DYNAMICAL EXTENDED RANGE FORECASTING (DERF) AT THE NATIONAL METEOROLOGICAL CENTER

by

M. Steven Tracton, Kingtse Mo, Wilbur Chen
Climate Analysis Center
and
Eugenia Kalnay, Robert Kistler, Glenn White
Development Division

National Meteorological Center, NWS/NOAA, Washington, D.C. 20233

1. INTRODUCTION

The National Meteorological Center (NMC), Washington, D.C., is engaged in a program to assess the feasibility of providing more useful operational monthly predictions by extension of the basic model used for medium range forecasts (MRF). The rationale for this effort in Dynamic Extended Range Forecasting (DERF) is twofold. First, a number of prediction and predictability studies (e.g., Miyakoda, 1986, and Shukla, 1981) have shown that useful information can be extracted from dynamic models beyond 10 days, at least in some cases. Second, models and computer power have improved enough to make feasible the extensive testing and experimentation which is required. The principal motivation for this effort is the fact that at present monthly forecasts have only marginal utility, so that even modest gains from DERF are potentially significant (Kalnay and Livezey, 1985). Additionally, it is expected that extended integrations will provide feedback to further development of the MRF through, for example, increased understanding of the evolution of the model's climate drift.

The basic elements of NMC's DERF program are i) model development, ii) generation of a data base of experimental extended range predictions, and iii) evaluation of results. With respect to the model, the strategy we have followed has been to utilize the latest available version of the MRF in the the basic experiments with comparisons with alternative configurations when

TABLE I. Medium Range Forecast (MRF) Model Description for DERF Phase II Experiment.

RESOLUTION:
Rhomboidal 40
18 unequally spaced layers

OROGRAPHY:
Silhouette mountains

CUMULUS CONVECTION:
Deep convection (Kuo, Anthes)
Shallow Convection (Tiedke)

LARGE SCALE CONDENSATION:
100% Criterion
Modified Kessler evaporation

AIR-SURFACE INTERACTION:
Analysis of SST (fixed anomaly)
Predicted land temperature

RADIATION:
Shortwave (Lacis/Hansen)
Longwave (Fels/Schwartzkopf-GFDL)
Diurnal cycle
3-layer climate zonal mean clouds

SURFACE ROUGHNESS:
Vegetation dependent over land, stress dependent over ocean

ALBEDO:
Monthly mean modified by snow cover

SOIL MOISTURE:
Monthly climate interactive with forecast precipitation

SNOW DEPTH/COVER:
Monthly climate interactive with forecast

SEA ICE:
Monthly climate fields

TABLE II. NMC DERF Experiment Strategy

PHASE I:
Nine 30-day "exploratory" runs

PHASE II:
Series of 108 contiguous 30-day runs, 24 hours apart, 14 Dec 1986 - 31 March 1987

PHASE III:
Series of 20 contiguous 30-day runs, 12 hours apart, 1 -10 Jan 1987
Series of 10 contiguous low resolution(R20) 30-day runs, 24 hours apart, 1-10 Jan 1987
Monte Carlo experiments (in progress)

PHASE IV:
Design and test of possible operational configuration for DERF (scheduled to begin sometime in 1989)

possible. The most important changes in the MRF over the past few years are outlined by White (1988), while the version of the MRF ("MRF86") used in the Phase II experiments discussed below is described in Table I.

The sets of experimental range predictions follow the strategy outlined in Table II. Phase I consisted of nine exploratory 30-day integrations whose purpose was to establish a benchmark of the model's capabilities in comparison to observations and other centers' (such as ECMWF and UKMO) predictions and to provide experience in the logistics, data manipulation, and evaluation of extended runs. Suffice it to say that the results of this effort were sufficiently encouraging to warrant proceeding to the next phase. Phase II, which will be the principal focus of this paper, was an intensive effort to address a number of outstanding questions on prediction, predictability, and the potential utility of DERF, especially issues related to the concept and application of lagged average forecasting (LAF). The data base, which has become an invaluable resource and will be made available to the research community (Schubert et al, 1988), consists of 108 contiguous 30-day integrations from initial conditions 24 hours apart between 14 December 1986 and 31 March 1987. The Phase III predictions are designed to complement Phase II in addressing questions on the optimal size and spacing of LAF ensembles, the relevance of a low-resolution runs, and the Monte Carlo (MC) versus LAF (or combined) approach for constructing forecast ensembles. Finally, based upon the experience acquired in Phases I to III and in context with available resources, Phase IV will consist of designing and testing possible operational configurations of DERF.

The balance of this paper is devoted to evaluation of the results. The subjects addressed are i) model performance, ii) postprocessing of model output, and iii) estimation of skill. Model performance is described from the perspective of systematic errors (Section 2.1), and in terms of verification scores as a function of initial conditions, length of forecast, and geographical region (Section 2.2). Postprocessing of model output refers to procedures for extracting and optimizing the useful information in the predictions. Here we discuss the effects of time averaging and selection of the optimum averaging period (Section 3.1), the influence of LAF on forecast skill (Section 3.2), correction of systematic errors (Section 3.3), empirical orthogonal function (EOF) filtering (Section 3.4) and objective/subjective specification of operationally relevant parameters, such as surface temperature anomaly (Section 3.5). This latter subsection

