what extent they are forced or naturally occurring internal modes of variation. Of these patterns, the PNA pattern appears to be the most robust, though the detailed structures of all the patterns show considerable variation and depend on the analysis techniques, quantity, and averaging periods. For example, Livezey and Mo (1987) examined the correlations between SST and northern hemisphere circulation patterns, and found that several different patterns including the PNA pattern are likely to occur during strong ENSO years. One pattern (the tropical/northern hemisphere pattern) showed an especially strong link to anomalies in the central Pacific during ENSO years. The PNA pattern also showed a strong link to non-ENSO-year SST anomalies 25° west of the dateline and just north of the equator. Hamilton (1988) examined the sensitivity of the extratropical ‘PNA-like’ response to El Niño for the period 1899 to 1982. He found, in contrast to the results of Livezey and Mo (1987), that the response during the mature phase of ENSO is ‘fairly insensitive’ to the precise magnitude of the SST anomalies in the central and eastern equatorial Pacific. Instead, the extratropical circulation during the mature phase was found to be strongly correlated with the SST in the far western Pacific/Indonesian region, such that the strongest extratropical teleconnections occur when the far western Pacific is anomalously warm or, at least, not too cold.

The North Atlantic Oscillation (NAO) is a mode of variation apparently distinct from ENSO and fluctuates on interannual to decadal time-scales (e.g. Wallace and Gutzler 1981; Wallace et al. 1990; Kushnir 1994). This mode, which impacts the climate throughout much of the North Atlantic and Europe, is still poorly understood, especially the role of the oceans (e.g. Battisti et al. 1995; Hurrell 1995). While there is some evidence for an ENSO signal over Europe (e.g. Fraedrich 1994), the NAO appears to be the leading source of predictive skill on seasonal to interannual time-scales. For example, Johansson et al. (1998), employed a canonical correlation analysis for the time period 1955–93, showing seasonal forecasts of surface temperature over northern Europe with wintertime skill comparable to that found over the USA. Skill was, in some cases, evident out to about 15-months lead time. They found that the dominant source of skill over Europe is, however, associated not with ENSO (or more generally global SST), but with the NAO, reflected in this case primarily by the winter-to-winter persistence of the 700 hPa heights. Palmer and Sun (1985) examined the impact of western North Atlantic SST anomalies on 50-day forecasts employing a hemispheric GCM. They found the response to SST anomalies (enhanced by a factor of 1.5 compared with observational values) over the North Atlantic and Europe during winter to be consistent with observations. Analysis of the observations suggested, however, that the SST anomalies develop initially in response to anomalous atmospheric forcing. They suggest that positive air/sea interaction could help anomalies persist, particularly during late autumn and early winter.

The above studies indicate that much uncertainty exists about the sensitivity of the atmospheric response to the location and amplitude of SST anomalies. GCM studies by Palmer and Mansfield (1984, 1986) suggest that heating in the far western equatorial Pacific is key to the strong response over the North Pacific and North Atlantic. Geisler et al. (1985) used a GCM to show that the pattern of the middle-latitude response to equatorial Pacific heating anomalies resembles the PNA pattern, and is insensitive to the location of the heating, while the amplitude of the response decreases as the anomaly is moved further east of the dateline. Hoerling and Kumar (1997) show that the amplitude of the extratropical response is approximately linearly related to the
amplitude of the SST anomaly. They further show that the blended SST data (Reynolds and Marsico 1993) give too weak an anomaly. The blended SST are deficient due to excessive trimming of the data. In the SST analysis cycle 30% of the observations were discarded, resulting in unrealistically low (1.5 °C) temperatures in the eastern equatorial Pacific. They found that experiments using the optimal interpolation (OI) SST estimates (Reynolds and Smith 1994) gave a stronger more realistic response. We will return to this issue in analysing our own results.

One of the key difficulties in understanding the sensitivity of the extratropical response to SST forcing lies in the role of the background flow. Simmons et al. (1983), for example, argue that much of the extratropical low-frequency variability (e.g. the PNA pattern) arises from instabilities of the zonally asymmetric background flow, and the primary role of the tropics is to excite these modes. Branstator (1985) and Sardeshmukh and Hoskins (1988) suggest that the strong horizontal shear on the equatorward side of the east Asian jet may play a key role in producing a strong extratropical response to tropical heating (see also Held and Kang (1987), and Rasmusson and Mo (1993) regarding the importance of the stretching term). Ting and Sardeshmukh (1993) further quantified the role of the basic state by examining the response to idealized forcing in the tropics for two different basic states (one was from an analysis and the other from a GCM run), and for both shallow and deep heating profiles. For zonally asymmetric basic states (for the case of the observed mean state) the stronger response occurred for shallow heating. For the simulated mean state the results are less sensitive to the vertical distribution of heating. Experiments to test the sensitivity to longitudinal position of the tropical heating showed an equatorial quadrupole response that moves with the heating, as well as a normal-mode response that is fixed (for heating within the Pacific region) but with amplitude depending on position; the strongest response occurs with the heating in the eastern Pacific. Kumar et al. (1996) examined two slightly different versions of the National Centers for Environmental Prediction (NCEP) medium-range forecast model, differing only in the treatment of clouds. They found a much stronger response in one of the models, and argue that this occurs largely because of improved zonal mean climatologies in the tropics and subtropics, which enhances its middle-latitude sensitivity to tropical forcing. They suggest the key difference is in the stronger Hadley Cell which leads to an improved effective Rossby wave source. They show that the boreal winter stationary waves are similar, but the climatological precipitation distribution is different; there is more precipitation south of the equator over land (associated with the monsoons), yet the anomalies in precipitation are similar.

