Forecast Skill and Low-Frequency Variability in NMC DERF90 Experiments

WILBUR Y. CHEN AND HUUG M. VAN DEN DOOL

Climate Analysis Center, NOAA/NWS/NMC, Washington, D.C.

(Manuscript received 26 October 1993, in final form 28 December 1994)

ABSTRACT

A series of 90-day integrations by a low-resolution version (T40) of the National Meteorological Center's global spectral model was analyzed for its performance as well as its low-frequency variability behavior. In particular, 5-day mean 500-mb forecasts with leads up to 88 days were examined and compared with the observations. The forecast mean height decreased rapidly as forecast lead increased. A severe negative bias of the mean height in the Tropics was caused by a negative temperature bias and a drop of the surface pressure of about 2 mb. The forecast variance also dropped rapidly to a minimum of 75% of the atmospheric standard deviation before being stabilized at day 18. The model could not maintain large anomalous flows from the atmospheric initial conditions. However, it is quite capable of generating and maintaining large anomalies after drifting to its own climatology and temporal variability.

At extended ranges, the model showed better skill over the North Pacific than North Atlantic when the season advanced to the colder period of the DERF90 (dynamical extended-range forecasts 1990) experiments. The model also displayed dependence on circulation regimes, although the skill fluctuated widely from day to day in general. Blocking flows in the forecast were found to systematically retrogress to the Baffin Island area from the North Atlantic. Therefore, improvements of the model's systematic errors, including its drift, appear to be essential in order to achieve a higher level of forecast performance. However, no generalization can be made due to the usage of a low-resolution model and the experiments being carried out over a rather short time span, from only 3 May to 6 December 1990.

1. Introduction

In December 1986 and early 1987, the National Meteorological Center (NMC) conducted a large experiment in dynamical extended-range forecasts (DERF). The primary objective was to assess the feasibility of extending operational numerical forecast beyond 10 days to monthly predictions. Comprehensive results of that DERF experiment can be found in Tracton et al. (1989), Roads (1987, 1988), and Chen (1989, 1992). Another opportunity at NMC to conduct a further large-scale numerical experiment developed in 1990 when a Cray supercomputer was newly installed. A long series of integrations, each out to 90 days, were then conducted by Drs. Saha, Kalnay, Kanamitsu, and Van den Dool. These so-called DERF90 experiments were run with a horizontal spectral resolution of T40 and 18 vertical sigma levels. In comparison with the global model's present resolution of T126 and 28 vertical levels, the DERF90 model had relatively low resolution. Complete documentation of the model and its updates can be found in Sela (1980), Kanamitsu et al. (1989), and Kalnay et al. (1990).

Corresponding author address: Dr. Wilbur Y. Chen, Climate Analysis Center, National Meteorological Center, W/NMC51, Prediction Branch, Washington, DC 20233.

The first DERF90 run was initiated on 3 May 1990, with 90-day integrations conducted daily thereafter until 6 September of the same year. Progressively shorter forecasts were also run from 7 September (an 89-day integration) to 6 December (a 0-day forecast) to complete a Lorenz data block. Using this large dataset, some conventional forecast performance was evaluated by Van den Dool (1994), and the intraseasonal variability was investigated by Toth et al. (1991). Anderson and Van den Dool (1994) looked into the return of the forecast skill, and Rukhovets et al. (1994) studied the most predictable areas and patterns. Van den Dool (1994) further addressed the predictability of the model atmosphere, considering sensitivity to both initial atmospheric conditions and lower boundary conditions. In this report we will focus on the behavior of the low-frequency fluctuations in the 5-day mean forecasts, with various leads from 3 days up to 88 days. For the sake of brevity, this report will be largely confined to the results of the Northern Hemisphere, and only the geopotential height at the 500-mb level ($Z500$) will be examined.

Particular attention is paid to the model's ability or inability to generate and maintain large blocking flows, since the last DERF run (Tracton et al. 1989) indicates that a large portion of the unskillful forecasts can be traced to the model's inability to predict the evolution of blocking events beyond a few days into the forecast.

Tibaldi and Molteni (1990) also found a severe underestimation of blocking frequency in the ECMWF (European Centre for Medium-Range Weather Forecasts) medium-range forecasts.

Some different characteristics exhibited by the model over the North Pacific and North Atlantic are also explored. In analyzing the earlier DERF experiment, Chen (1992) found a distinct difference in model performance between the Pacific and Atlantic sectors. The difference was attributed to the influence from the tropical Pacific, because this earlier DERF experiment happened to fall within a period when a strong warm phase of the El Niño–Southern Oscillation (ENSO) was operating. During the present DERF90 experiments, the ENSO effect is considered to be negligible. We would like to know whether some difference in forecast performance might still exist.

The systematic biases of both the mean height and standard deviation in the forecasts are investigated first, in sections 2 and 3. Issues of the maintenance of large forecast anomalies are then examined in detail in section 4. Dependence of the forecast skill on geographical regions as well as on circulation regimes is discussed in section 5. A systematic forecast error caused by an erroneous retrogression of Atlantic blocks is presented in section 6. Last, section 7 summarizes the results. A detailed examination of the MRF model characteristics using a 10-yr-long integration is reported in Chen and Van den Dool (1995).

2. Systematic bias in the forecast mean height

Figure 1 shows the systematic bias of the mean height at various forecast leads. Since we are more interested in the medium- to extended-range forecasts, 5-day mean forecasts were used, where $D + 3$, $D + 8$, . . . , $D + 88$ denote day 1–5, day 6–10, . . . , day 86–90 mean forecast, respectively. In DERF90 experiments, there are altogether 213 cases of $D + 3$ forecasts, and as forecast lead time increases, eventually there are only 128 cases of $D + 88$ forecasts. The 5-day mean anomaly of both forecast and verifying analysis were obtained first; the time mean over all available cases of each forecast lead was then calculated and shown in Fig. 1. For the calculation of the Z500 anomaly, the Climate Analysis Center 14-yr Z500 climatology (1979–92) was used. As shown in Fig. 1 where the negative anomalies are shaded for easy inspection, the NMC analyzed Z500 during the DERF90 period (first panel) had slightly higher height over most of the midlatitudes except over part of Siberia. Other panels of the figure display systematic bias of the mean height at various forecast leads. The $D + 3$ forecasts show similar height anomaly patterns as the analyzed over the extratropical latitudes. However, −20-m anomalies begin to appear over the subtropical belt. At $D + 8$, the subtropical average height anomalies decrease to −40 m, but the positive height anomalies over the East Asian, western Pacific, North American, and Arctic regions increase in magnitude. At increased forecast lead times, the mean forecast bias further intensifies over the subtropical belt. By $D + 28$, the negative height anomalies over the subtropical belt settle at about −60 m, while the extratropical positive height anomalies stay at about 40 m.

The seasonal march of the global mean systematic bias of the forecasts is shown in Fig. 2. The top solid curve is the time series of the verifying analyses during the DERF90 period. While small positive anomalies are observed during most of the first half of the DERF90 period, the analyzed anomalies fluctuate about the climatological mean for the rest of the DERF90 experiments. The $D + 3$ global mean anomalies (the top dashed curve) are seen to drop dramatically by about 10–15 m. With increasing forecast lead, (i.e., $D + 8$, $D + 13$, $D + 28$, $D + 58$, and $D + 88$ represented by the alternating solid and dashed curves, respectively), larger systematic drops can be clearly seen. The rate of decrease becomes gradual from $D + 28$ to $D + 58$. Beyond $D + 58$ there is very little further decrease. The systematic biases appear to have approached the model's climate. We note that during the second half of the DERF90 period when the season gets colder the forecast negative biases are seen to be much larger than those of the first period.

Most of the forecast mean bias has developed in the Tropics as shown in Fig. 3, where the average bias for the tropical belt from 20°S to 20°N is compared with that of the extratropical latitudes that exclude 20°S–20°N. Note that the units are a factor of 5 different, 10 m for the Tropics versus 2 m for the extratropics. The leftmost bar is for the verifying analysis. The others represent forecasts from $D + 3$ to $D + 88$. At $D + 28$ the tropical average bias is more than −60 m, while the extratropical average bias is only about −10 m. But while the tropical biases approach a steady value beyond $D + 28$, the extratropical negative biases keep increasing. This increase is related to the large negative biases found beyond $D + 28$ in Fig. 1 in the subtropics. Figure 1 also indicates some positive biases over the Arctic. Therefore a similar chart like Fig. 3 but for the midlatitudes (40°–70°N) and Arctic latitudes (70°–90°N) was also obtained for further detail. As shown in Fig. 4, positive biases develop rapidly over the Arctic latitudes for $D + 3$ and $D + 8$, reaching a maximum of 5 m at $D + 23$. Beyond $D + 23$, the positive bias gradually decreases, to near the climatological value at $D + 88$. The midlatitudes also show slight increase in positive mean bias for $D + 8$ and $D + 13$. Beyond $D + 13$ the tendency of the forecast mean height is a steady decrease, ending up with a negative mean bias of about 5.5 m.

Note that the mean biases over the middle and Arctic latitudes are within +5 and −5 m, versus −60 m for the Tropics. The severe negative systematic bias in the Tropics was caused primarily by a negative tempera-

ture bias (Van den Dool et al. 1993). A substantial drop in the surface pressure of about 2 mb by $D + 23$, possibly due to the neglect of the "water mass forcing" term in the vertically integrated continuity equation used in the forecast model (Van den Dool et al. 1993), also contributed substantially to the negative systematic bias in the Tropics.

3. Systematic bias in the forecast standard deviation

Figure 5 presents the distribution of the standard deviation (SD) for the verifying analyses and the forecasts at various leads. The deviations were calculated from those means presented in Fig. 1. The number of cases in these statistics vary from 213 for $D + 3$ forecasts to 128 for $D + 88$ forecasts. The $D + 3$ SD looks very similar to the observed (ANAL), except a little bit stronger over the North Pacific and a little bit weaker over the Arctic Siberia. As the forecast lead time increases, the large-amplitude fluctuations are seen to decrease rapidly, reaching a minimum at $D + 18$, while the 40-m SD contour does not seem to shrink much, implying that about the same amount of smaller-amplitude fluctuations prevail throughout all forecast leads. Beyond $D + 28$, the SD pattern varies very little. It appears to be approaching the model's own variability characteristics.

Separate contributions to the globally averaged SD from Tropics and extratropics are shown in Fig. 6. The extratropical SDs show rapid decrease in magnitude to a minimum of about 80% of the observed SD at $D + 18$ and remain approximately the same magnitude beyond this lead. Since these extratropical SDs include those of the Southern Hemisphere, together with those results of Fig. 5, they suggest that the systematic bias of SD in the Southern Hemisphere is much the same as that of the Northern Hemisphere. The systematic SD bias in the Tropics shows a rapid decrease to a minimum at $D + 8$ and then a rapid increase to approximately the original level. Another decrease in the SD magnitude can be observed beginning at $D + 53$. These tropical SDs are, however, much smaller than those of the extratropical ones.

Figure 7 shows the drop of the SD magnitude for the Arctic latitudes and the northern midlatitudes. Both charts show decreases to a minimum at $D + 18$ with only very slight increase thereafter. However, the drop from the observed SD to $D + 18$ SD is very dramatic for the northern midlatitudes, from 44 to 33 m, a 25% reduction from its original SD magnitude.

4. Maintenance of large anomalies

As mentioned when Fig. 5 was presented, the SD decrease occurs over the areas where the SD maxima are located. Since the SD maxima are contributed primarily by the large-amplitude fluctuations, we expect much fewer large-amplitude fluctuations taking place in the forecasts than in the real atmosphere. Persistence of large anomalies is another characteristic we would like to study. For instance, long-lasting blocking flows can be observed in the real atmospheric flow. Would blocking flows in the forecasts, although with much reduced magnitude, be as likely and as persistent as those of the real atmosphere?

For this query, we focus our attention first on the North Pacific where, from Fig. 5 $D + 3$ forecast SD chart, we might find most of the blocking flows. We ask whether those blocking flows would continue to hold on at $D + 8, D + 13, \ldots, D + 28$?

A composite of $D + 3$ forecast anomalies—when each one exceeds 100 m at 50°N, 180°—was constructed first. We found 36 cases meeting this criterion out of 188 $D + 3$ forecasts. Although there were 213 $D + 3$ forecasts available, there were only 188 cases that had a corresponding $D + 28$ forecast. Hence, the pool to choose from consisted of only 188 cases. As the forecast lead time increased, the chosen forecasts evolved. The composites of the evolved forecasts were then obtained for $D + 8, D + 13, \ldots, D + 28$ forecasts. The left-hand panels of Fig. 8 show the temporal evolution of these composites. The right-hand panels show their corresponding verifications. Positive anomalies are shaded for easy inspection. Over the North Pacific, the real atmosphere maintains a long-lasting composite ridge, while the forecasts show a rapid demise of the composite ridge. On average, the blocking flows found in $D + 3$ forecasts last for only about five days. Beyond that no trace of blocking ridges can be found over the North Pacific.

One might suspect the inability of the model to maintain a blocking flow in the forecast to be a consequence of the rapid drop in anomaly amplitude as seen in Fig. 5. In other words, the swift drifting of the model's flow variabilities from the real atmospheric magnitude to the model's own magnitude might cause the disappearance of large flow anomalies. If this is the case, we might still be able to find persistent flows in the forecasts after the model settles down to its own climate and temporal variability, albeit with smaller than observed blocking flow magnitude. This has been shown to be the case with the DERF90 data and with another general circulation model run by Anderson (1993). Specifically, we may be able to find persistent flows after $D + 18$ and beyond, when without further drifting the model is generating its own dynamically consistent fluctuations.