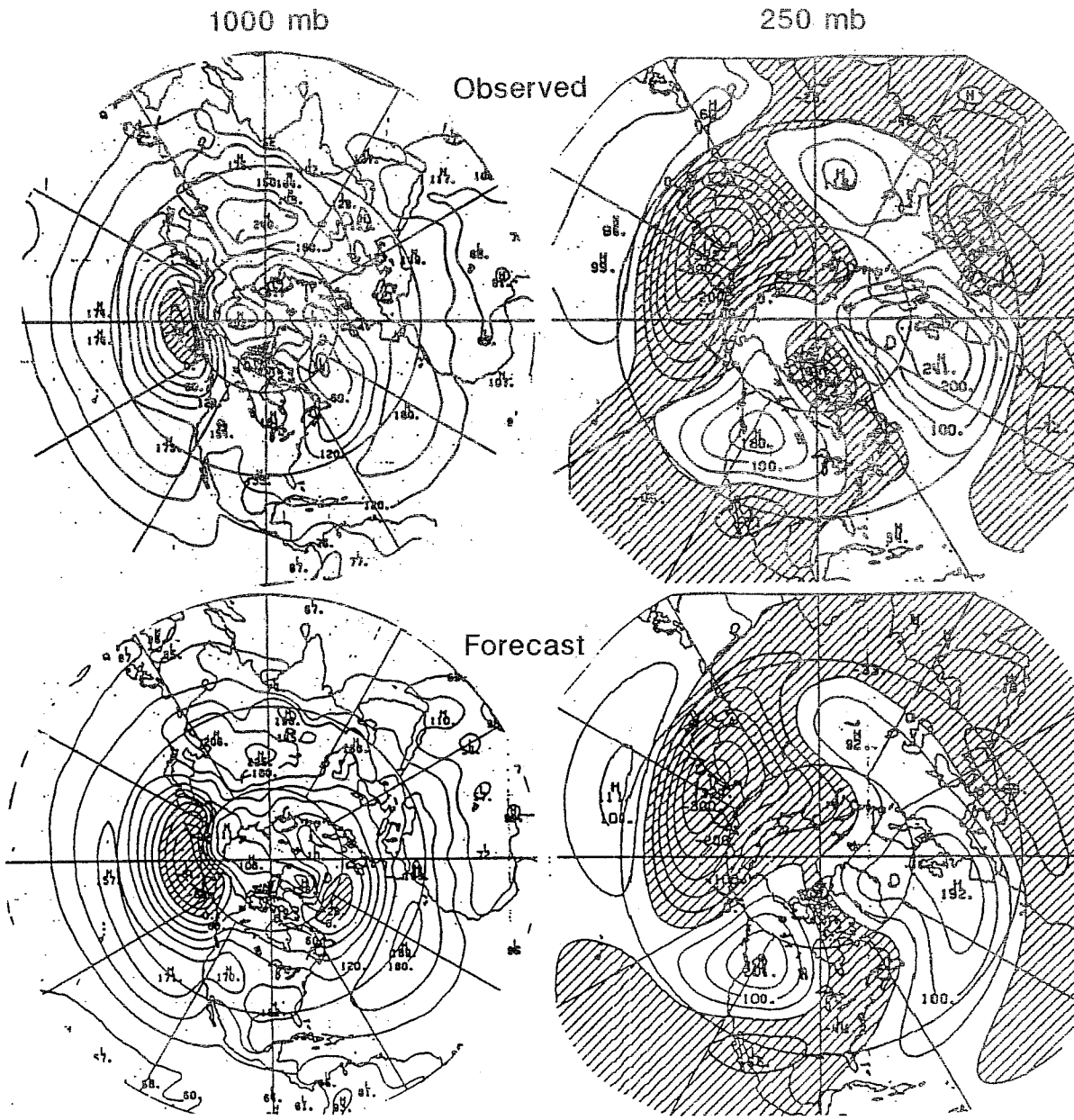


Fig. 2.1.1 Ensemble mean of 90 forecasts averaged over forecast days 1-30 (bottom) and verifying analyses (top) of 1000 mb height (left) and 250 mb height (right) with zonal mean removed. Contour interval 30 m (left), 50 m (right).

includes direct comparison of DERF predictions and the Climate Analysis Center (CAC) operational Monthly Outlooks. Estimation of skill addresses one of the most crucial problems in medium and extended range forecasting, the variability of skill and its possible prediction. We discuss the relationships between skill and the dispersion of predictions within LAF ensembles (Section 4.1), skill versus forecast persistence and magnitude of the predicted anomaly (Section 4.2), and the dependence of forecast quality upon circulation regime especially the occurrence of blocking and the PNA index (Section 4.3). A summary and discussion concludes the paper (Sec. 5).

2. MODEL PERFORMANCE

2.1 Systematic Behavior

This section examines the systematic behavior of DERF II forecasts and reviews the effect of model changes on systematic (i.e., time mean) errors. The forecasts appear to predict the zonal mean wind and stationary waves with reasonable accuracy, but display a stronger midlatitude zonal flow than observed. Fig. 2.1.1 displays the forecast and verifying time-mean atmospheric circulation for a 90-case ensemble with initial conditions from 1 January to 31 March averaged over forecast days 1 to 30. At 1000 mb the forecasts maintain the positions of the major quasi-stationary features but display too strong oceanic lows and highs over North America and the subtropical Atlantic and too weak a polar high. At 250 mb (with the zonal mean removed) the forecasts correctly position most major ridges and troughs, but have less success with their amplitudes. The largest error develops in the east Atlantic where a strong blocking ridge is observed and a weaker ridge much more like climatology is forecast. The forecasts also develop too strong a ridge over western North America. The ability of the model to predict stationary waves declines markedly from forecast days 1-10 to forecast days 6-15 (not shown). The forecasts also tend to weaken stationary waves at 250 mb; however, the weakening occurs largely in the first 10 days.