Another aspect of the response to anomalous SST forcing lies in the feedback with shorter time-scales, such as the synoptic-scale eddies. For example, Kok and Opsteegh (1985) suggest that anomalous transient eddy fluxes during ENSO may play an important role in the seasonal mean extratropical response. The diagnosis of the role of transients in the extratropical response to El Niño in a GCM carried out by Held et al. (1989) suggests a rather complicated interaction, in which the direct response to the heating is enhanced by feedback with the transient eddies, and the initial low-latitude forcing due to latent heating is itself modified as a result of changes in the Rossby wave penetration into the tropics.

The above results suggest several promising lines of research aimed at improving monthly and seasonal forecasts. The sensitivity of the amplitude of the extratropical response to the amplitude of the tropical SST anomalies suggests that we must look more closely at both potential bias in the SST fields and at the veracity of the mechanisms by which the model converts the SST anomalies into atmospheric heating anomalies. In addition, the sensitivity of the extratropical response to the background flow appears
to place more stringent requirements than currently achieved on the accuracy of the simulated climatological mean flow. In order to make progress in these areas we need to have a firm understanding of the basic responses achieved by the current generation of atmospheric models. It is important to know, for example, how much of the difference we see in the responses is the result of statistical sampling. As discussed earlier, the seasonal-forcasting problem is inherently probabilistic, requiring that we carry out a sufficiently large ensemble of forecasts to be able to clearly distinguish the forced signal from the noise due to the sensitivity to initial conditions.

The current study was carried out as part of the DSP project (Shukla et al. 2000), which has as its primary goal a comprehensive study of the predictability of seasonal anomalies in the extratropics. The project’s first objective is to assess the ability of atmospheric GCMs to produce skilful forecasts when they are forced by the observed oceanic boundary conditions (SST and sea ice). This should provide an upper bound on the skill of seasonal forecasts, and a natural baseline for further studies of the impact of uncertainties in the boundary forcing and on the requirements for SST forecasts. This is also consistent with the two-tiered approach to seasonal forecasting (Bengtsson et al. 1993) in which the SST and atmospheric predictions are carried out sequentially and with different models.

As mentioned earlier, perhaps the most important unresolved issue in the seasonal atmospheric-forecast problem in the extratropics concerns the relative contributions of the forced signal and the internal ‘noise’ to the extratropical seasonal variability. In fact, one of the main outcomes of the DSP project has been the realization that the amplifications of both the signal and the noise are highly model dependent, leading to a tremendous range of estimates of this ratio. The primary purpose of the current study is to assess the skill of seasonal forecasts employing version 2 of the Goddard Earth Observing System (GEOS-2) GCM. We examine both the robustness of the forced response and the potential impact of initial conditions. While much of the focus is on the seasonal time-scale and the PNA region (the PNA response), we also present results for other regions, shorter time-scales, and other known modes of variability. The role of the initial conditions is assessed by comparing the ensemble seasonal forecasts from observed initial conditions, with a set of companion ‘AMIP’ simulations for the same time periods; these runs were started from observed initial conditions in late 1978, thus they must have lost all memory of the initial conditions for the time periods of interest here.

Section 2 gives an overview of the model and forecast experiments. The results are presented in section 3. The first part of section 3 compares the impact of initial conditions with the forcing from the boundary conditions on forecast skill at various forecast ranges. This is followed by the analysis of the signal-to-noise ratio. The discussion and conclusions are given in section 4.

2. Model description and experiments

The current study employs an early version of the GEOS-2 GCM. The model represents a major upgrade from the GEOS-1 GCM documented in Takacs and Suarez (1996) and Molod et al. (1996). The dynamics (the ARIES-GEOS dynamical core) has been upgraded to a 4th-order accurate scheme (Suarez and Takacs 1995). The vertical extent and resolution has been substantially enhanced to 70 sigma levels extending to 0.01 mb. For this study the model was, however, truncated at 10 mb with 43 levels to reduce computational costs. The model includes a gravity-wave drag scheme (adapted from Zhou et al. 1996), and new short-wave and long-wave radiation schemes (Chou and Suarez 1994).
The penetrative convection originating in the boundary layer is parametrized using the Relaxed Arakawa–Schubert scheme (Moorthi and Suarez 1992), and includes a relative-humidity trigger to improve the diurnal cycle of precipitation over land. The model also includes a parametrization that models the evaporation of falling convective rain, as described in Sud and Molod (1988). Cloud properties are linked to the diagnosed cloud water. Non-precipitating clouds are included as in Slingo and Ritter (1985). The planetary boundary layer uses the second-order closure model of Helfand and Labraga (1988). For the experiments described here, GEOS-2 was integrated at a resolution of 2° latitude by 2.5° longitude. Since this version of GEOS-2 is run without a land surface model*, soil moisture is computed off-line based on a simple bucket model using monthly mean observed surface air temperature and precipitation (Schemm et al. 1992). The snow line and surface albedo are also prescribed but vary only with season. The SST is prescribed from observations, as described below.