To check out this possibility, the same procedure used to generate Fig. 8 is now used to obtain another set of charts but starting with $D + 18$ forecasts instead of $D + 3$ forecasts. Figure 9 shows the results of the composite and the subsequent evolutions. This time only 30 cases out of a total of 173 exceed the 100-m criterion and enter into the composite. Both a smaller pool to choose from and beginning at a longer forecast lead make the selected cases smaller—30 instead of 36. Immediately, we see the persistence of the blocking

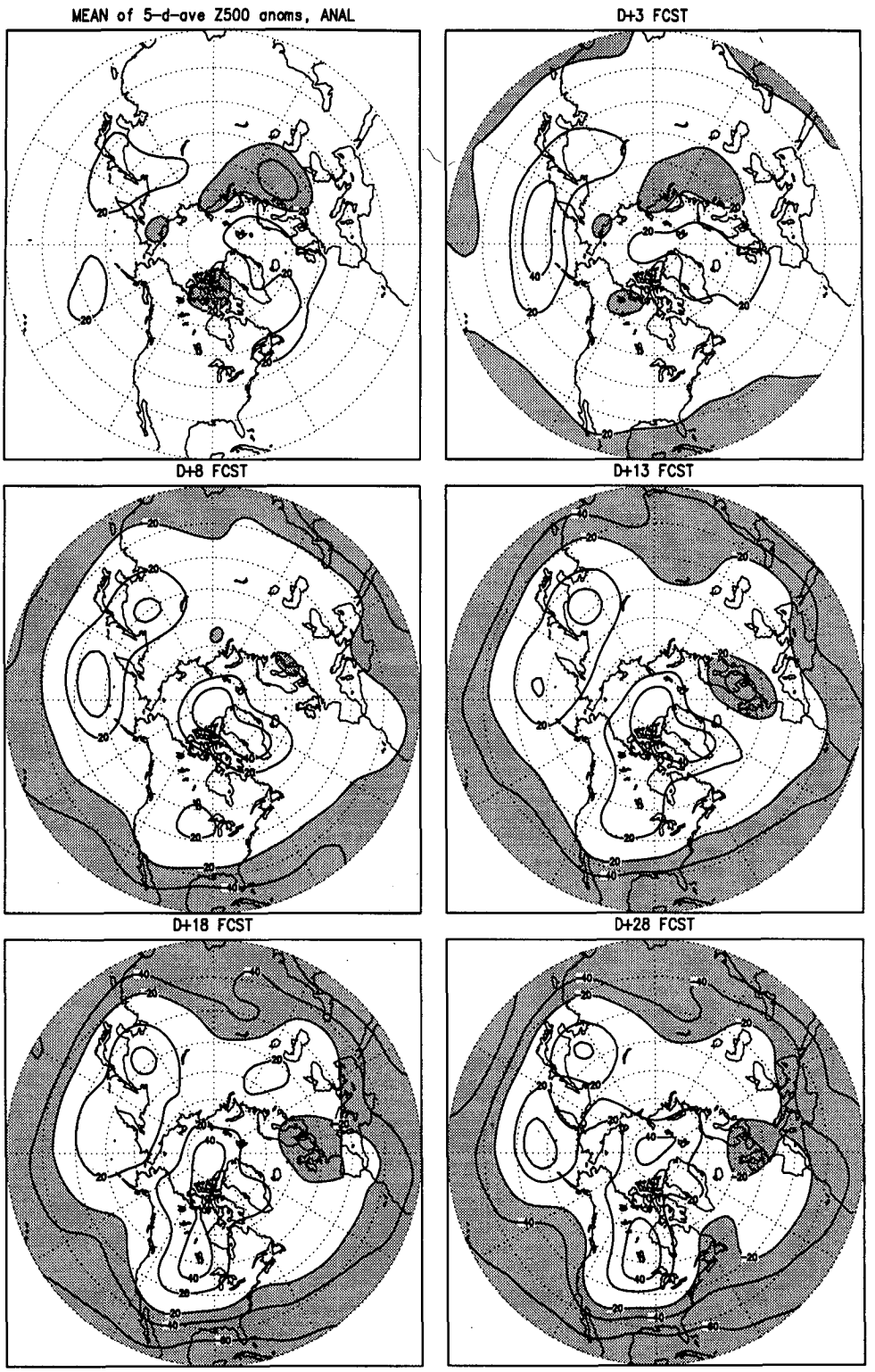

FIG. 1. Analyzed and forecast Z500 mean anomalies. CAC climatology, 1979–92, was used to obtain anomaly; $D + 3$ denotes days 1–5 mean forecast. All other forecast leads have the same convention. The contour interval is 20 m. Negative anomalies are shaded.

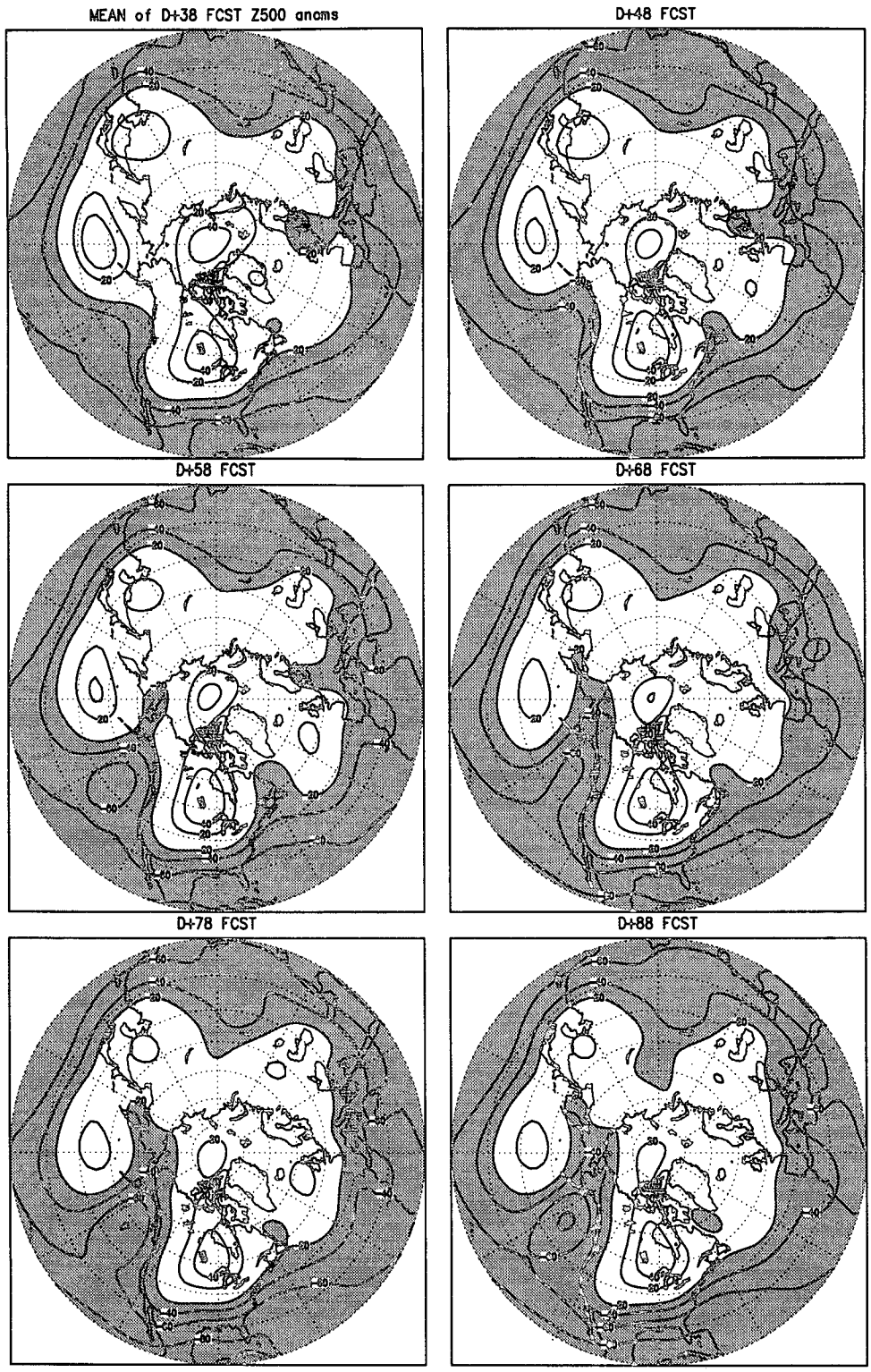

FIG. 1. (*Continued*)

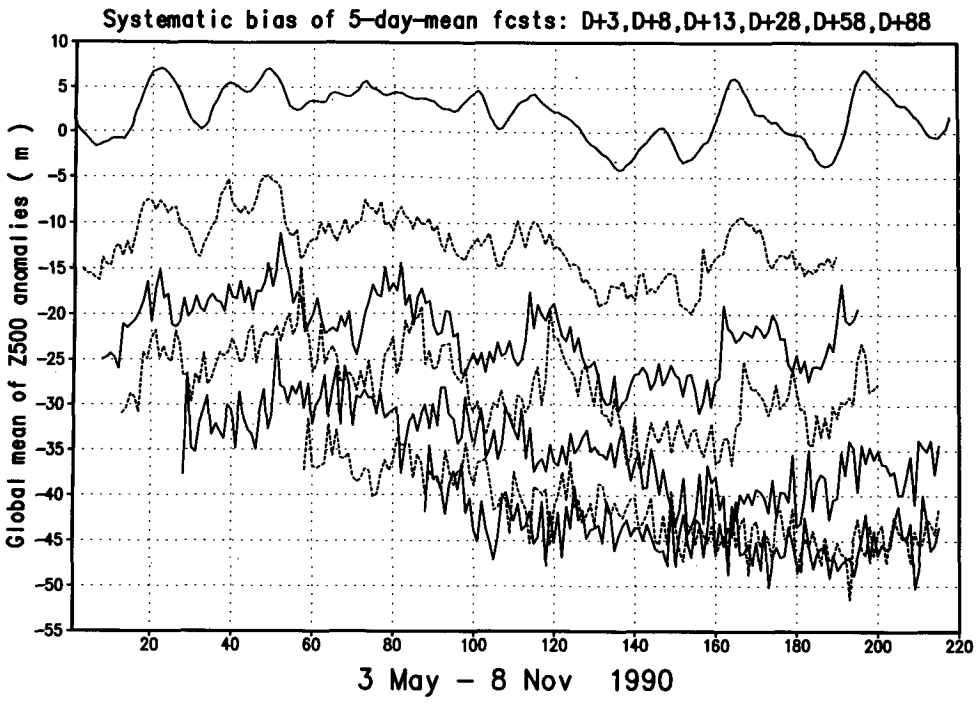

FIG. 2. The Z500 global mean anomaly time series, analyzed (top solid curve) versus forecasts with various leads. The other alternating dashed and solid curves from top to bottom represent $D + 3$, $D + 8$, $D + 13$, $D + 28$, $D + 58$, and $D + 88$, respectively.

ridge over the North Pacific. The model indeed has the dynamical capability to generate its own large positive anomalies and surprisingly has a magnitude comparable to that shown in Fig. 8. The $D + 18$ blocking flow composite is seen to persist for more than 30 days. Note also that in addition to the ridge over the North Pacific the entire four-ridge pattern is being maintained throughout all 30 days.

The next inquiry is about the maintenance of large negative anomalies. Starting from $D + 3$ forecasts and compositing those with anomalies exceeding -100 m, only 23 cases could be found over the North Pacific, versus 36 cases for the large positive anomalies. A highly skewed anomaly distribution is therefore present in the model, as in the real atmosphere (e.g., White 1980; Chen and Van den Dool 1995). This negative composite did not persist (not shown). Starting from $D + 18$ forecasts, only 5 cases of large negative anomalies could be found, versus 30 cases for the positive anomaly shown in Fig. 9, and again, this negative composite did not persist (not shown).

To investigate further, we place our attention on another area. The anomaly amplitude distribution at 50°N, 150°W, another forecast maximum SD center shown in Fig. 5, was also found to be highly skewed, but this area was skewed toward the negative instead of the positive. At this grid point and starting from $D + 18$ forecasts, only 10 cases were found to exceed 100 m but 23 cases exceeded -100 m. While the positive composite did not persist (not shown), the negative composite, as shown in Fig. 10, persisted for more than 30 days.

Persistent flows over the Atlantic sector will be examined in section 6.

From the above results, the model evidently has the dynamical capability to generate and maintain its own positive and negative large-amplitude anomalies. What degrades the model's capability to maintain a large anomaly appears to be the drastic drift and a possible interim dynamical inconsistency that the model has to undergo and adjust to before arriving at its own consistent fluctuation characteristics from the real atmospheric conditions. If this is the case, alleviation of the model's drastic drift becomes an essential task before better performance of the model can be realized.