The systematic errors at 1000 and 250 mb in the 1 to 30 day means for the 90 case ensemble (Fig. 2.1.2a,b) tend to be zonally symmetric and imply a substantial westerly bias in the midlatitude zonal wind, a common error in high-resolution general circulation models without gravity wave drag. The ensemble error in 1000-500 mb thickness (Fig. 2.1.2c) is less zonally

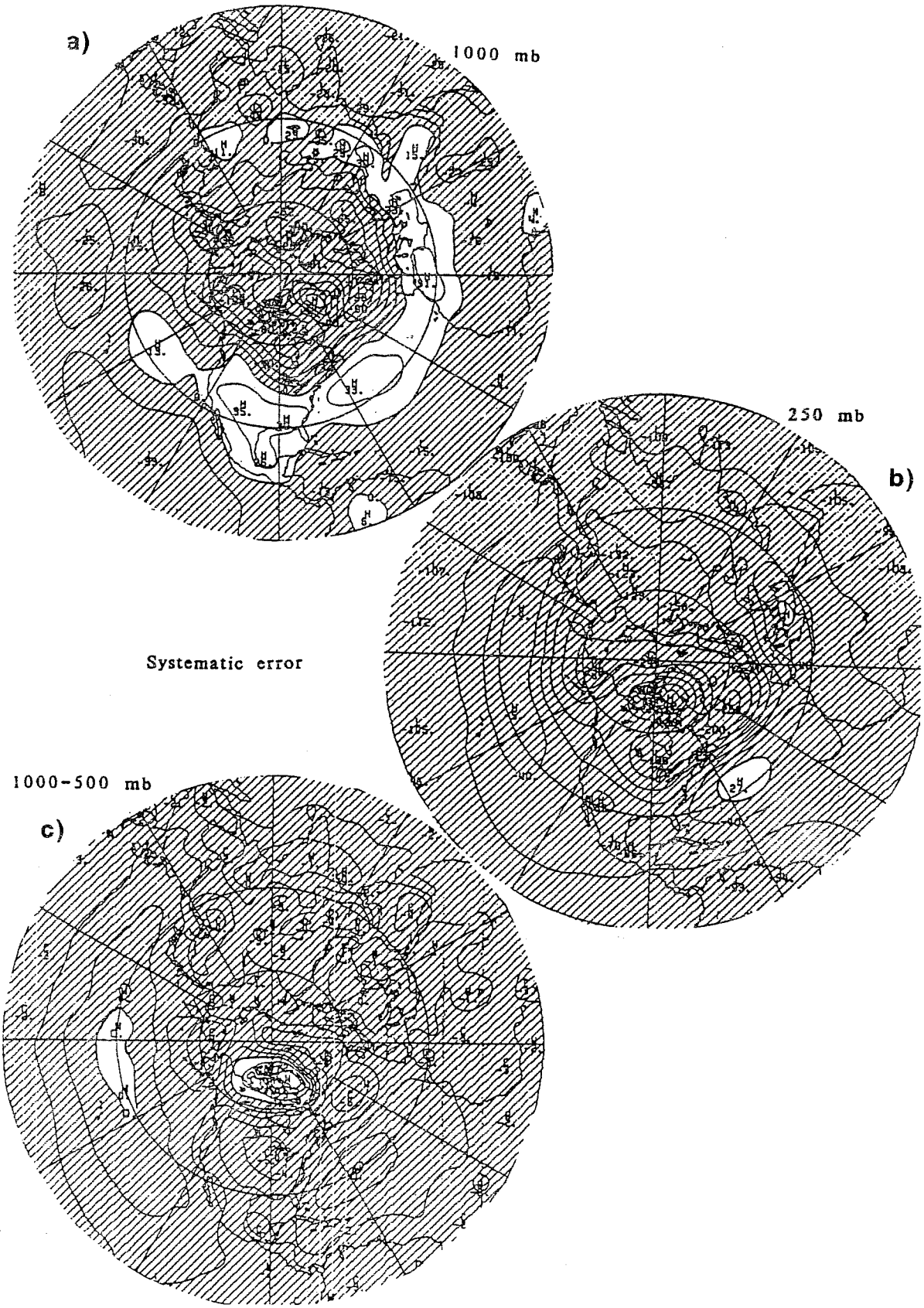


Fig. 2.1.2 Ensemble mean error of forecast days 1-30 of 1000 mb (a) and 250 mb (b) height, and 1000-500 mb thickness (c). Contour interval 20 m, 40 m and 1°C, respectively; negative values shaded.

symmetric and displays a cold bias nearly everywhere that is particularly strong over North America and near the oceanic lows. Experimental 30-day forecasts with gravity wave drag indicate that it reduces the midlatitude westerly bias and improves the forecast stationary waves (A parameterization of gravity wave drag was implemented operationally in the fall of 1987.)

Fig. 2.1.3 shows the relationship between the systematic and total error of zonal mean height at 1000 and 500 mb in a 31 case ensemble from initial conditions for 1 to 31 January (results from ensembles from initial conditions for February and March are quite similar). The negative systematic errors in high latitudes show continuous growth throughout the forecasts (2.1.3a,b) and so does the root-mean-square error at 500 mb (2.1.3d). Removal of the systematic error from the root-mean-square error at 500 mb (2.1.3f) produces considerable reduction in total error in the tropics and in high latitudes in the Northern Hemisphere. The ratio of systematic error squared to the total error variance is shown for 500 mb (2.1.3e) and 1000 mb (2.1.3c). At 500 mb systematic error dominates the tropical errors by day 2, but is less important in midlatitudes and at 1000 mb.