The design of the forecast experiments is that agreed upon by the DSP project (Shukla et al. 2000) to allow a consistent comparison of the model results from the various participating organizations. The forecast experiments consist of an ensemble of nine runs for each winter season for the period 1980/81 through 1994/95. The nine members of the ensemble were generated by starting the forecasts 12 hours apart, centred on 15 December. Global SST and sea ice are prescribed. We note that, while the DSP experimental design calls for the use of the OI SST data (Reynolds and Smith 1994), our DSP runs were carried out with the Reynolds and Marsico (1993) blended data. As noted earlier, these data have for some periods unrealistically weak SST anomalies in the eastern equatorial Pacific and, as we shall see later, impact the strength of our model’s ENSO response. In addition to the DSP forecast experiments, we have also carried out an ensemble of AMIP-style integrations with the same model. In this case, six simulations were started with boundary conditions for September 1978 (with initial conditions from previous model runs separated by 10 days) and integrated for 16 years with prescribed SSTs and sea ice. These experiments employ the AMIP-II boundary conditions which incorporate the Reynolds and Smith (1994) SST data†. In all runs soil moisture is prescribed as described above. The initial conditions for the DSP runs are based on the GEOS-1 re-analysis (Schubert et al. 1993); however, for each year a re-assimilation is carried out starting one week prior to the earliest initial time of the ensemble. This short re-analysis was performed with the model version used throughout this study.

3. RESULTS

We begin by showing in section 3(a) some basic forecast-skill results based on anomaly correlations. These results are used to establish the time-scales and regions which show significant skill, and to determine the forecast range at which the initial conditions appear to play a substantial role. They include a comparison with the AMIP integrations described above. This is followed in section 3(b) by an assessment and analysis of the signal-to-noise ratios of the seasonal-mean forecasts and of the AMIP simulations, with a special focus on the PNA region.

* The final version of GEOS-2 does include a land surface model, but this was not available at the start of the DSP project.
† After submitting this study, it was discovered that an error in our interpolation scheme for the SST fields resulted in a one grid point westward shift in the AMIP SST fields. This shift was confined to the eastern hemisphere. We have since rerun the first three years of the AMIP ensemble with the corrected SSTs, and found only minor differences in the results from the original runs.
Figure 1. Forecast skill computed as an anomaly correlation of the 500 mb height for the northern hemisphere (20°N-80°N). Correlations are computed for (a) zonal wave numbers 0–70; (b) zonal wave numbers 0–3. The results are based on 6-hourly data for all Decembers (9 × 15) for the period 1980/81 to 1994/95. The thick lines show the average of the correlations between the individual forecasts and the GEOS-1 re-analysis (see text). The forecast and verification anomalies are computed with respect to the 15-year mean forecast and the 15-year mean re-analysis climatology, respectively. The thin lines are the ‘perfect model’ results following Lorenz (1982). The different curves represent the correlations between all model forecasts with initial conditions separated by 12 hours, 24 hours, 36 hours, etc.

(a) Anomaly correlations and the role of the initial conditions

As noted in the introduction, the average skill of numerical weather forecasts using current operational models is insignificant beyond about 15 days (Van den Dool 1994). We show in Fig. 1(a) (thick line) that GEOS-2 has forecast skill similar to the operational models, with the 15-year average correlations for the northern hemisphere during December dropping below 0.6 after about six days, and below 0.2 after about 11 days. There is year-to-year variability in the correlations. For example, during December 1994 the correlation remained above 0.6 through day 8, while during December 1988 the correlation dropped to 0.6 by day 5. The largest scales (wave numbers 1–3) remain skilful several days longer than the full fields (Fig. 1(b)). We also show in Fig. 1 the ‘perfect model’ results (thin lines) following Lorenz (1982). The different curves represent the correlations between all model forecasts with initial conditions separated by 12 hours, 24 hours, 36 hours, etc. For example, forecasts with initial conditions separated by 12 hours (top curve in Fig. 1(a)) remain well correlated (greater than 0.6) out to 9 days. The largest scales (wave numbers 1–3) remain skilful out to 10 days, but become essentially uncorrelated (0.3 correlation) after two weeks. These results confirm that obtaining skilful predictions beyond two weeks must rely on isolating
regions or modes of variability and/or time periods which exhibit predictive skill beyond that evident in the average skill numbers shown in Fig. 1.

Next we examine the skill of several of the well-known patterns of low-frequency variability. This is done by defining indices (following Wallace and Gutzler 1981) corresponding to the centres of action of the 5 patterns (east Atlantic, PNA, western Atlantic, western Pacific, Eurasian) identified by Wallace and Gutzler (1981, see their Fig. 26), as well as a sixth index corresponding to the NAO as defined in Johannsson et al. (1998). These are defined here in terms of the 500 mb unnormalized height anomaly ($z'$) based on the 15-year mean of the forecasts and re-analysis data. Specifically, the east Atlantic index is $0.5z'(55^\circ N, 20^\circ W) - 0.25z'(25^\circ N, 25^\circ W) - 0.25z'(50^\circ N, 40^\circ E)$, the PNA index is $0.25[z'(20^\circ N, 160^\circ W) - z'(45^\circ N, 165^\circ W) + z'(55^\circ N, 115^\circ W) - z'(30^\circ N, 85^\circ W)]$, the western Atlantic index is $0.5[z'(55^\circ N, 55^\circ W) - z'(30^\circ N, 55^\circ W)]$, the western Pacific index is $0.5[z'(60^\circ N, 155^\circ E) - z'(30^\circ N, 155^\circ E)]$, the Eurasian index is $-0.25[z'(55^\circ N, 20^\circ E) + 0.5z'(55^\circ N, 75^\circ E) - 0.25z'(40^\circ N, 145^\circ E)]$ and the North Atlantic index is $z'(65^\circ N, 23^\circ W) - z'(38^\circ N, 26^\circ W)$.