5. Region and regime dependency of forecast skill

Figure 11 presents the time history of 5-day mean forecast skill measured by anomaly correlations (AC) (e.g., Hollingsworth et al. 1980; Chen 1990). These ACs were evaluated for the Northern Hemisphere (20°–80°N) only. In general, the Southern Hemisphere yields lower skill scores (not shown). At longer forecast leads, the forecast skill in Fig. 11 can be seen to fluctuate widely from day to day but with a visible trace of low-frequency oscillation, implying certain regime dependence. For 203 cases considered here, the average

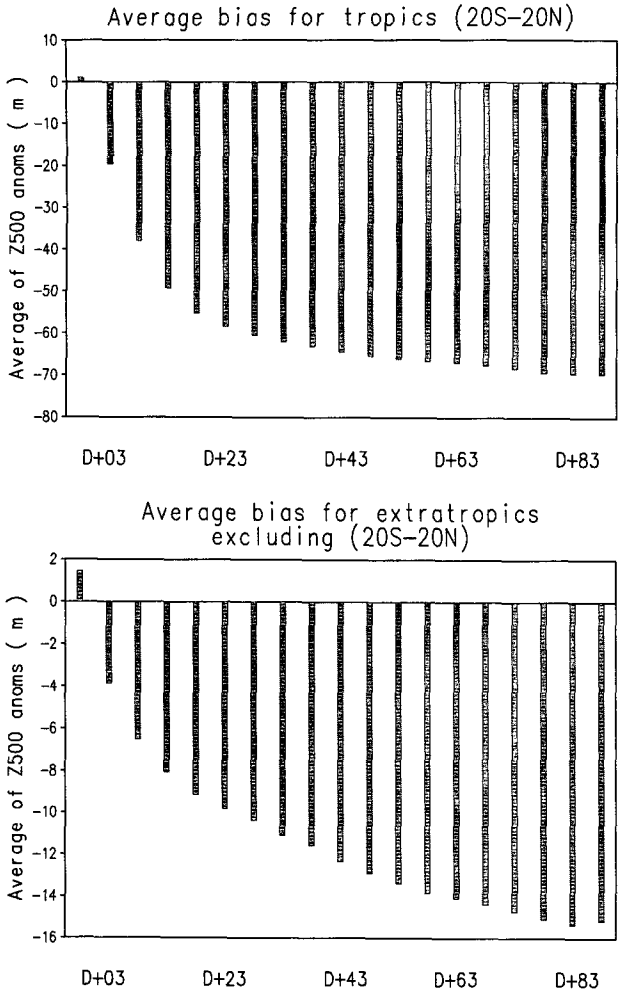

FIG. 3. Zonally and latitudinally averaged systematic bias at various forecast leads, for Tropics and extratropics. Areal weighting with cos(latitude) has been applied. The analyzed anomaly is shown by the leftmost bar.

AC scores are 0.86, 0.37, and 0.12 for $D + 3$, $D + 8$, and $D + 13$ forecasts, respectively. At leads longer than $D + 13$, the 5-day mean forecast skill fluctuates around zero (not shown), implying lack of skill in those long-range forecasts.

Analyzing an earlier DERF experiment, Chen (1992) observed some distinct features in forecast skill between the Pacific sector and the Atlantic sector. The difference was attributed to the influence from the tropical Pacific. This earlier DERF experiment happened to fall within a period when a strong positive phase of ENSO was prevailing. For the present experiment, during which no significant ENSO activity was observed in the Pacific Tropics, we would like to know whether some differences in forecast skill might still exist. The $D + 3$ AC scores between these two sectors are compared in Fig. 12. Here the Pacific sector (PAC) is defined from 20° to 80°N and 120°E and 120°W (solid curve) and the Atlantic sector (ATL) from 20° to 80°N and 80°W to 40°E (dashed curve). Note that both sectors have the same domain size.

At a quick glance, it appears not much difference can be identified except the ATL sector had a few comparatively lower scores. Both sectors show lower scores during the middle of this experiment and much better scores for the last 40 cases, a seasonally dependent feature not uncommon to practicing forecasters. A careful examination, though, reveals some differences in skill characteristics between the PAC and ATL sectors. Greater differences can be detected when the season advances from May to November and from $D + 3$ to $D + 13$ forecasts. Table 1 lists the comparative statistics. Taking $D + 8$ for an example, the PAC average AC score is 0.37 for the whole DERF90 period, increasing to 0.42 for the second half of the experiment and to a fairly significant average score of 0.56 for the last quarter when the season entered into late autumn

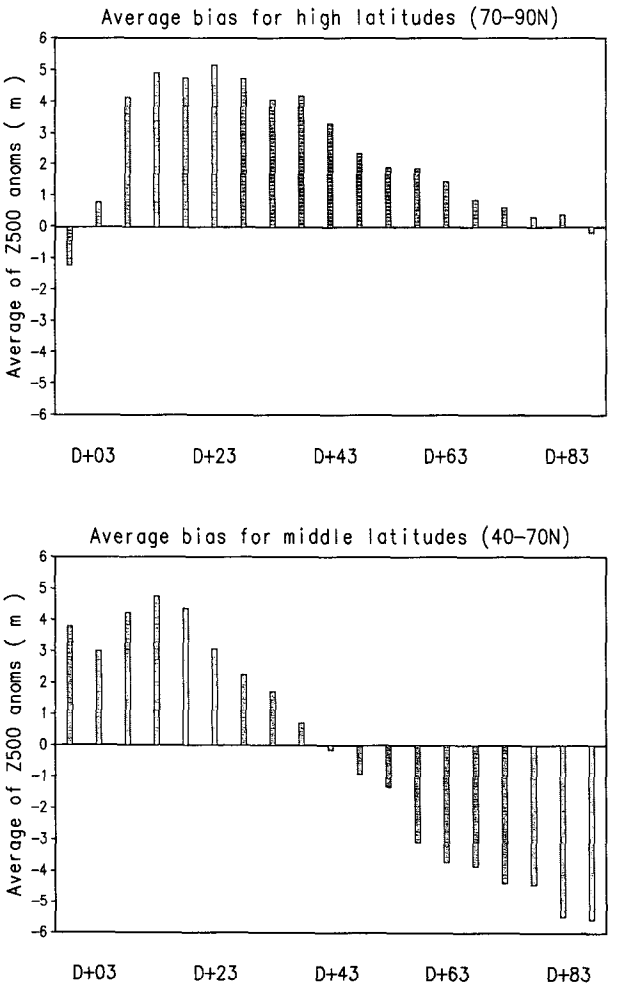

FIG. 4. Similar to Fig. 3 except for high and middle latitudes.

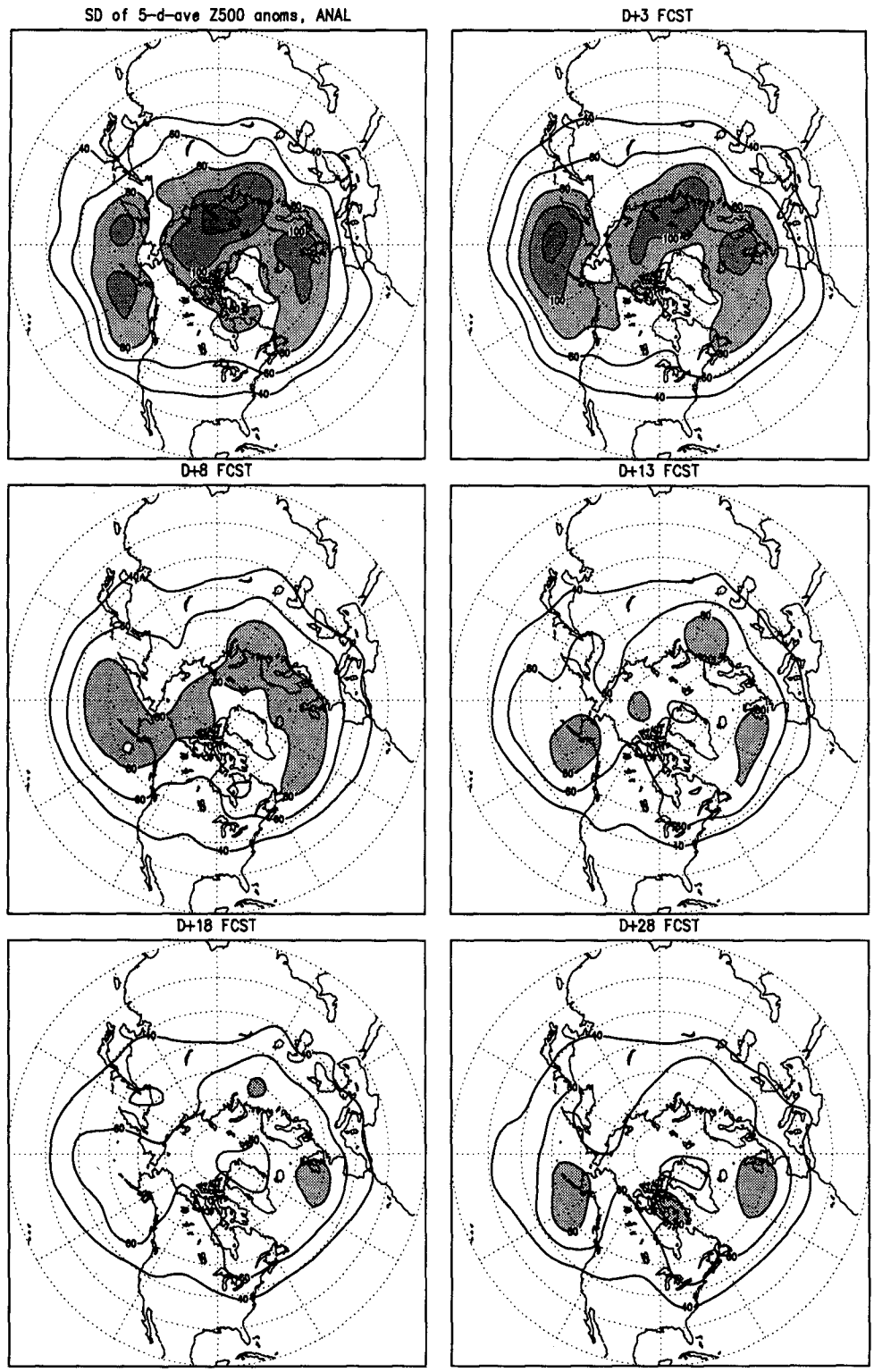

FIG. 5. Distribution of Z500 standard deviations for analysis and forecasts. The contour starts at 40 m with an interval of 20 m. Areas above 80 m are shaded.

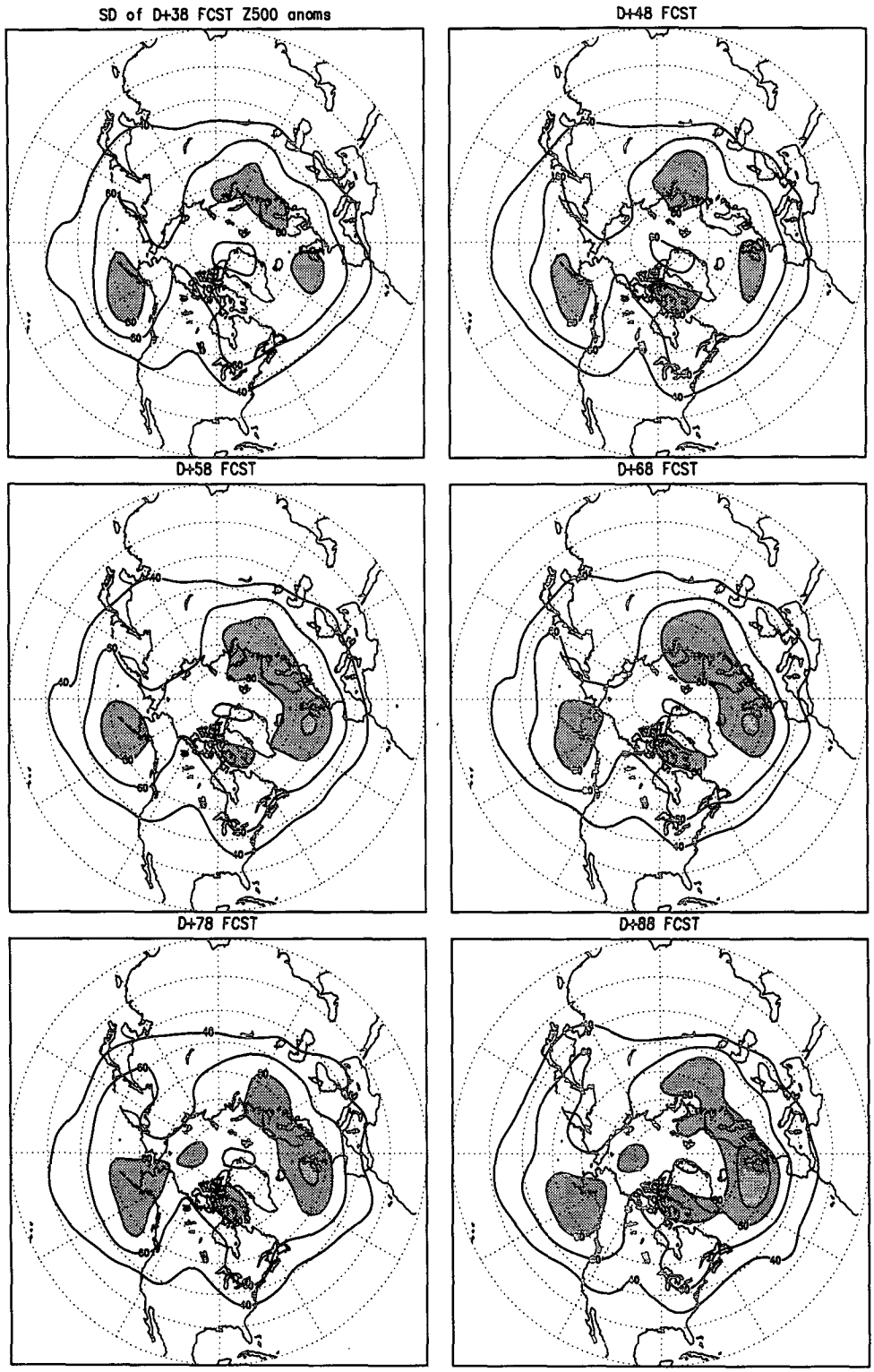

FIG. 5. (*Continued*)