Fig. 2.1.4 illustrates the growth rate in temperature error with the 30-day forecast from 9 February. The zonal mean error in temperature in the 30-day averaged forecast (2.1.4a) displays strong cold biases in the stratosphere and weaker cold biases in the lower troposphere. The drift of temperature averaged over 30-60^{o}N from initial conditions (2.1.4b) shows that the lower tropospheric cold bias develops largely within the first 5 days, while the stratospheric cold bias continues to grow throughout the forecast. Experiments with gravity wave drag and with 18 levels of humidity show some reduction in the stratospheric cold bias, while preliminary experiments with interactive clouds and improved surface fluxes of moisture over land yield a reduced cold bias in the lower troposphere (new parameterizations of surface evaporation and the use of interactive clouds are currently being tested for operational implementation).

2.2 Verification Scores

Figure 2.2.1 displays the 500 mb height anomaly correlation (AC) scores over the Northern Hemisphere (20^{o} to 80^{o}N) for the sequence of 108 DERF

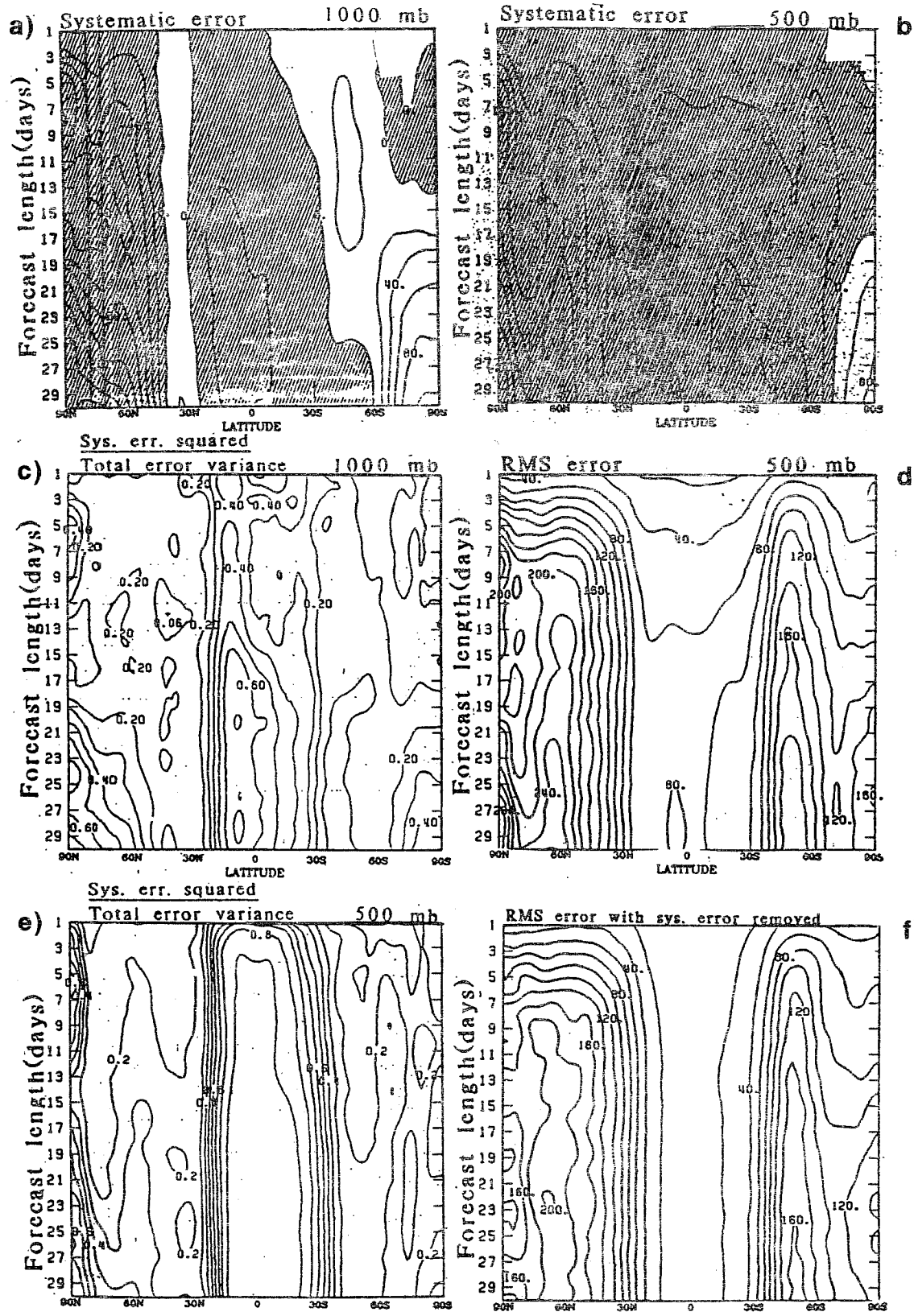


Fig. 2.1.3 Zonal mean error in 1000 mb (a) and 500 mb (b) height as function of forecast length averaged over 31 January cases. Contour interval 20 m, 40, respectively. Zonal mean RMSE of 500 mb height (d) and same with systematic error removed (f). Contour interval 20 m. Zonal mean ratio of systematic error squared to total error variance of 1000 mb (c) and 500 mb (e) height. Contour interval 0.1.