The ensemble mean prediction of the index is correlated with the observed value of the index for several different time averaging periods (5 days, 10 days, 15 days, 30 days, and the seasonal (January–March, JFM) mean). These results (Fig. 2) show considerable variability in the skill among the patterns. Values less than 0.5 are not significant at the 95% level based on a Fischer's $z$-transform statistic (e.g. Stuart and Ord 1994). The Atlantic region appears to have the least skill at monthly and longer time-scales. The eastern Atlantic region has the least skill, with no skill beyond 10 days, while the western and North Atlantic patterns show skill for 15-day averages, but little beyond that. The Eurasian pattern shows skill for monthly averages (correlation of about 0.5),
though there is no skill at the seasonal time-scale. The Pacific patterns both show skill at monthly time-scales. The PNA index alone shows skill on the seasonal time-scale. Here we focus on the 500 mb height field, though qualitatively similar results are obtained using sea-level pressure (SLP; not shown). The SLP does show somewhat greater skill over the Atlantic and Eurasia, and somewhat less skill over the Pacific at the longer time-scales.

Figures 3, 4 and 5 show the time series of the predicted and observed indices for the 15-day, 30-day and seasonal (JFM) averages which were used to generate the correlations in Fig. 2. Figure 3 shows that, with the exception of the east Atlantic index, the 15-day averages are well predicted in terms of both amplitude and phase. For 30-day averages (Fig. 4) the predicted east Atlantic index shows little phase coherence (correlation) while the amplitude is comparable to that observed. The longer-term variation associated with negative values during the late 80s and early 90s in the Eurasian index, and the general trend toward negative values after the mid 80s in the North Atlantic index are both captured in the predictions. For the seasonal-mean predictions (Fig. 5) the amplitudes of the anomalies predicted for the Atlantic and Eurasia are near zero, suggesting very little useful skill in the predictions. The Pacific indices still show substantial phase coherence though the amplitudes are also somewhat reduced.
Figure 6 is the same as Fig. 2, but for the AMIP runs. This shows that the initial conditions are important. At the shorter time-scales the variance is dominated by fluctuations unrelated to variations of the boundary forcing. For 5-day means none of the regions shows any skill. As the averaging interval is lengthened, the SST-forced signal begins to rise relative to the noise. This is most dramatic for the PNA region for which significant correlations appear even in 15-day means. Both the western Atlantic and western Pacific show substantial positive correlations for the seasonal means, though they fall below the 95% significant level (0.5). The Eurasian and North Atlantic patterns show no evidence of an influence from the SST at any time-scale.

We focus next on the North American region. Figure 7 shows the spatial correlations between the observed (re-analysis) and ensemble-mean predictions of the January–March mean 500 mb height anomalies for each year. Here, in addition to the forecast results, we show the correlations for the AMIP-style ensembles. There is clearly a great deal of year-to-year variation in the correlations. Large positive correlations (greater than 0.7) are found in the forecast results for 1983, 1989, 1990 and 1992. Of these, 1983 and 1992 are El Niño years, while 1989 is a La Niña year; see Niño3 (area 5°N–5°S, 150°W–90°W) SST in Fig. 7. The results from the AMIP integrations show a similar behaviour. Notable differences include a stronger correlation during 1991 and a weaker correlation during 1990, though these two years have rather weak signals (Fig. 8). In fact, all non-ENSO years are characterized by relatively weak signals in this region. Differences in the correlations between the AMIP and DSP results during ENSO events are small. There are, however, considerable differences in the amplitude of the mean
response over the PNA region. Over North America (Fig. 8) the AMIP runs show 40% to 50% increases in amplitude during 1983 and 1989 compared with the response in the DSP experiments. This difference in amplitude is apparently due to the differences in the SST forcing, though the differences in ensemble size (6 in AMIP versus 9 in DSP experiments) could also contribute (see next section). Figure 7 shows, for example, that during 1983 the OI SST anomalies are about 20% larger in the Niño3 region. During 1989 the differences in the Niño3 SST are, however, smaller suggesting the sampling bias (or SSTs in other regions) are contributing to this difference. We shall examine further the impact of the SST differences in the next section, which looks at the signal-to-noise ratio.

(b) Signal-to-noise ratios

We next examine the atmospheric ‘signal’ coming from the prescribed boundary conditions. This is compared with the ‘noise’ coming from the uncertainties in the initial conditions. The focus of this section is on the seasonal (JFM) mean results and the PNA region. The signal is measured by the interannual variance of the ensemble mean about the climate mean, and the noise is measured by the variance of the individual ensemble members about each year’s ensemble mean. The estimate of the signal of a quantity $x$ is

$$ s^2_g = \frac{n}{n-1} \left[ (\bar{x} - \bar{\bar{x}})^2 \right], $$

where an overbar denotes a mean over the $m = 9$ ensemble members, the subscript $g$ indicates signal, and the square brackets denote a mean over the $n = 15$ winters. The
Figure 6. As Fig. 2, but for the Atmospheric Model Intercomparison Project (AMIP) results.