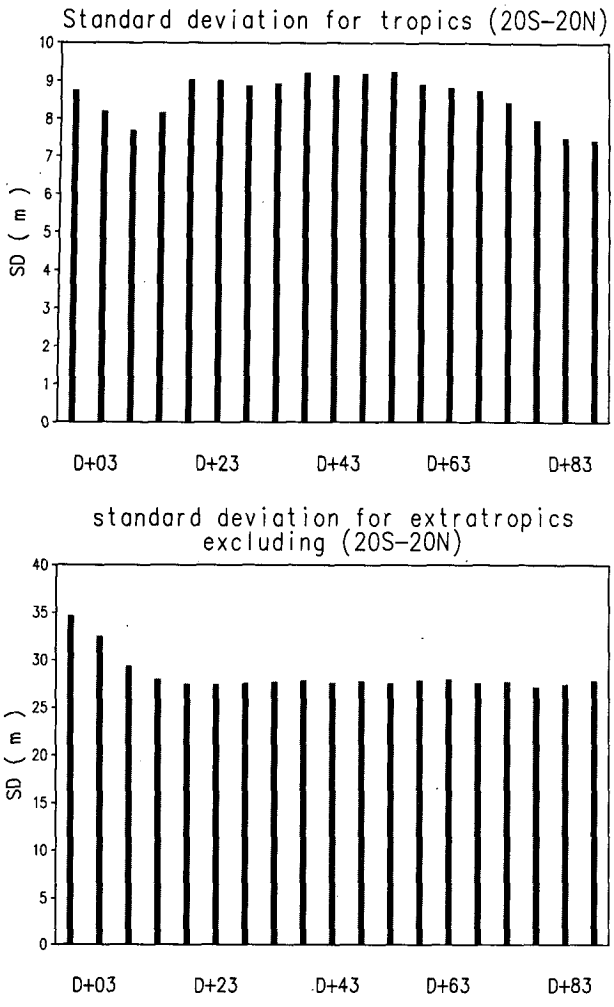

FIG. 6. Zonally and latitudinally averaged standard deviations (SD), separately for Tropics and extratropics. Areal weighting with cos(latitude) has been applied. The analyzed SD is shown by the leftmost bar.

and early winter. Meanwhile, the ATL average $D + 8$ AC scores stayed rather unchanged around 0.3. For the last quarter of the experiment, the long-lead forecasts, $D + 13$, showed some skill though marginal over the PAC sector, while over the ATL sector the scores fluctuated about zero.

Since the last period yielded significant forecast skill at extended ranges for the PAC sector, there was opportunity to examine further some special characteristics, such as whether these extended forecasts were regime dependent (e.g., Molteni et al. 1990a,b). Using the AC scores, Fig. 13 displays and compares the $D + 8$ and $D + 13$ skill scores between the PAC and ATL sectors. As expected from the Table 1 statistics, the PAC sector yielded much better skill than the ATL sector at extended ranges, especially for the last quarter of the DERF90 period, during which the $D + 8$ and $D + 13$ ACs were frequently (rarely) above 0.6 for the PAC (ATL) sector.

Focusing on the PAC sector, there are a few interesting features one can observe. Taking case 190 (i.e., forecasts initiated on 8 November 1990), for instance, both $D + 8$ and $D + 13$ have ACs near 0.85. We would like to know how skillful this forecast is at $D + 18$ or longer. For this case, Fig. 14 displays the predictions (right panels) and their verifications (left panels) from $D + 3$ up to $D + 23$. A large positive anomaly was predicted amazingly well up to $D + 18$, including the weakening and progression during the $D + 8$ period, the maintenance of the positive anomaly, and its retrogression during the $D + 13$ and $D + 18$ periods. The replenishment of the major Pacific high was not predicted at D + 18. Nevertheless, even at $D + 23$, the relative positive–negative anomaly positions were fairly well predicted.

However, it should also be pointed out that its neighboring integrations (i.e., cases 189 and 191 of Fig. 13),

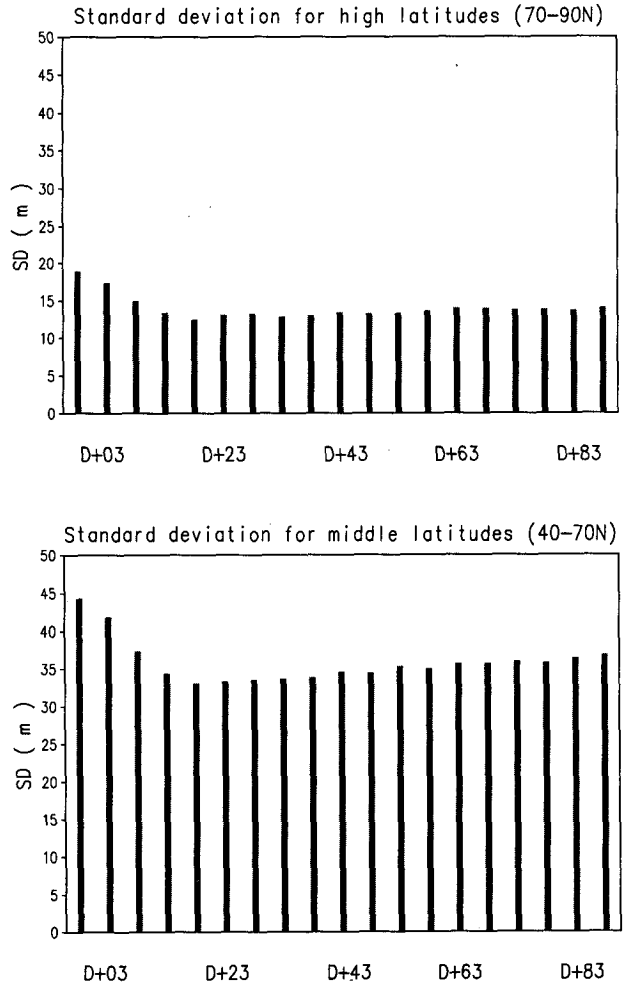

FIG. 7. Similar to Fig. 6 except for high and middle latitudes.

although producing very skillful $D + 8$ forecasts, did not yield good $D + 13$ forecasts. Other cases (184, 186, 201, and 202, for example) also show similarly large day-to-day skill fluctuation in $D + 13$ forecasts. Therefore, the highly skillful $D + 8$ forecasts in these cases are likely not regime dependent and are not something one can rely on from one case to the next. With a slight initial perturbation (butterfly effect of Lorenz) the forecast model though deterministic can lead to a grossly different solution in less than two weeks. This point of view is consistent with the lack of any systematic "return of skill" in DERF90 (Anderson and Van den Dool 1994). In this situation, the ensemble forecast approach (Tracton and Kalnay 1993) should be quite helpful.

The upper panel of Fig. 13 also reveals another interesting phenomenon. The forecast skill at extended ranges can sometimes depend on what we loosely call the "weather regime," and all forecasts can have similar skill. Note that for 10 consecutive days, from case 159 up to case 168 (8–17 October initial conditions), all of the $D + 8$ ACs and more than half of the $D + 13$ ACs were considerably above 0.5, but a short while later, beginning on 22 October (case 173), the $D + 8$ forecasts began to encounter great difficulty and continued to have problems for at least 10 days. The corresponding $D + 13$ forecasts fluctuated only about zero. Figure 15 contrasts the AC scores obtained from these two weather regimes with forecasts initiated on 8–12 October shown in the upper panel and those initiated on 24–28 October in the lower panel. While most of the upper-panel ACs stayed above 0.5 beyond day 15, all of the lower-panel ACs dropped below 0.5 at about day 7. The regime dependency of these consecutive forecasts is clearly illustrated. Sensitivity to the initial error may be, relatively speaking, of smaller importance during both of these regimes. Nevertheless, the ensemble forecast approach could still be useful, because from the magnitude of spread among members at extended ranges it may tell in advance whether the solution is to be trusted or distrusted.

6. Systematic forecast error over the North Atlantic

We further investigated forecast performance and a recurrent error for the ATL sector. A long-lasting blocking flow prediction over the North Atlantic was found in the forecasts initiated on 6 November 1990. The right-hand panels of Fig. 16 present the forecasts with increasing forecast lead time up to $D + 28$. The verifications are shown on the left-hand panels. The $D + 3$ forecast correctly predicted a strong blocking flow between Greenland and Norway. The blocking flow was maintained in the forecasts, while in the real atmosphere it disintegrated during the following week. In the forecasts, the block retrogressed and increased in magnitude. On $D + 28$, the positive height anomaly in the forecast split into two parts, with its main portion over the area north of Hudson Bay and the smaller portion over the East Siberian Sea (not shown). During this period, the real atmosphere had a new and strong blocking anticyclone established south of Greenland (the third and fourth left panels). It then moved slowly eastward over the North Atlantic. However, the model entirely missed this newly developed block. So, what we see in this case is that the model was able to maintain a blocking flow for more than 30 days, but this large positive anomaly evolved incorrectly and moved to the wrong location. The associated predicted weather events would have been grossly different from the observed.

The question is, Does this kind of forecast error occur often? And most importantly, does it show up in consecutive forecasts so that an ensemble forecast formed by them would also contain the same error? Figure 17 presents four adjacent integrations: three immediately prior to and one immediately after the 6 November initial conditions. The first column shows the initial conditions, the second column the verifications for the $D + 8$ forecasts, and the third column the $D + 8$ forecasts. In the initial conditions, a major blocking high was positioned southeast of Greenland on 3 November. The block strengthened and progressed slightly eastward over the ensuing four days. With these initial conditions, every one of the five $D + 8$ forecasts predicted retrogression of the blocking flow toward Greenland or even farther west to Baffin Island, while the verifications showed weakening and then disintegration of the blocking high. So during this particular period the forecasts showed a consistent error in the sense that the blocking flow retrogressed to Greenland or farther west and the block remained as strong or became even stronger. This forecast error is consistent with the results we see in Fig. 1, where the long lead forecasts consistently showed much larger positive height anomalies over the Baffin Island area while the NMC analyses showed negative anomalies.

For completeness, we searched through all 188 cases of $D + 3$ and their $D + 8$–$D + 28$ forecasts for persistent blocking flows. The same procedures used in section 4 were adopted here. That is, a composite of $D + 3$ forecast anomalies exceeding 100 m at 65°N, 20°W was constructed first. Then the evolution of the composite of the chosen cases was followed for the $D + 8$ through $D + 28$ forecasts. Figure 18 shows the evolution of the blocking high present at 65°N, 20°W at $D + 3$, contrasted with the verifications. At $D + 8$, while the real atmospheric ridge progresses toward Europe the forecast blocking high retrogresses to Greenland. At $D + 13$, the verification shows only a weak high, while the forecast blocking flow remains almost as strong and keeps retrogressing toward the Baffin Bay area. The $D + 18$ through $D + 28$ forecasts show traces of blocking high maintenance, although its strength becomes weaker progressively. The remnant eventually spreads over both sides of the North Pole.