Phase II predictions. Also shown are the corresponding scores of persistence forecasts, where persistence is the anomaly of the respective averaging periods (10 or 30 days) preceding the initial time. For the 1-30 day means (Fig. 2.2.1a) the average score over the sample is 0.39. In virtually all cases the AC is greater than zero, and the dynamic predictions score higher than the corresponding persistence forecasts. The latter is an especially relevant control, since operationally persistence is difficult to beat in forecasting monthly means. Note also that in many cases the AC exceeds the often cited 0.50 - 0.60 criteria (Hollingsworth et al, 1980) for categorizing forecasts as synoptically useful. From the AC of overlapping 10-day means (Figs. 2.2.1 b-f) it is apparent that much of the skill in the 30-day means reflects the influence of the first half of the extended runs and is strongly associated with persistence in the latter half. Nevertheless, even over the last 10 days the AC is generally positive, better than persistence in a majority of cases, and reaches the 0.50 to 0.60 level in some instances. Some of the exceptionally good cases at extended ranges reflect the so called "return of skill", a phenomenon which has been noted elsewhere (Molteni, et al, 1986) and which appears related to circulation regime (see Section 4.3). Finally, the AC varies regionally (not shown) with forecasts generally more skillful over North America than Europe (1-30 day means, 0.40 and 0.34, respectively), and the variability from one case to the next is larger on a regional basis than for the Northern Hemisphere as a whole.

Here, as is true throughout our evaluation (except where otherwise indicated), the essential elements of the results are independent of whether the AC or root-mean-square error (RMSE) are used as measures of model performance. Thus, for example, the sequence of 1-30 day mean 500 mb height RMSE (Fig.2.2.2) shows that DERF is better than persistence in almost all cases. Additionally, in many cases the RMSE of DERF predictions is less than the level of the climatological standard deviation, a criterion found comparable to the 0.50 to 0.60 value of AC for useful skill. Keep in mind, however, that standard verification scores and procedures do not necessarily convey a complete picture of the operational utility of numerical predictions. This reflects both inadequacies of the scores themselves (e.g., AC & RMSE) and the the fact that verified fields (e.g., 500 mb height) may not relate directly to forecast parameters of interest (e.g., precipitation).

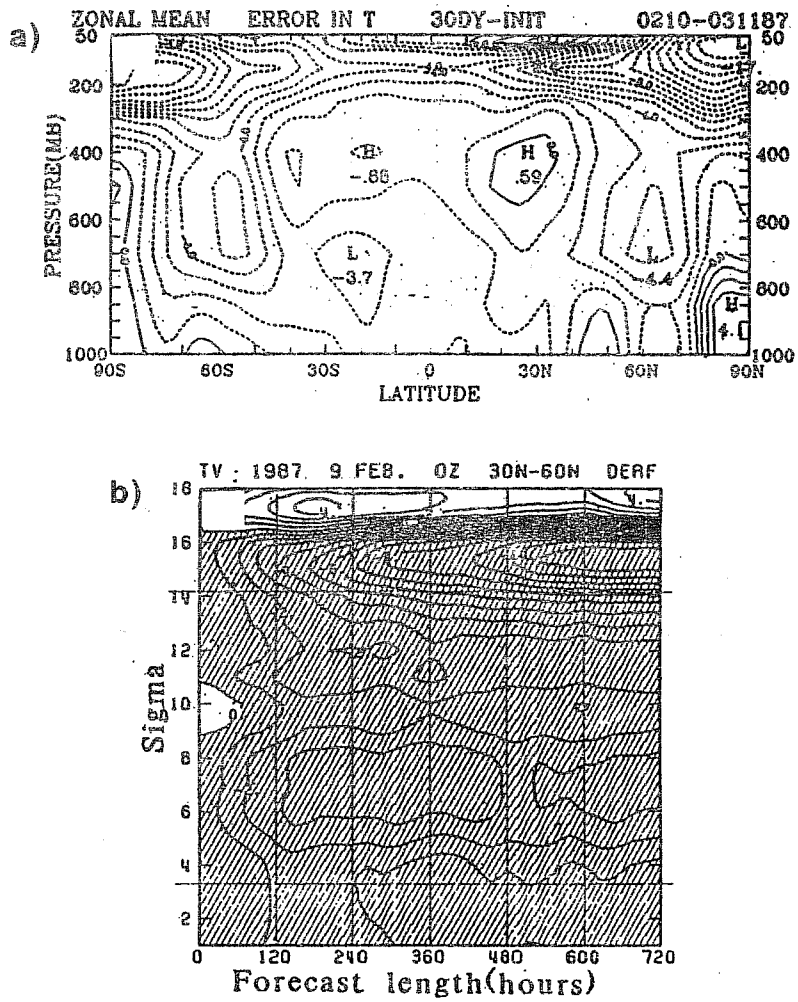


Fig. 2.1.4 Zonal mean temperature error in 1-30 day mean forecast from 9 Feb 1987 (a) and drift in forecast temperature averaged over 30-60°N from initialized values (b). Contour interval 1°C.