Figure 7. The spatial correlations (left-hand ordinate) between the re-analysis and the 9-member ensemble mean predictions (Dynamical Seasonal Predictions, DSP), and between the re-analysis and the 6-member ensemble mean simulations (Atmospheric Model Intercomparison Project, AMIP), for the mean January–March 500 mb height field encompassing the region 25°N–70°N, 60°W–150°W. Line plots show the Niño-3 (area 5°N–5°S, 150°W–90°W) sea surface temperature anomalies (degC, right-hand ordinate) used for the DSP and AMIP runs.
estimate of the noise of \( x \) (indicated by subscript \( w \)) is
\[
s^2_w = \frac{m}{m-1} [(x - \bar{x})^2]. \tag{2}
\]
The total variance is the sum of (1) and (2), and the signal-to-noise ratio is
\[
T = \frac{s^2_s}{s^2_w}. \tag{3}
\]

We note that the estimate of the signal (1) is biased for finite sample size (see e.g. Rowell et al. 1995). In particular,
\[
\frac{E(s^2_s)}{E(s^2_w)} = \frac{(1/m)\sigma^2_w + \sigma^2_s}{\sigma^2_w}, \tag{4}
\]
where
\[
\sigma^2_s = \frac{n}{n-1} [(\mu - [\mu])^2], \tag{5}
\]
is an unbiased estimate of the signal. Here \( E(\bar{x}) = \mu \) is the population mean for a particular winter, and \( E(s^2_w) = \sigma^2_w \), is the population variance. From (4), it is clear that experiments with few ensemble members will tend to have inflated signal-to-noise ratios, especially in regions of large noise. For example, in the absence of forcing, the expected value of the forcing estimate from the AMIP runs would be 50% larger (9/6) than the estimate from the DSP runs.

A standard test for the signficance of the signal (e.g. DeGroot 1975) is based on the ratio
\[
s^2_s \sqrt{\frac{s^2_w}{m}} = mT. \tag{6}
\]
For normal random variables, and under the null hypothesis that all the signals ($\mu$ for each winter) are zero, (6) has an $F$-distribution with $(n-1, n(m-1))$ degrees of freedom. The 5% significance level is obtained at $F_{14,120}(0.05) = 1.78$. This translates to a significance level for $T$ of $1.78/9 = 0.2$.

We begin by showing, in Fig. 9(a)–(c), the global distribution of the JFM mean 500 mb height field noise, signal, and total variance from the DSP integrations. We note that the quantities in Fig. 9(a)–(c) are actually the standard deviations. The signal-to-noise ratios are shown in Fig. 9(d). The largest noise occurs in the winter hemisphere at high latitudes and in the oceanic storm tracks where the values exceed 50 m. The signal, by comparison, is considerably weaker, with values exceeding 30 m only in the Gulf of Alaska and the high latitudes of the northern hemisphere (north of 60°N, between 60°W and 60°E). The distribution of the total variance thus largely reflects the distribution of the noise. The signal-to-noise field (Fig. 9(d)) highlights the relatively large values (ranging from 3 to greater than 10) in the tropics, where the variation in the seasonal-mean height field (while small compared to that in middle latitudes) is largely forced by the SSTs. In the extratropics the signal-to-noise ratio is generally less than one. In both hemispheres the largest values occur over the subtropical Pacific. In the northern hemisphere the largest middle-latitude values occur over the northeastern Pacific, central Canada, and the region north of 60°N, between 60°W and 60°E. The southern hemisphere middle and high latitudes show a surprisingly strong signal-to-noise ratio compared with the northern hemisphere.
As mentioned earlier, our signal is likely impacted by the weaker SST anomalies in the blended SST dataset compared with the OI SST data. Figure 10 compares the signals in the DSP and AMIP runs as percentages of the total variance. For these results we limit the DSP ensemble to include only the first 6 members, to avoid any bias in the estimate of the signal due to differences in the ensemble size (see (4) above). The DSP results (Fig. 10(a)) are consistent with the signal-to-noise ratios (Fig. 9(d)) discussed earlier, with the forced variance in the tropics exceeding 90% of the total over much of the tropical Pacific Ocean, while considerably smaller values occur in the extratropics. In the northern hemisphere middle latitudes, the forced variance exceeds 30% of the total only over the eastern North Pacific. The results for the AMIP runs, while generally similar, show substantial increases in the forced variance over much of the PNA region. For example, much of the eastern North Pacific now shows fractions greater than 40%,
Figure 11. Variance of the mean January–March precipitation (mm day$^{-1}$) for the years 1981–94 for: (a) the DSP runs; (b) the AMIP runs; and (c) the merged Xie-Arkin rainfall data. See text for details.

with values exceeding 50% extending well north of 30° latitude. Also, a substantial region of eastern North America (extending from the east coast into the Great Lakes region) has values exceeding 40%.

The above differences in the extratropical response are associated with changes in the tropical forcing. This is illustrated in Fig. 11(a) and (b) in terms of the variability of the JFM mean tropical precipitation for the DSP and AMIP runs, respectively. The AMIP results show generally larger variance, especially east of the dateline, and in a narrow band just north of the equator. For comparison, Fig. 11(c) shows the precipitation variance from observations (Xie and Arkin 1996). While the observations are themselves subject to uncertainty, the comparison suggests that the magnitude of the precipitation variability in the AMIP results is somewhat closer to that of nature. On the other hand,
the patterns of high variability in both model results show considerable differences from the observations. The model results have a clear minimum at the equator, with a band of high variability north of the equator, and another region of high variability south of the equator and west of about 160°W. The observations show, instead, a single broad region of high variability within about 10° of the equator.