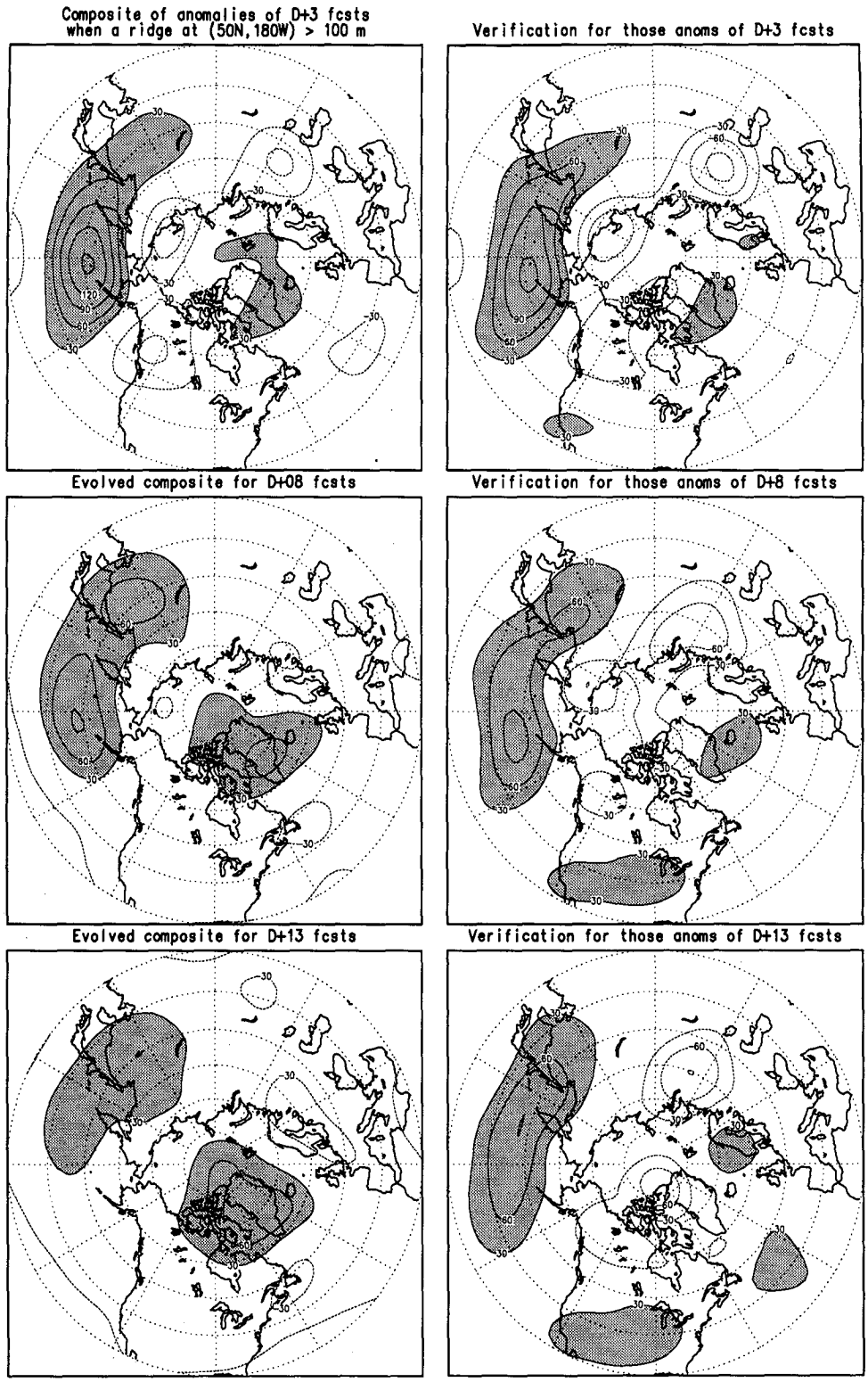

FIG. 8. For top two panels, composite of $D + 3$ forecasts for those with a positive anomaly at grid point 50°N, 180° exceeding 100 m, and its verification shown on the right-hand side. The evolution of those forecasts at longer leads and their verifications are shown by the subsequent panels. The composite with the net positive anomaly is shaded. The contour interval is 30 m.

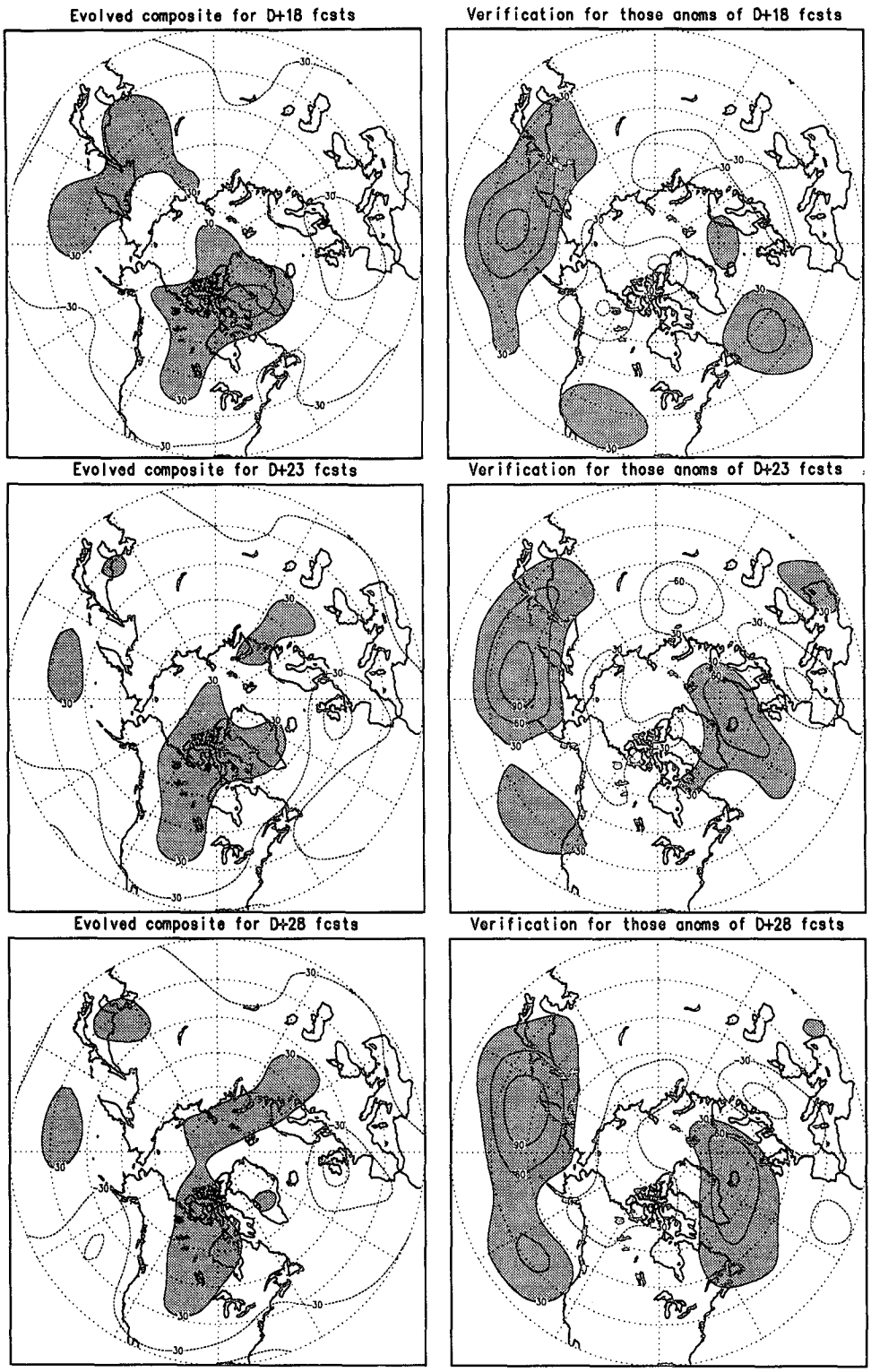

FIG. 8. (*Continued*)

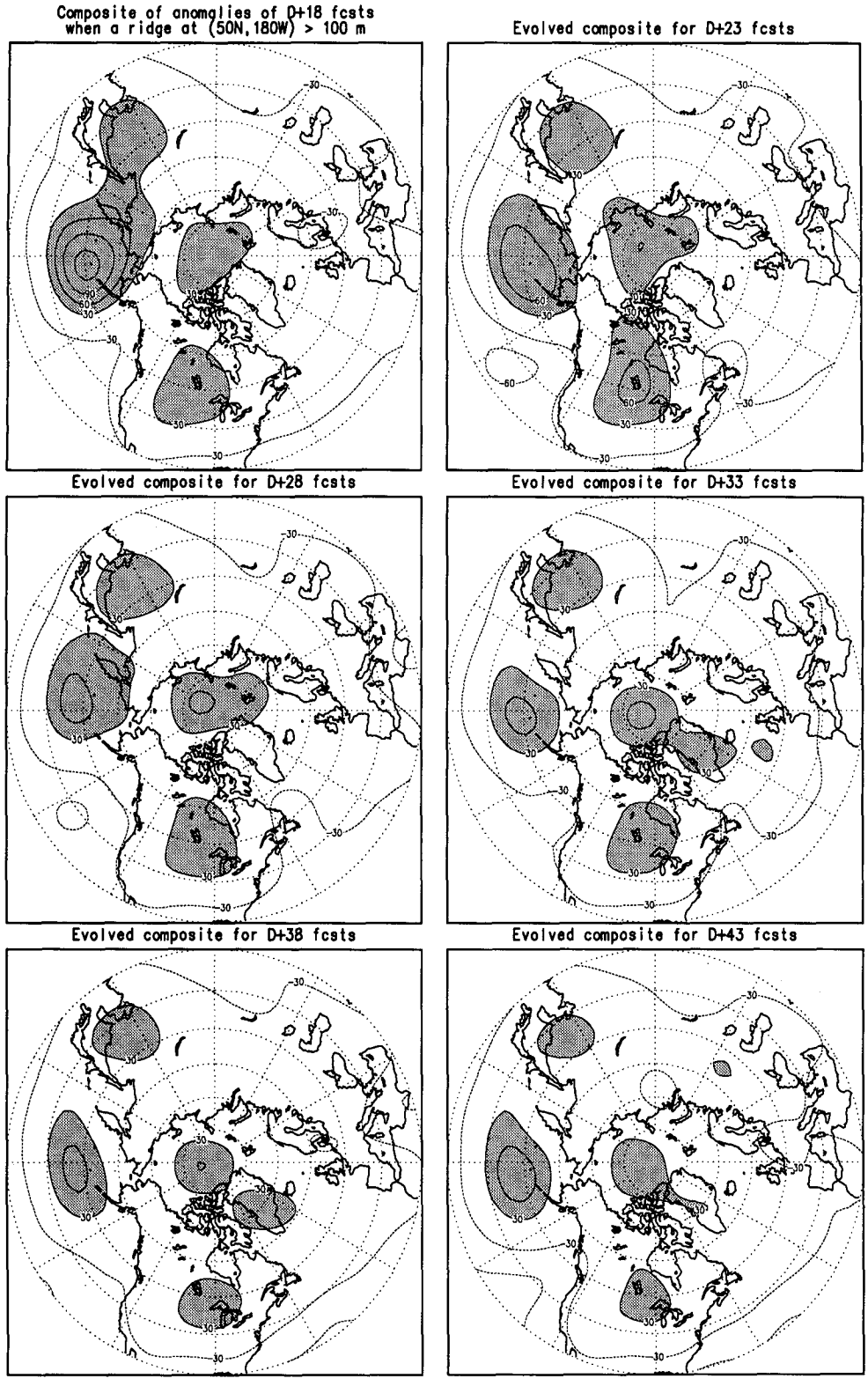

FIG. 9. Similar to Fig. 8 except using $D + 18$ forecasts instead of $D + 3$ forecasts to construct the composite. The temporal evolutions of the selected forecasts from $D + 23$ up to $D + 43$ are also shown.

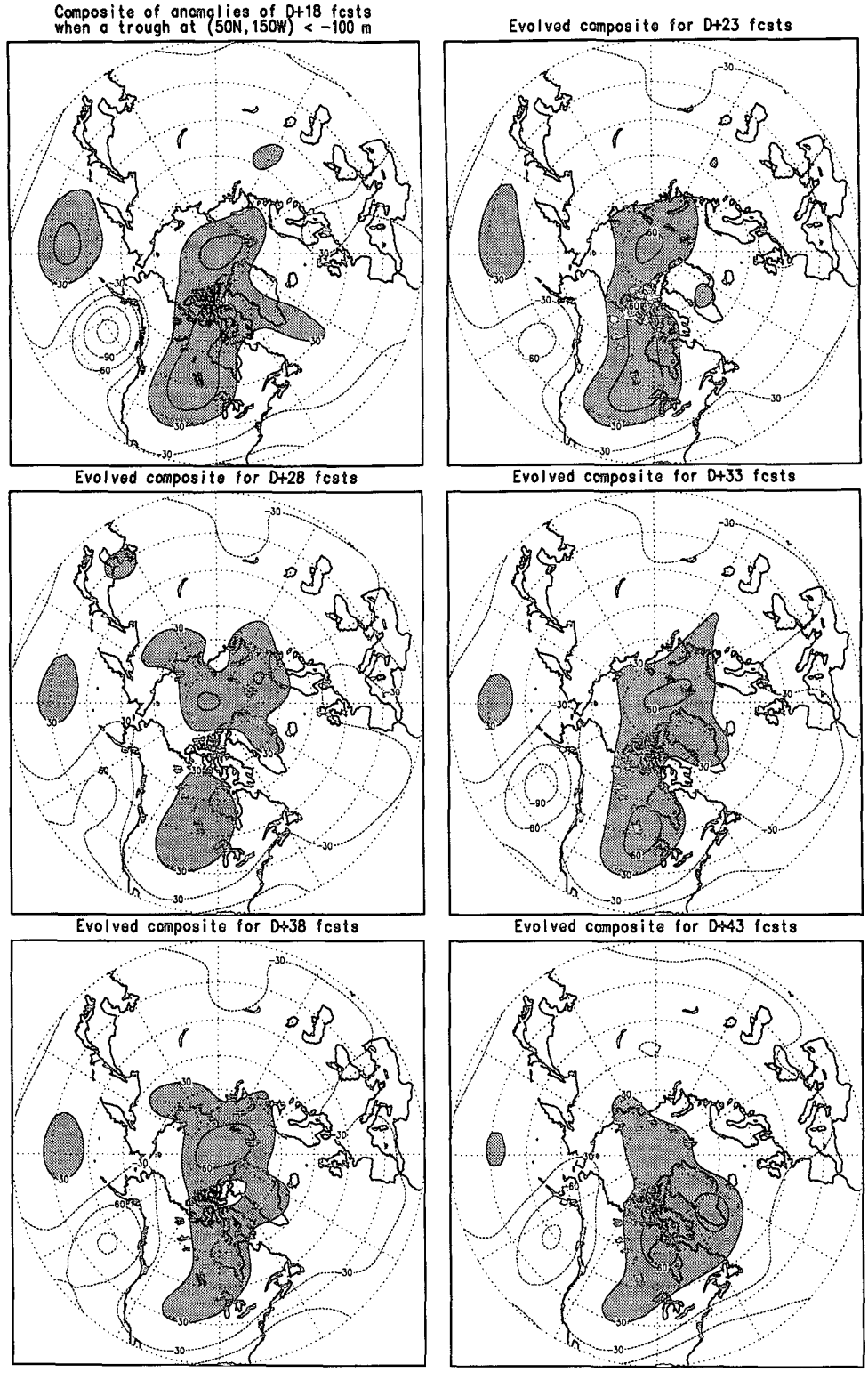

FIG. 10. Similar to the Fig. 9 composite except for a negative anomaly at grid point 50°N, 150°W that exceeds −100 m, showing persistence of large negative anomalies.