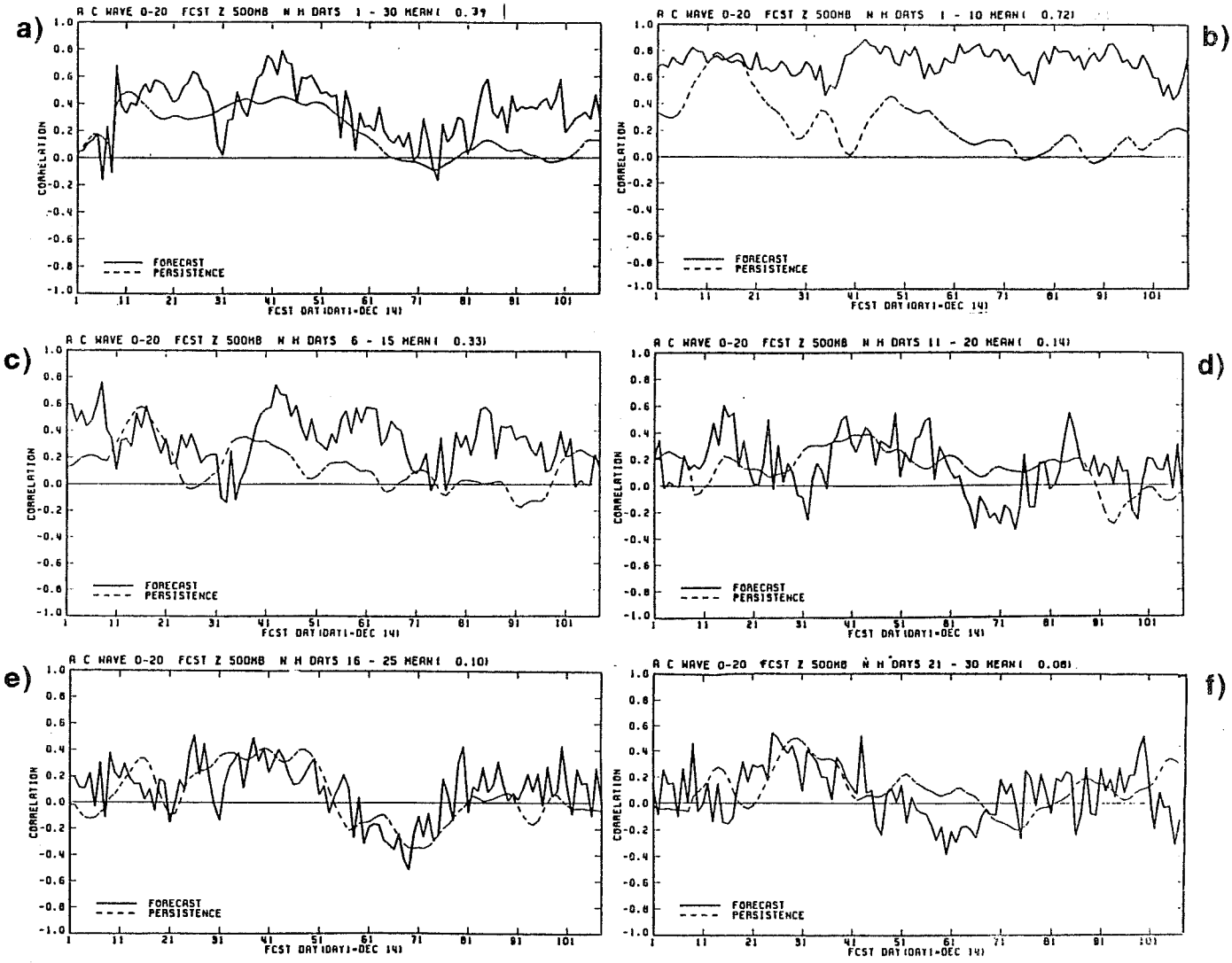


Fig. 2.2.1 AC of NH 500 mb height for the contiguous series of 108 cases (case 1, 14 Dec 1986 initial conditions): (a) 1-20, (b) 1-10, (c) 6-15, (d) 11-20, (e) 16-25, (f) 21-30 day means.

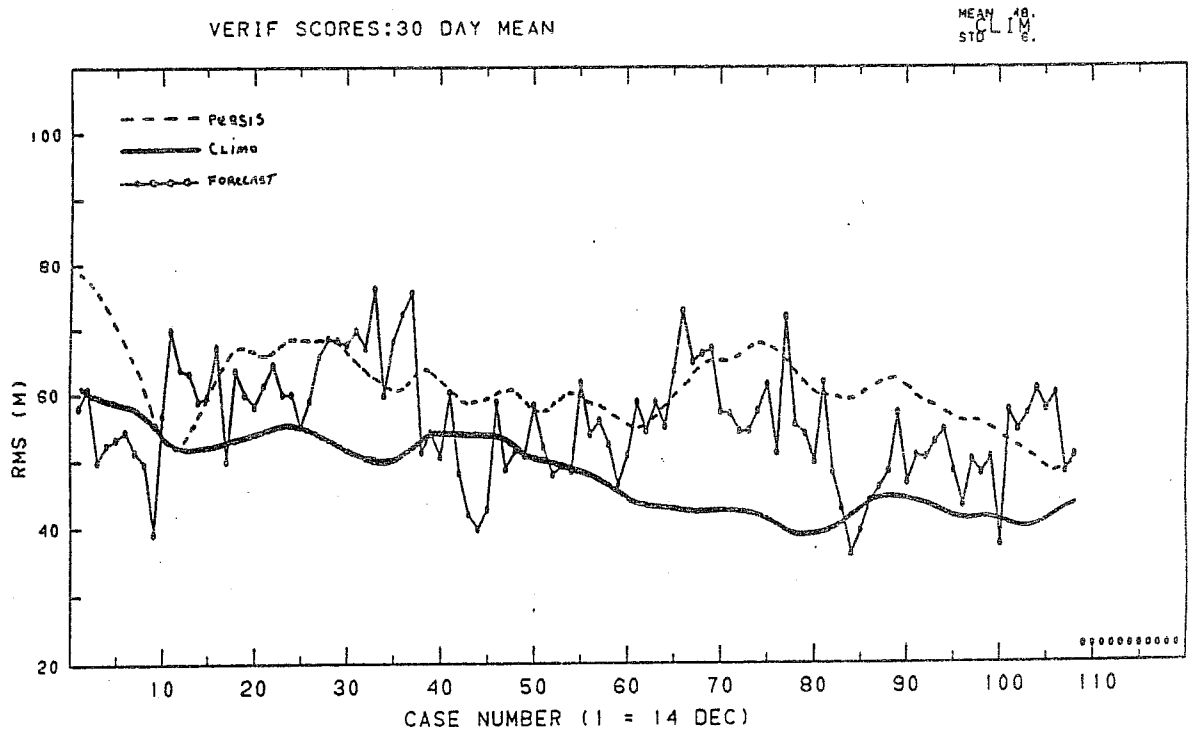


Fig. 2.2.2 As Fig. 2.2.1(a), except RMSE climatological standard deviation added.