To further assess the impact of the differences in the SST data, we have rerun the nine DSP experiments for 1983 using the OI SST (which has a 20% larger anomaly in the Niño3 region, see Fig. 7) and, indeed, find a stronger (20% increase) mean response in the region of maximum response in the North Pacific (not shown). These results are consistent with the study by Kumar and Hoerling (1997) which found that the amplitude of the signal in the extratropics increases nearly linearly with the strength of the SST anomaly.

One of the key issues that needs to be addressed is whether the signal and noise fields obtained by the GEOS-2 model are realistic. While we cannot compute ensembles from nature (there is only one realization), we can compare the total JFM variance generated by the model with the estimate of nature’s variance from the GEOS-1 re-analysis. We note the GEOS-1 re-analysis estimates are virtually identical to those obtained from the National Center for Atmospheric Research (NCAR)/NCEP re-analysis (not shown). The only issue in such a comparison concerns the sampling error. We do not know the sampling error of the variance estimate obtained from the single 15-year realization from nature. On the other hand, we can generate multiple 15-year realizations from the model ensemble members, so we can ask the question whether the one realization from nature falls within the sampling distribution of the model results. Note that this is somewhat different from our previous utilization of the ensemble members, in that now we constrain the statistics to only those we can also compute from the observations. In this case we are interested in the sampling distribution of

\[ s^2 = \frac{n}{n-1} [(x - \bar{x})^2]. \]  

To illustrate the variability of \( s^2 \) due to sampling, we present in Fig. 12 several different 15-year realizations of the square-root of \( s^2 \) from the DSP runs. Eight of them are from the model, and one of them is the single realization from nature (the re-analysis). The realizations from the model were determined simply by taking all the first members (those starting at the earliest times) of the ensemble for each year, then taking all the second members of the ensemble for each year, and so on, to construct separate 15-year estimates of the variance. This shows a considerable degree of variability in the variance estimates, especially over the North Pacific region, ranging from very weak (upper-right panel) to very strong (lower-middle panel). The estimate from the re-analysis is in the centre panel. This shows that, while many of the model estimates are somewhat weak compared with nature, at least some of the estimates do have variances approaching those obtained by nature over the period 1980/81 to 1994/95.

The above results can be further quantified by generating a very large number of different realizations of 15 years from the 9-member ensembles generated by the model (in principle all permutations of the 9 x 15 realizations). This was done for an area over the North Pacific where the signal is strongest. In this case one million permutations were generated to produce a histogram of the total variance (Fig. 13). Comparisons with the single realization from the re-analysis (heavy vertical line in Fig. 13), shows that nature falls in the upper tail area of the model’s distribution (though still within 2 standard deviations). This is basically in agreement with the much more limited results shown above, indicating that the realization from nature would be a
somewhat rare occurrence. We note that the histogram of the variance has a somewhat narrower distribution than would be expected to result from randomly sampling a normal distribution (a Chi-squared distribution with 14 degrees of freedom; not shown). This is not inconsistent with the presence of a forced component, since in the limit of a purely forced signal the histogram would approach a delta function.

Figure 13 also shows the results for the AMIP integrations. This curve shows a distribution for the total variance that is narrower (smaller variance in the variance) and slightly shifted to the right towards larger values compared with the DSP results. This suggests a somewhat stronger (on average) and more robust ENSO response in the AMIP runs. This is consistent with our previous results. In order to determine if these differences are real, we again sub-sampled the 9-member DSP results to be consistent with the AMIP sample size (6 ensemble members). This allows the generation of a large number of different 6-member ensemble results. The histogram generated in this way almost coincides with that of the 9-member DSP results, indicating that the estimate of the total variance from the smaller sample is not substantially biased with respect to the 9-member results. The shading represents the spread in the 6-member DSP results (+/− one standard deviation). The comparison of the shaded region with the AMIP results suggests that the AMIP runs do produce a more consistent or robust response to the ENSO forcing than the DSP results (smaller variance in the variance).
Figure 13. Histograms of the variance of the January–March (JFM) mean 500 mb height (z, m) field for the period 1980–94 over the North Pacific region (40°N–60°N, 180°W–135°W). The histograms were obtained by generating one million different estimates of the variance, where each estimate consists of a different realization of the 15 years. Values are binned into 1 m² intervals from 0 to 7000 m². The vertical axis is the number of cases falling within each bin. The two thin-line curves are from the 9-member Dynamical Seasonal Predictions (DSP) ensembles and the 6-member Atmospheric Model Intercomparison Project (AMIP) runs. The thick-line curve is the average of a large number of different 6-member ensembles sub-sampled from the 9-member DSP runs. The shading represents the uncertainty in the 6-member DSP results (+/− one standard deviation). The heavy solid vertical line denotes the one realization from nature (from the re-analysis; the vertical extension of the line is for clarity).

also evidence of a stronger signal in the AMIP runs though the shift of the distribution to the right (with respect to the DSP distribution) falls within the sampling error of the 6-member ensemble.