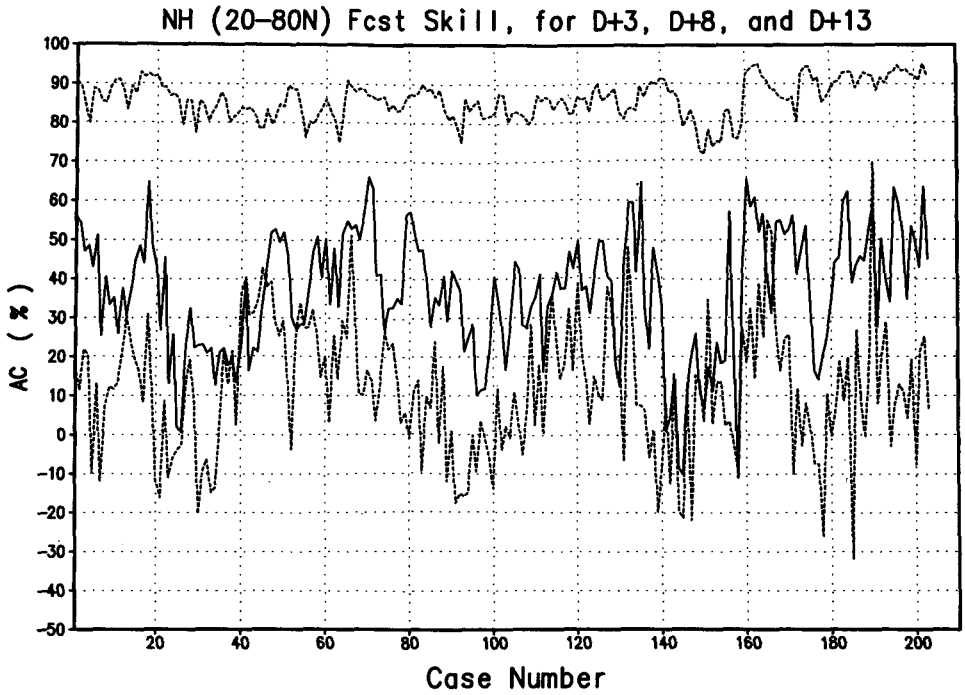

FIG. 11. Time history of 5-day mean forecast skill at $D + 3$ (top dashed curve), $D + 8$ (solid curve), and $D + 13$ (lower dashed curve). Case 1 was initiated on 3 May 1990.

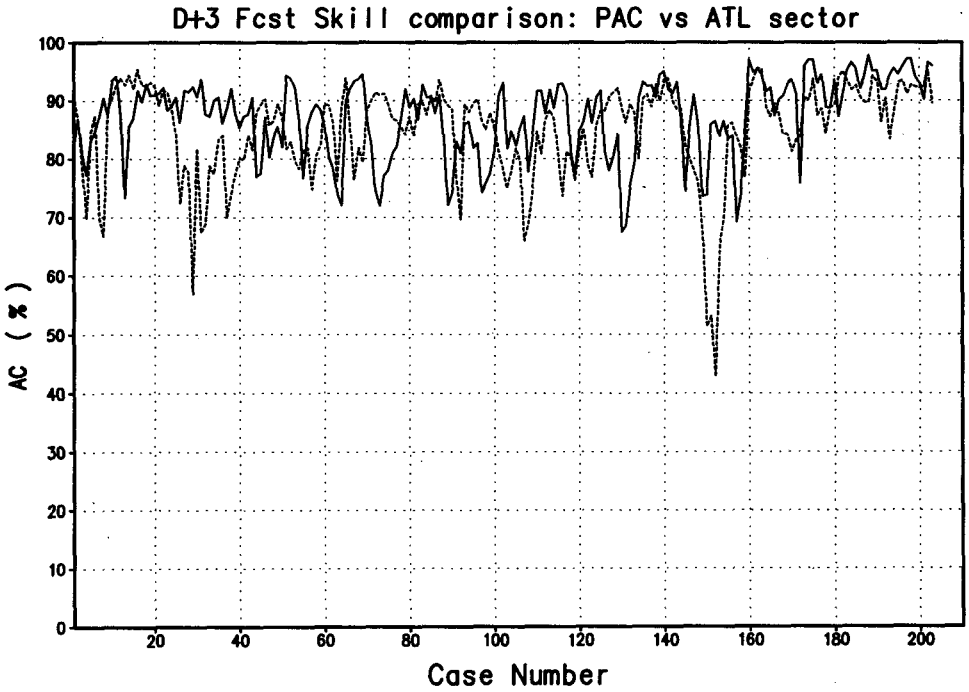

FIG. 12. Comparison of $D + 3$ forecast skill between PAC sector (20°–80°N, 120°E–120°W) and ATL sector (20°–80°N, 80°W–40°E); solid curve for PAC sector and dashed curve for ATL sector.

TABLE 1. Comparisons of anomaly correlation scores between PAC and ATL sectors. Anomaly correlations evaluated separately for the whole DERF90 period (cases 1–203) and the second half (case 101–203) and last quarter (cases 152–203) of the DERF90 period.

Average AC over	$D+3$		$D+8$		$D+13$	
	PAC	ATL	PAC	ATL	PAC	ATL
Whole period	0.87	0.85	0.37	0.34	0.14	0.13
Second half	0.89	0.85	0.42	0.28	0.17	0.11
Last quarter	0.92	0.88	0.56	0.30	0.28	0.03

From the previous three figures, the implication is that an ensemble forecast, consisting of nine integrations with initial conditions 6 h apart or otherwise, will also contain the same error described above for a North Atlantic blocking flow. In this situation, we suspect that the ensemble forecasts may not be of great help. Consistency does not necessarily offer a skillful forecast. In this case, it is the improvement of the model's systematic error that appears to be the essential factor for achieving a much more skillful forecast at extended ranges.

7. Summary

A series of 90-day integrations by a low-resolution version (T40) of the NMC's global spectral model was analyzed for its forecast performance and low-frequency variability behavior. Five-day mean forecasts at the 500-mb level with lead times up to 88 days were examined. During the period investigated, which is 3 May–6 December 1990, the forecast Z500 mean height decreased rapidly as the forecast lead time increased. The cold bias was particularly severe in the Tropics. The forecast variance of the 5-day mean fluctuations also showed a rapid drop to only 75% of the observed standard deviation at day 18. Beyond this lead time the forecast variance stabilized, showing only slight and gradual increases for the northern latitudes. Careful examination reveals that the swift decrease was mainly associated with those of the large-amplitude fluctuations. The small-amplitude fluctuations in the forecasts showed very little decrease.

Large-amplitude anomalies, both positive and negative, could not be maintained by the model from the real atmospheric initial conditions. However, after a drastic drift to its own climatology and variability characteristics, the model was capable of generating large-

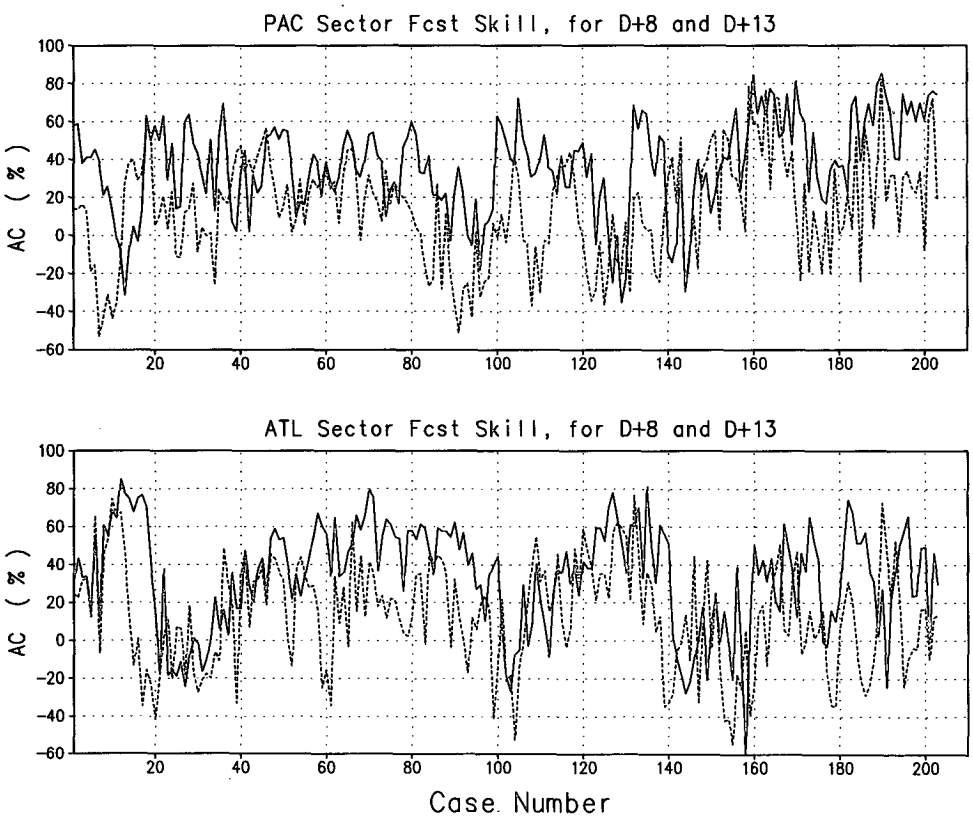

FIG. 13. Comparison of AC skill scores between PAC and ATL sectors. The solid curves are for $D+8$ forecasts and the dashed curves are for $D+13$ forecasts.

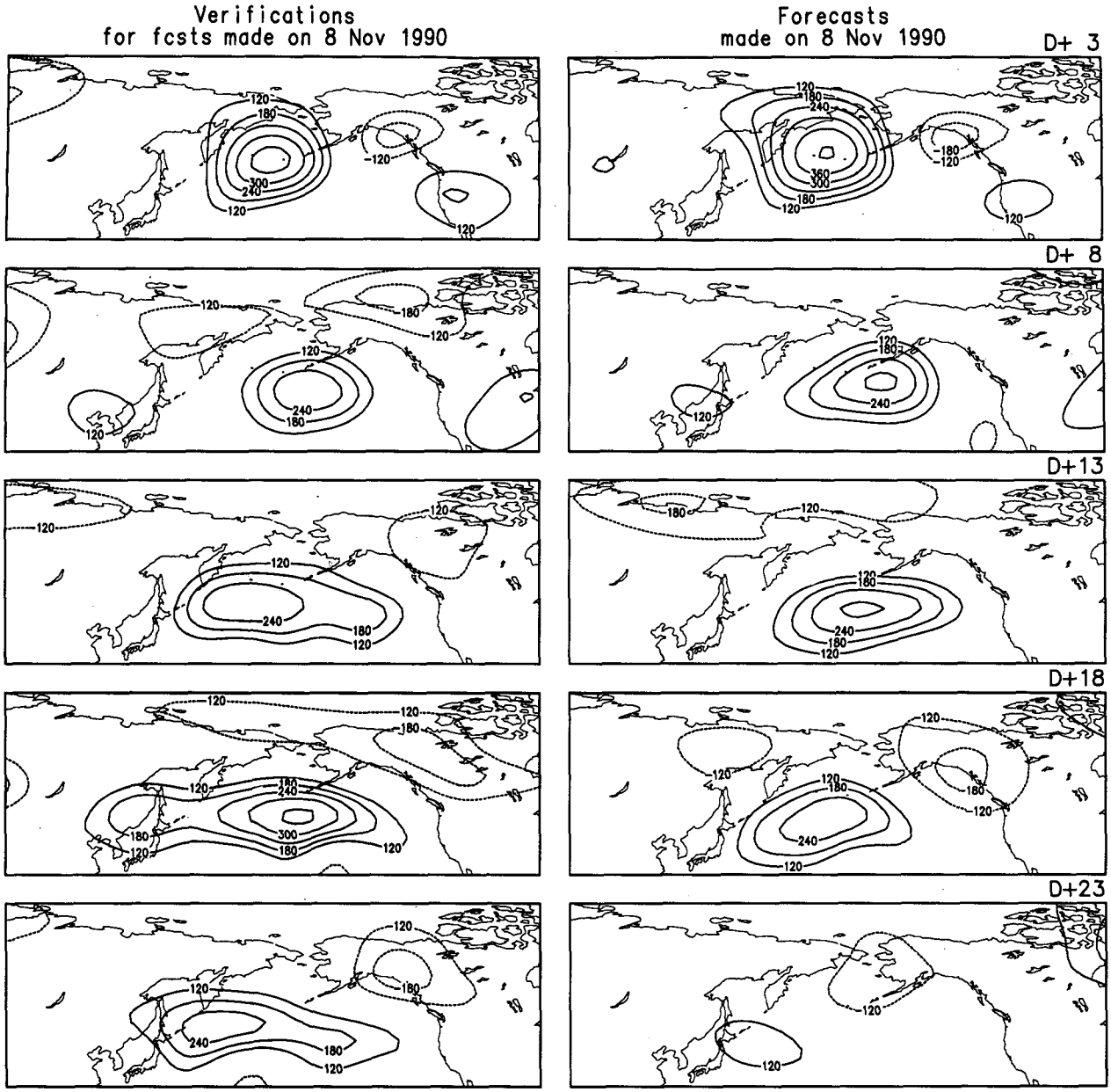

FIG. 14. Forecasts up to $D + 23$ made on 8 November 1990 and their verifications. The contour starts at 120 m with an interval of 60 m. The solid contours are for positive anomalies and the dashed contours are for negative anomalies.

amplitude anomalies and maintaining them for longer than 30 days. Over the North Atlantic, although the model is also capable of maintaining large anomalies, it tends to systematically retrogress the large blocking ridges westward to the Baffin Island area, resulting in large forecast error at extended ranges over the North Atlantic. Therefore, preventing the model from drifting away from the atmospheric characteristics of climatology and variability appears to be an essential task for achieving much more skillful forecasts at extended ranges.