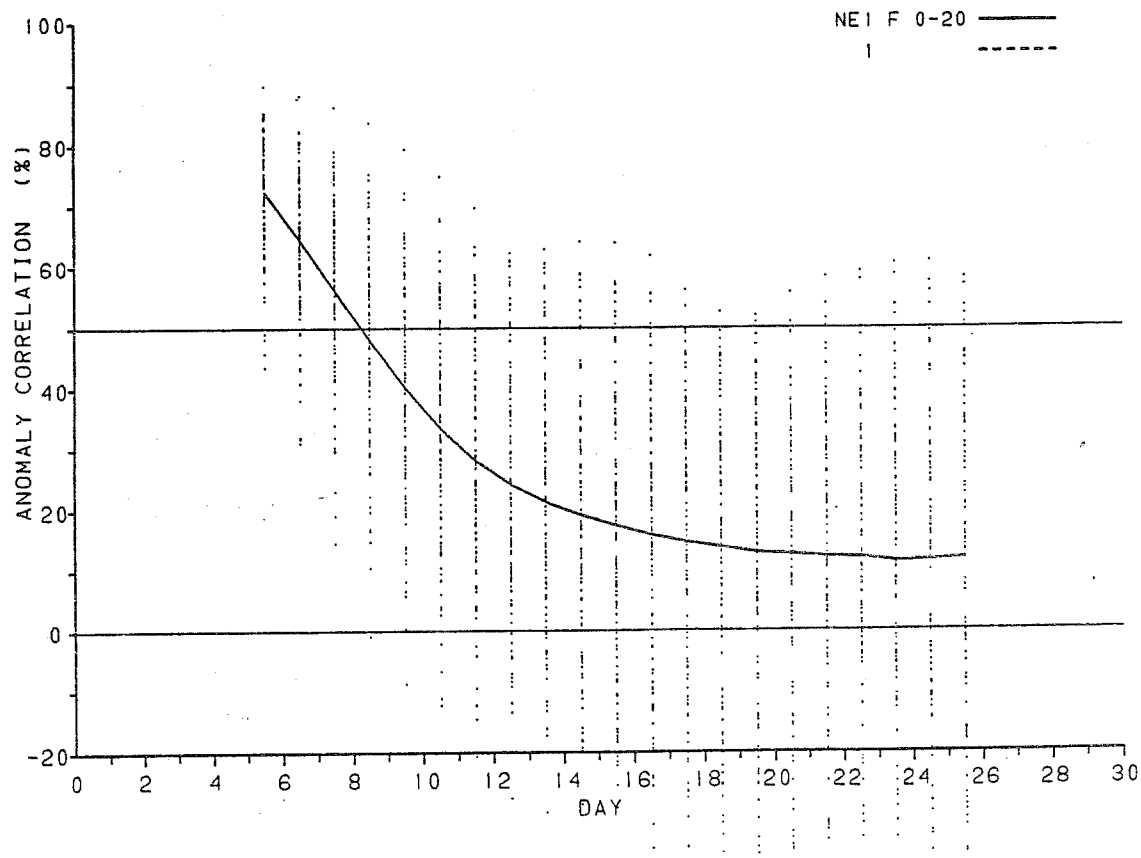


Fig. 2.2.3 AC (108 case average) of NH 500 mb height DERF (solid) and persistence (dashed) forecasts vs. length of prediction for overlapping 10 day means. The distribution of the 108 scores at each time range shown by dots.

The variability in skill is also apparent in Fig. 2.2.3, which displays the distribution of scores about the 108 case average AC of 10-day means versus length of forecast (1-10 plotted at 5.5 days, 2-11 at 6.5 days, etc.). On average the scores of the DERF predictions reach the .5 level at day 8 and are more skillful than persistence through around day 19. The envelope of scores is rather wide even at medium ranges and widens further at longer ranges. The average results thus convey little information relative to the success or failure of individual cases, an observation certainly not unique to this study (e.g., Mansfield, 1986). It is clear that an essential requirement for the success of DERF is diagnosing, understanding, and ultimately predicting the variability in forecast skill. This subject is discussed in Section 4. Finally, as shown below, corrections for systematic errors and various weighting and filtering procedures have relatively little influence on average verification scores and do not change the basic character of the case to case variation in skill scores.

3. POSTPROCESSING OF MODEL OUTPUT

3.1 Time Averaging

According to classic predictability studies (e.g., Lorenz, 1982), the theoretical limit of deterministic predictability is about two weeks for individual weather systems. That limit naturally will vary, for it depends upon the dominant modes of atmospheric instability which are related to factors such as season, location, and scale. Time averaging is an attempt to extend the range of predictability by suppressing the less predictable, high frequency components of the circulation. It assumes, of course, that the evolution of the low-frequency circulation is not strongly dependent upon details of smaller-scale features.

The influence of time averaging on forecast skill is illustrated in Fig. 3.1.1 by comparison of the 108 case average (Northern Hemisphere) AC for daily and 10-day mean 500 mb height fields versus length of forecast. Also shown is the score referenced above for the 30-day means. The choice of 10-day means follows the rationale outlined by Miyakoda (1985) which suggests this filter is adequate to eliminate the "noise" of the high frequency baroclinic waves while retaining the "signal" of the stationary and low-frequency planetary and some synoptic-scale waves. Consideration of 30-day means largely reflects operational precedent (i.e.,tradition) and user

requirements.