The above results provide indirect evidence that the model response to the tropical Pacific SST anomalies over the North PNA region, while not inconsistent with the observations, is likely to be weaker than found in nature. We can obtain a more direct assessment of the signal-to-noise ratio by making some assumptions about the nature of the signal. The previous results suggest that most of the response in middle latitudes is due to the SST forcing associated with El Niño. We assume a simple linear-regression model with the SST in the Niño3 region as the predictor of the 500 mb height. The sum of squares in (7) is then partitioned into two parts: the sum of squares due to the regression (the signal); and the sum of squares of the residuals about the regression (the noise). The signal-to-noise ratio for the regression can be written as

$$T_{\text{reg}} = \frac{r^2}{1 - r^2},$$

where $r$ is the correlation between the Niño3 SST and height field at each grid point. Also, for normally distributed residuals, and under the null hypothesis that the slope of the regression line is zero, $(n - 2)T_{\text{reg}}$ has an F-distribution with $(1, n - 2)$ degrees of freedom (see e.g. Draper and Smith 1966). The 5% significance level is obtained at $F_{1,13}(0.05) = 4.67$. This translates to a 5% significance level for $T_{\text{reg}}$ of 4.67/13 = 0.36.
Figure 14. (a) The temporal correlation between the Niño-3 SST (5°N–5°S, 150°W–90°W) and the predicted 500 mb height at each grid point computed from the mean January–March fields for 1981–95. The correlations are based on the averaged variances and covariances from one million different 15-year realizations (based on different permutations of the ensemble members). (b) The same as (a), but for the single realization of nature (from the re-analysis). (c) The signal-to-noise ratio for the model. (d) The signal-to-noise ratio from the observations. See text for details.
Figure 15. The individual ensemble members of the Dynamical Seasonal Predictions (DSP) predictions for 1983 (an El Niño year). Results are shown for the mean January–March 500 mb height field (m) in the North American region. The upper-left panel shows the ensemble mean, the upper-middle panel shows the standard deviation about the ensemble mean, and the upper-right panel shows the observations (re-analysis) for that time period. The remaining panels show the individual ensemble members.

Note that $T_{\text{reg}}$ is a biased estimate of the population signal-to-noise ratio. The ratio of the expected value of the signal to the expected value of the noise is similar in form to (3), but with $m$ replaced by $n - 2$, and the population signal is that obtained from the regression with the true parameters. The above results hold for a single 15-year realization. For the model results, we can again generate a large number (one million) of 15-year realizations. These realizations are used here to generate improved estimates of the parameters of the regression, so that the significance point for $T_{\text{reg}}$ reduces to the limiting value of $1/(n - 2) = 1/13$. Note, however, that this does not alleviate the bias of the estimated signal.

The results from the regression are shown in Fig. 14 for both the model and nature. The maps of the correlations are similar (cf. Fig. 14(a) and (b)), with both the model and observations showing moderately strong correlations (greater than 0.5) over the whole
globe, though the regions of high correlation tend to be larger for the observations. The signal-to-noise ratios (Fig. 14(c) and (d)) reflect the correlation patterns. The model shows the highest signal-to-noise ratios (greater than 2) largely confined to the eastern tropical Pacific, while the observations show high signal-to-noise ratios throughout most of the tropics. Both the model and nature show evidence of a wave response in both hemispheres emanating from the tropical Pacific. The southern hemisphere wave train shows exceptionally large middle-latitude signal-to-noise ratios (values greater than 2 in the observations). In general, the model generates weaker and less extensive regions of signal-to-noise ratios. Over the PNA region, in particular, the observations show a much larger region in which the signal-to-noise ratios exceed 0.5, though the maximum values do not exceed 1.

A further assessment of the noise (variability between different ensemble members) can be obtained by simply inspecting the nine ensemble members for individual years. We show next the results for two of the most strongly forced years over the PNA region.
Figure 17. As Fig. 15, but for 1989 (a La Niña year).

Figure 15 shows the ensemble members for the 1983 El Niño event. The ensemble mean, the standard deviation among the ensemble members, and the results from nature (the reanalysis) are shown in the upper-left, upper-middle, and upper-right panels respectively. The following 9 panels show the individual members. The ensemble-mean response of the model is similar in pattern, though considerably weaker than the observed anomalies. There are clearly some ensemble members with a very weak signal while a few are quite strong, approaching that of nature. Figure 16 is the same as Fig. 15, but shows precipitation over the USA. The major observed positive precipitation anomalies over the west coast and over the south-eastern USA are evident in the ensemble-mean model response, though also of reduced amplitude. This again reflects a considerable variability among the ensemble members. For example, the first ensemble member shows no evidence of enhanced rainfall along the west coast: this is consistent with the very weak height anomaly off the west coast for that case (Fig. 15). Another year (a La Niña year, 1989) is presented in Figs. 17 and 18. Again we find that there are ensemble members with a weak response, but at least a few have amplitudes similar to that found in nature.
Compared with the 1983 El Niño event, the La Niña event shows somewhat greater intra-ensemble variability (noise), suggesting a less robust response to the SST forcing.

4. Discussion and Conclusions

This study has shown that, despite the approximate 2-day average doubling time for small errors, skilful seasonal (JFM) forecasts can be achieved in middle latitudes. The results, based on an ensemble of atmospheric GCM seasonal forecasts and long-term simulations (1980/81 to 1994/95) forced by observed SSTs, show that, while all memory of the initial conditions is lost beyond about one month, SST forcing can produce skilful forecasts at seasonal time-scales. The skilful extratropical seasonal forecasts are, however, largely confined to the wave-like ENSO responses emanating from the tropical Pacific into both hemispheres. In the northern hemisphere, this is similar to the well-known PNA pattern. Other low-frequency modes of variability previously identified...
in the literature (e.g. Wallace and Gutzler 1981) are less strongly influenced by SST anomalies. In particular, the eastern and North Atlantic patterns and the Eurasian pattern show no significant skill from SST anomalies, while the western Atlantic and western Pacific patterns show marginal skill at the seasonal time-scale. Somewhat surprisingly, the PNA pattern has significant forecast skill associated with SST anomalies even at sub-monthly time-scales.