The model yields better forecast skill for the North Pacific sector than the North Atlantic sector when the season advanced to late fall and early winter. In this regard, the result is in good agreement with an earlier DERF experiment conducted in late 1986 to early 1987 when a warm phase of ENSO was prevailing (Chen 1992). At extended ranges, the model's skill fluctuated in general with a very wide range, yielding sometimes very high skill score (AC = 0.85) up to $D + 13$, while the forecasts initiated on its previous day and the day after showed very little skill. Under these nonregime-

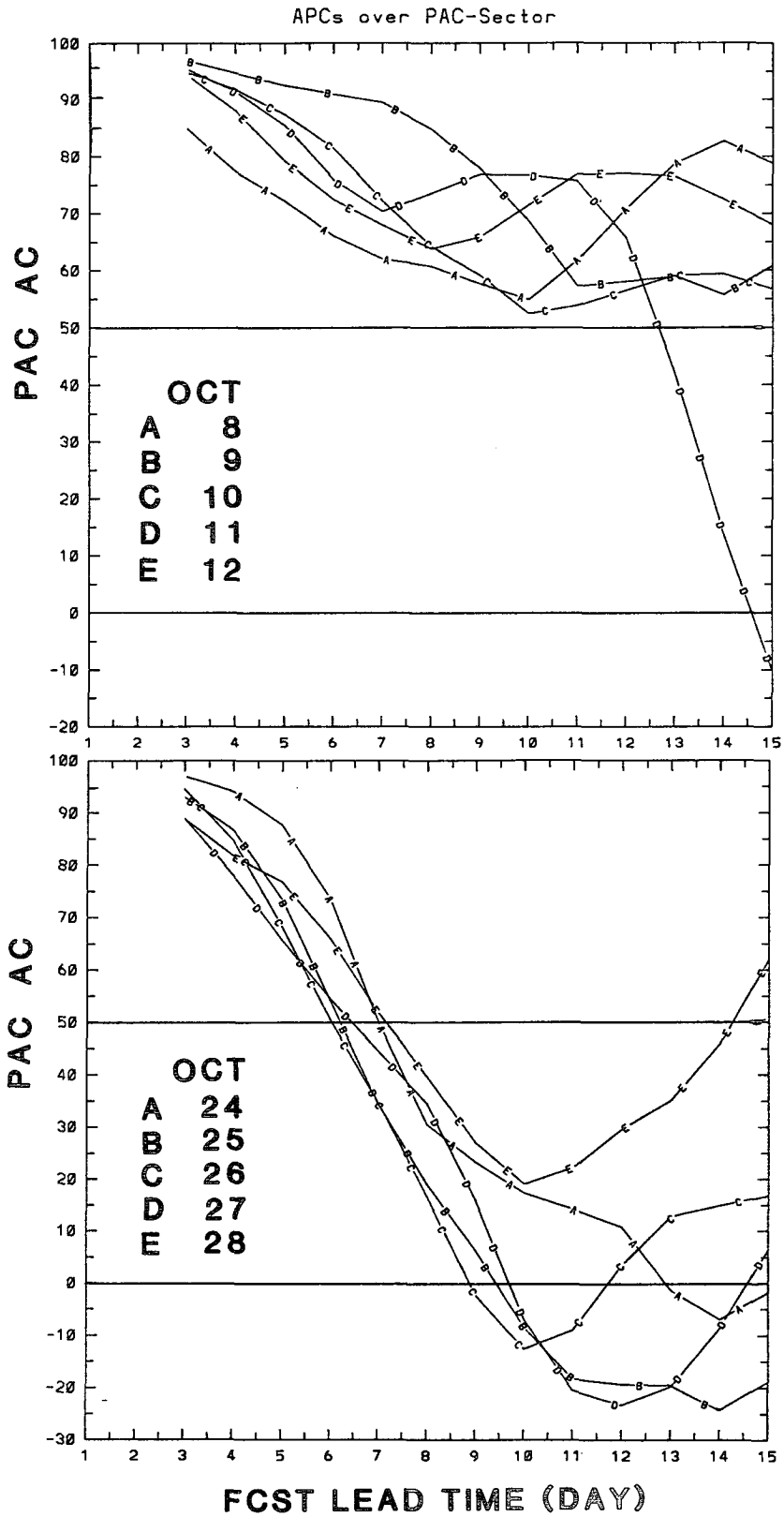

FIG. 15. Comparisons of forecast skill scores obtained from favorable (upper panel) and unfavorable (lower panel) circulation regimes.

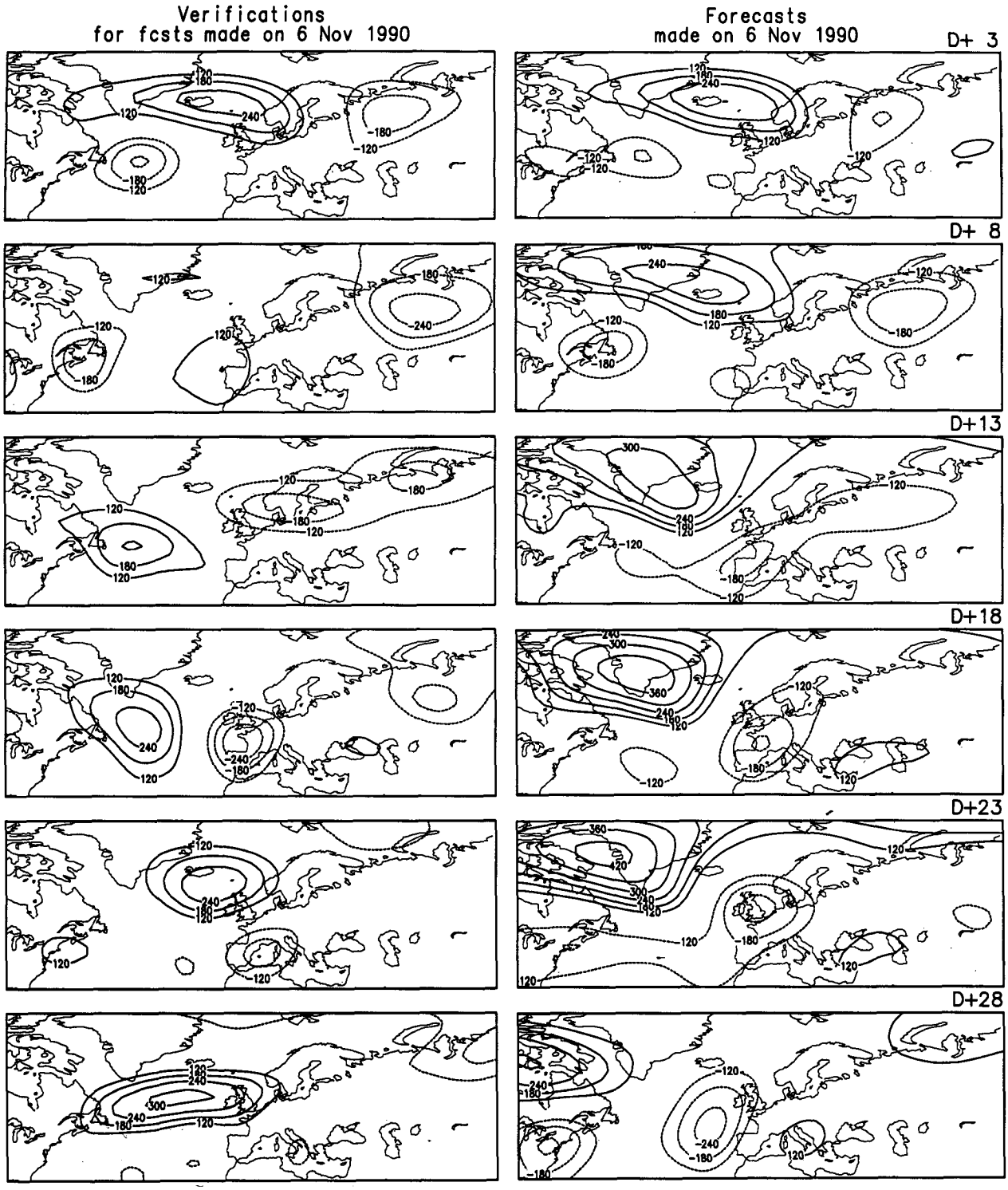

FIG. 16. Forecasts up to $D + 28$ made on 6 November 1990 and their verifications. The contour starts at 120 m with an interval of 60 m. The solid contours are for positive anomalies and the dashed contours are for negative anomalies.

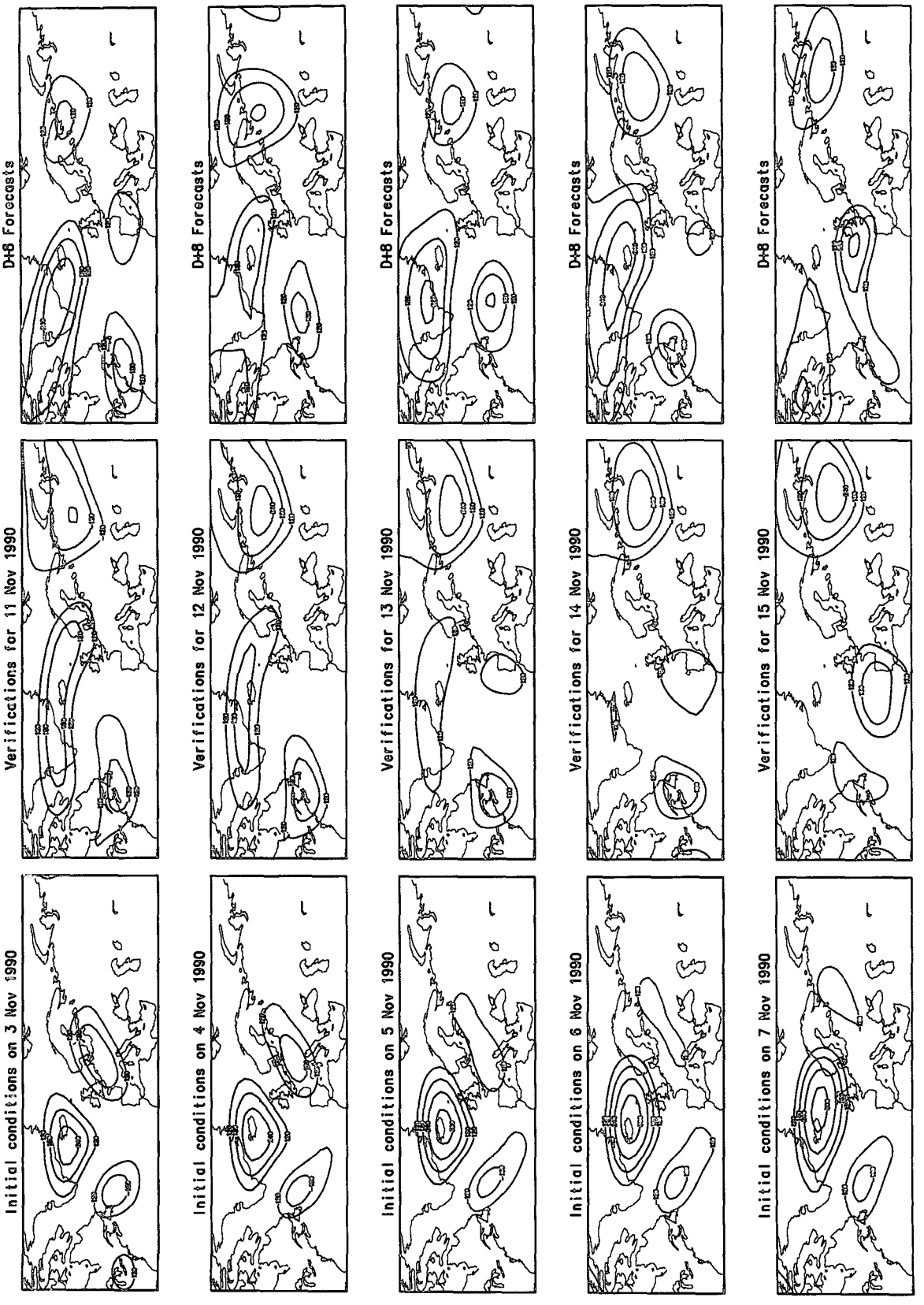

FIG. 17. *Still another five consecutive D + 8 forecasts, made from 3 to 7 November 1990, and their verifications and initial conditions.*

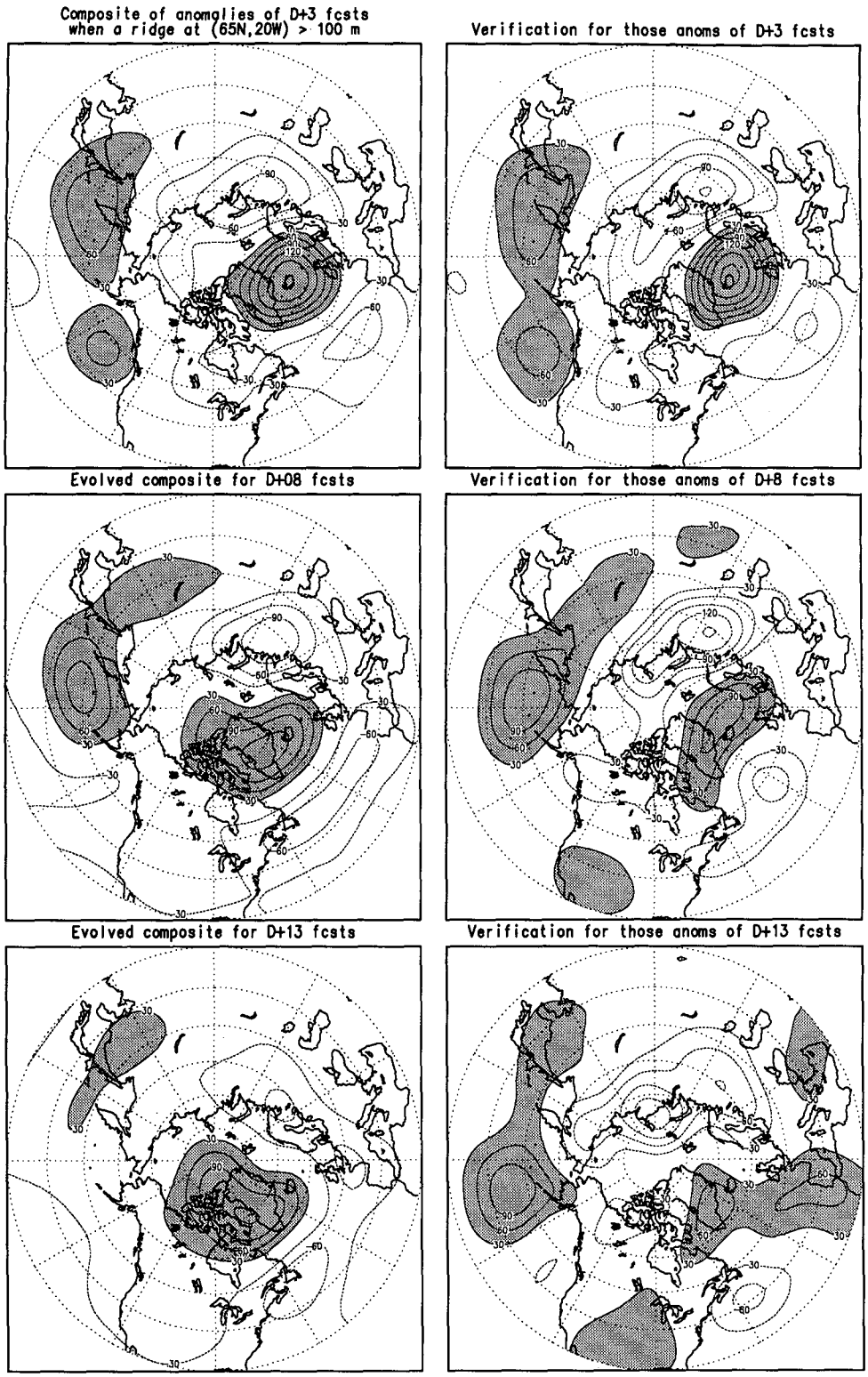

FIG. 18. Similar to Fig. 8 except for the composite of the large positive anomalies at grid point 65°N, 20°W, showing retrogression of this persistent composite.

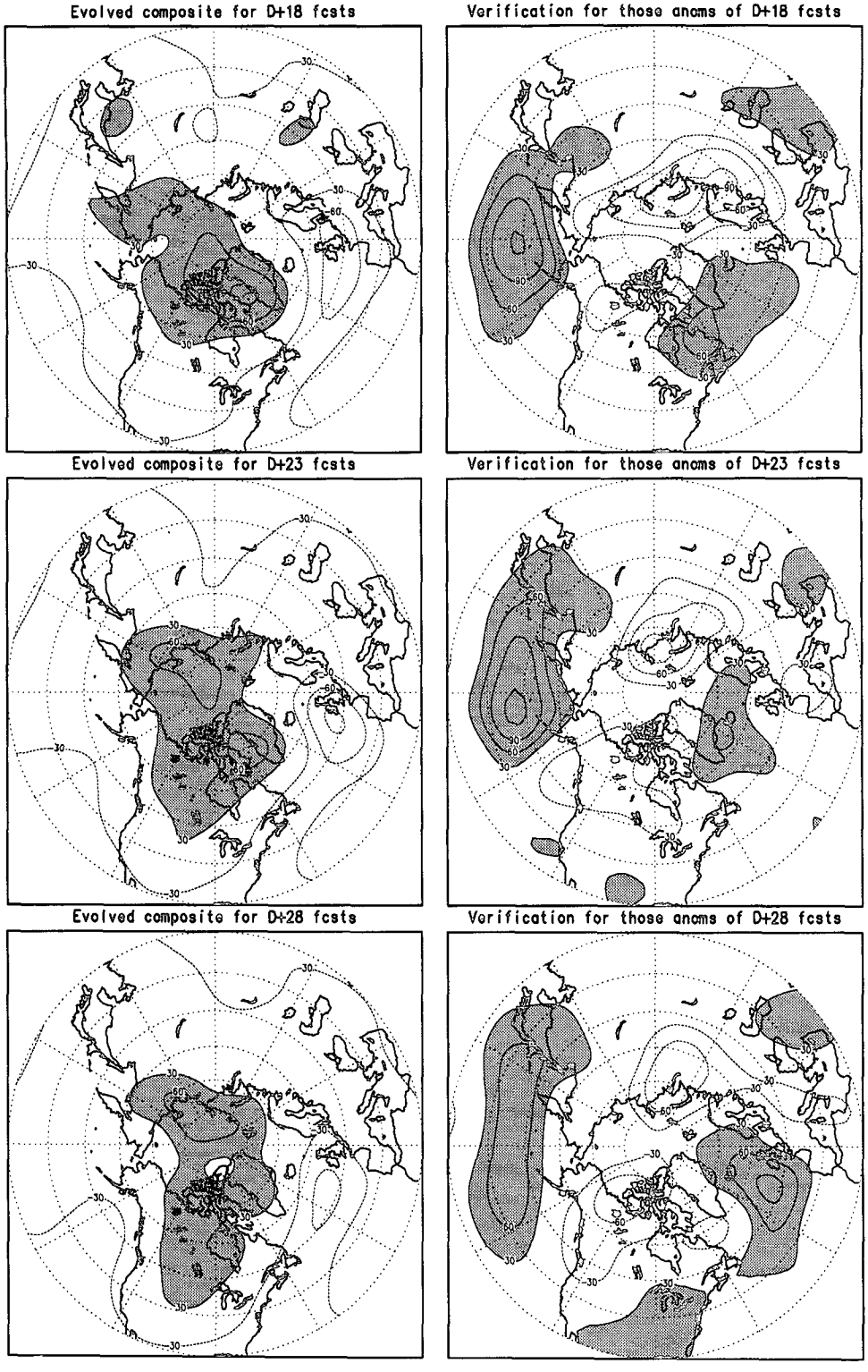

FIG. 18. (*Continued*)

dependent situations, the ensemble forecast should be of great utility. At other times, the model's skill does exhibit strong dependency on the atmospheric circulation regime, producing very high forecast skill for a string of consecutive days but at other times very little skill for a long period of time. In these regime-dependent situations, the ensemble forecast could also be useful. It might tell in advance whether the solution is to be trusted or not.

On the other hand, if the model keeps generating error in a systematic way, such as the erroneous retrogression of an Atlantic block toward the Baffin Island area, the usefulness of an ensemble forecast could be drastically curtailed. A reduction in the model's erroneous displacement of these large positive anomalies must be achieved before significant improvement in the extended-range forecasting can be realized.

The main purpose of this investigation is to become more familiar with the behavior of this NMC global spectral model, not the atmosphere. These experiments were run with a low-resolution version of the model over a relatively short span of time. Therefore, we have to stress the limited usefulness of these results. They might be suitable only for this version and resolution of the NMC MRF model. A generalization of the results may not be proper, and the usage of the results must be applied with caution.

Acknowledgments. Without the dedication of many NMC staff members to produce the DERF90 experiment, this investigation would not have been possible. The authors are also grateful to Drs. Edward Epstein, Steve Tracton, and David Walker for their helpful comments of an earlier version of this paper. This work also benefited tremendously from the suggestions and constructive criticism of two anonymous reviewers as well as Professor Fred Sanders.

REFERENCES

Anderson, J. L., 1993: The climatology of blocking in a numerical forecast model. *J. Climate,* **6,** 1041–1056.

——, and H. M. Van den Dool, 1994: Skill and return of skill in dynamic extended range forecasts. *Mon. Wea. Rev.,* **122,** 507–516.

Chen, W. Y., 1989: Estimate of dynamical predictability from NMC DERF experiments. *Mon. Wea. Rev.,* **117,** 1227–1236.

——, 1990: Interannual variability of skill of NMC medium-range forecasts over the Pacific/North America sector. *Mon. Wea. Rev.,* **118,** 179–188.

——, 1992: Dynamical prediction at medium/extended ranges employing low-frequency PNA mode. *Mon. Wea. Rev.,* **120,** 2641–2652.

——, and H. M. Van den Dool, 1995: Low-frequency anomalies in the NMC MRF model and reality. *J. Climate,* **8,** 1369–1385.

Dole, R. M., and N. D. Gordon, 1983: Persistent anomalies of the extratropical Northern Hemisphere wintertime circulation: Geographical distribution and regional persistence characteristics. *Mon. Wea. Rev.,* **111,** 1567–1586.

Hollingsworth, A., K. Arpe, M. Tiedtke, M. Capaldo, and R. H. Savijarvi, 1980: The performance of a medium-range forecast model in winter: Impact of physical parameterization. *Mon. Wea. Rev.,* **108,** 1736–1773.

Kalnay, E., M. Kanamitsu, and W. E. Baker, 1990: The NMC global forecast system. *Bull. Amer. Meteor. Soc.,* **71,** 1410–1428.

Kanamitsu, M., 1989: Description of the NMC Global Data Assimilation and Forecast System. *Wea. Forecasting,* **4,** 335–342.

Molteni, F., S. Tibaldi, and T. N. Palmer, 1990: Regimes in the wintertime circulation over northern extratropics. I: Observational evidence. *Quart. J. Roy. Meteor. Soc.,* **116,** 31–68.

——, and S. Tibaldi, 1990: Regimes in the wintertime circulation over northern extratropics. II: Consequences on dynamical predictability. *Quart. J. Roy. Meteor. Soc.,* **116,** 1263–1288.

Roads, J. O., 1987: Predictability in the extended range. *J. Atmos. Sci.,* **44,** 3405–3527.

——, 1988: Dynamical extended-range forecasts of the lower tropospheric thickness. *Mon. Wea. Rev.,* **117,** 3–28.

Rukhovets, L. V., H. M. van den Dool, and A. G. Barnston, 1995: Principal component and singular value decomposition analysis of a series of 90 day forecasts. *J. Climate,* submitted.

Sela, J. G., 1980: Spectral modeling at the National Meteorological Center. *Mon. Wea. Rev.,* **108,** 1279–1292.

Tibaldi, S., and F. Molteni, 1990: On the operational predictability of blocking. *Tellus,* **42A,** 343–365.

Toth, Z., E. K. Kalnay, and S. Saha, 1991: A study of the predictability of the 20–60 day waves using the DERF90 dataset of forecasts. *Proc. Ninth Conf. on Numerical Weather Prediction,* Denver, CO, Amer. Meteor. Soc., 771–774.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting,* **8,** 379–398.

——, K. Mo, W. Chen, E. Kalnay, R. Kistler, and G. White, 1989: Dynamical extended range forecasting (DERF) at the National Meteorological Center. *Mon. Wea. Rev.,* **117,** 1604–1635.

Van den Dool, H. M., 1994: Long range weather forecasts through numerical and empirical methods. *Dyn. Atmos. Oceans,* **20**(3), 247–270.

——, S. Saha, and Z. Toth, 1991: The climate in a multi-year NMC model run. *Proc. Fifth Conf. on Climate Variations.* Denver, CO, Amer. Meteor. Soc., 511–514.

——, ——, A. H. Oort, and W. Ebizusaki, 1993: On the role of atmospheric water in the continuity equation. *Proc. 18th Annual Climate Diagnostics Workshop,* Boulder, CO, Climate Analysis Center, 244–247.

White, G. H., 1980: Skewness, kurtosis and extreme values of Northern Hemisphere geopotential heights. *Mon. Wea. Rev.,* **108,** 1446–1455.