While the basic results presented here concerning the ENSO-related seasonal-forecast skill are similar to those from other models (Shukla et al. 2000), the key difference among the various models is in the strength of the signal-to-noise ratios in the extratropics. For example, the maximum signal-to-noise ratio in the North Pacific varies among the models by almost an order of magnitude. The GEOS-2 model falls on the side of weaker signal-to-noise ratios (near one). Determining which, if any, of the models provide realistic signal-to-noise ratios is difficult since nature provides only a single realization. We can, however, determine whether statistics computed from nature's single realization (the observations) are consistent with each model's distribution of the same statistic (determined from the ensemble). For example, a comparison in the North Pacific of the distribution of the variance of the GEOS-2 model response with the observed variance, shows the single realization from nature is only marginally consistent with the model's distribution (falls about two standard deviations from the mean), suggesting the GEOS-2 model response to ENSO forcing may be too weak. On the other hand, a simple regression model employing the Niño3 SST as a predictor suggests that signal-to-noise ratios in the northern extratropics near 1.0 may not be unreasonable.

The response obtained in our DSP experiments was impacted to some extent by the use of the blended SST data (Reynolds and Marsico 1993), which have somewhat weaker SST anomalies in the tropical Pacific than the more recent OI SST data (Reynolds and Smith 1994). Our AMIP results, which do employ the OI SST, show consistently larger forced signals in the PNA region. Also, a rerun of our 1983 DSP ensemble using the OI SST produced a 20% increase in the strength of the response in the North Pacific. This is consistent with the Kumar and Hoerling (1997) finding that the response increases nearly linearly with the strength of the SST anomaly.

The issue of the sensitivity of the response to the SST requires further study, both for understanding the interannual variability of the ENSO response and for defining the accuracy requirements of SST predictions. While it appears that the amplitude of the tropical SST anomaly is important (e.g. Kumar and Hoerling 1997, and this study), previous modelling and observational studies leave a rather conflicting picture as to the importance of the location and details of the SST to the extratropical ENSO response. Given the generally small signal-to-noise ratios in the extratropics, it is likely that part of the confusion is simply the result of statistical sampling errors. While most recent GCM ensemble climate simulation and prediction studies have employed ensembles of order 10 or less (including the current study), this is unlikely to be adequate to address many remaining questions involving the sensitivity of the response to SST forcing. For example, the basic issue of determining whether the extratropical response is different for different ENSO events is likely to require considerably more than 10 ensemble members. In fact, one of the future objectives of the DSP project is to carry out a much larger number of forecasts (approximately 100) for a few key cases.

The wide range of signal-to-noise ratios among models, mentioned above, suggests that there are still substantial deficiencies in the modelled response to SST forcing. Recent studies have shown sensitivities to rather subtle technical details in the implementation of parametrizations (e.g. Kumar et al. 1996), making it difficult to design
definitive experiments to assess physical improvements to models based on their responses to SST forcing. Ultimately, the answer to the question of what is the correct response must await an improved understanding gained from further analyses of the modelled and observed atmospheric response to previous and future ENSO events. The apparently complicated nature of the response, involving feedbacks with the transients (e.g. Held et al. 1989), local energy sources and modifications to wave propagation characteristics associated with the background flow (e.g. Trenberth et al. 1998), suggest we must find new, more sensitive measures for validating climate models.

ACKNOWLEDGEMENTS

This work was carried out under NASA's Seasonal-to-Interannual Prediction Program (NSIPP) with support from the NASA Global Modeling and Analysis Program (UPN 622-24-47).

REFERENCES


1994 Forecasting global ENSO-related climate anomalies. Tellus, 46A, 381–397

Barnett, T. P., Arpe, K., Bengtsson, L., Ji, M. and Kumar, A.

1997 Potential predictability and AMIP implications of midlatitude climate variability in two general circulation models. J. Climate, 10, 2321–2329

Battisti, D. S., Bhatt, U. S. and Alexander, M. A.

1995 A modeling study of the interannual variability in the wintertime North Atlantic Ocean. J. Climate, 8, 3067–3083

Bengtsson, L., Schleser, U., Roeckner, E., Latif, M., Barnett, T. P. and Graham, N.

1993 A two-tiered approach to long-range climate forecasting. Science, 261, 1026–1029

Branković, Č., Palmer, T. N. and Ferranti, L.

1994 Predictability of seasonal atmospheric variations, J. Climate, 7, 217–237

Branstator, G.


Chou, M.-D. and Suarez, M. J.


DeGroot, M. H.

1975 Probability and statistics. Addison-Wesley Publishing Co., Reading, MA, USA

Draper, N. R. and Smith, H.


Friedrich, K.

1994 An ENSO impact on Europe? Tellus, 46A, 541–552

Gates, W. L.


Geisler, J. E., Blackmon, M. L. and Bates, G. T.


Halpert, M. S. and Ropelewski, C. F.

1992 Surface temperature patterns associated with the Southern Oscillation. J. Climate, 5, 577–593

Hamilton, K.


Held, I. M. and Kang, I.-S.


Held, I. M., Lyons, S. W. and Nigam, S.


Helfand, H. M. and Labraga, J. C.


Lorenz, E. N. 1982 Atmospheric predictability experiments with a large numerical model. *Tellus*, 34, 505–513


Xie, P. and Arkin, P. 1996 Analyses of global monthly precipitation using gauge observations, satellite estimates and numerical model predictions. *J. Climate*, 9, 840–858