Fig. 3.1.1 indicates that the 10-day means gain about 1.5 days in predictability (6.5 to 8) with respect to the 0.50 crossover and approximately 5 days (14 to 19) relative to scoring better than persistence. Overall, the AC of the 10 and 30-day means is larger than the average of the daily scores over the respective averaging periods. Note also that the effect of the temporal filtering varies from case to case. In particular, as illustrated by Fig. 3.1.2, the gain from time averaging is generally much larger when the individual daily forecasts retain higher skill. This agrees with van den Dool's (1985) argument that the advantage of time averaging occurs only when the daily forecasts have skill. Finally, we note that the application of spatial filtering to temporal averages results in negligible additional improvement (not shown).

It is clear from the results thus far that most of the skill in the 1 to 30 day average forecasts is concentrated in the earlier time ranges. This suggests that it may be preferable to truncate the time averaging at less than 30-days in order to retain the information of the 30-day average and eliminate the influence of the non-skillful later part of the integration. To address this question Fig. 3.1.3 presents the AC of forecasts averaged from days 1 to N (N = 1 to 30) and verified as 30-day means. The solid curve represents the 108 case mean, while the dots indicate for each case the maximum AC and N at which it occurs. The curve reflects two competing influences. As the the averaging period is extended, increasingly more shorter scale, high frequency components are removed, and the skill score improves. However, as N increases the mean includes additional less accurate daily forecasts of all scales, so the curve tends to level and then decline.

Overall, averaging the first 7 or 8 days of the integrations provides the best estimate of the complete 30-day mean. The range of optimum N's, though, extends from using the day 1 prediction as the 30-day mean to averaging the forecast over the entire month. In each case the best proxy (optimum N) beats persistence and in about half the cases results in an AC greater than 0.50. The large variability in the optimum value of N, of course, is related the intrinsic variability in skill discussed above. If one knew a priori the optimum N for each case, the average AC in predicting 30-day means would increase to 0.54 from the 0.39 value of the full 30-day

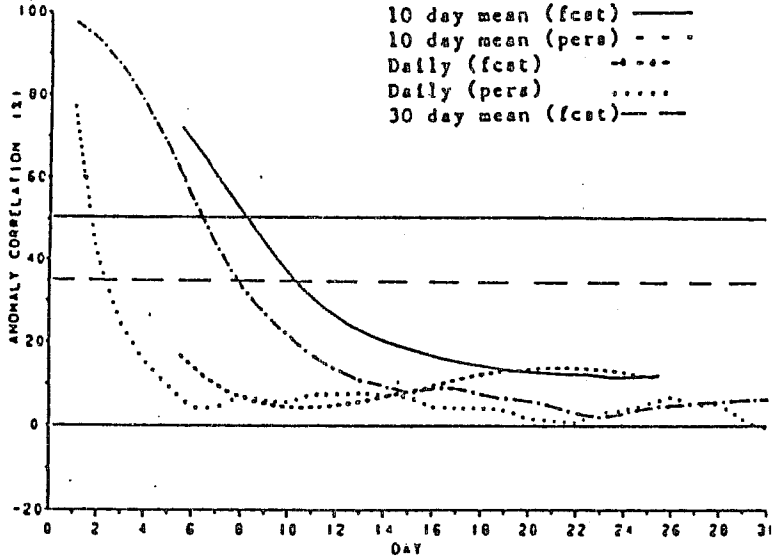
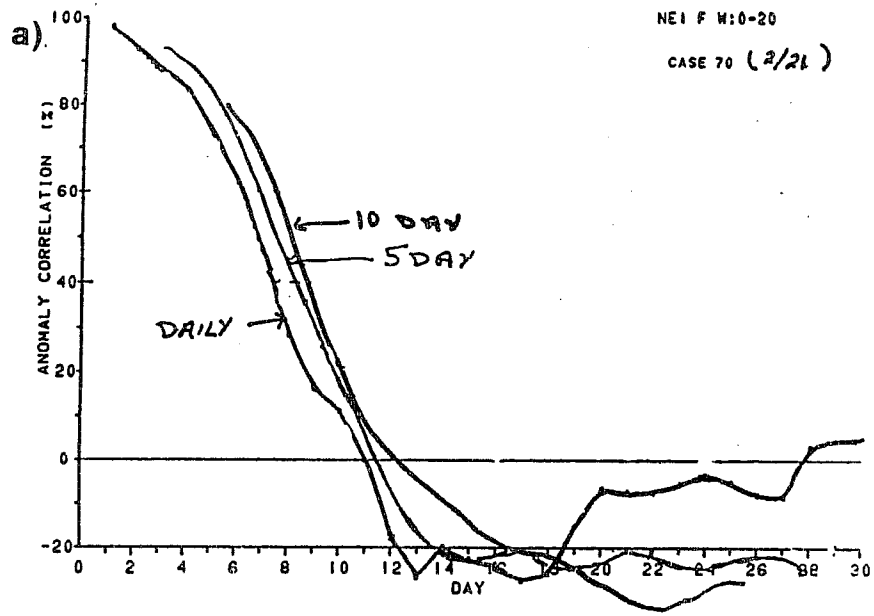


Fig. 3.1.1 AC (108 case averaged) of NH 500 mb height DERF and persistence forecasts vs. length of prediction for daily fields and 10-day means. Also shown is the 108 case average AC for 30 day means.



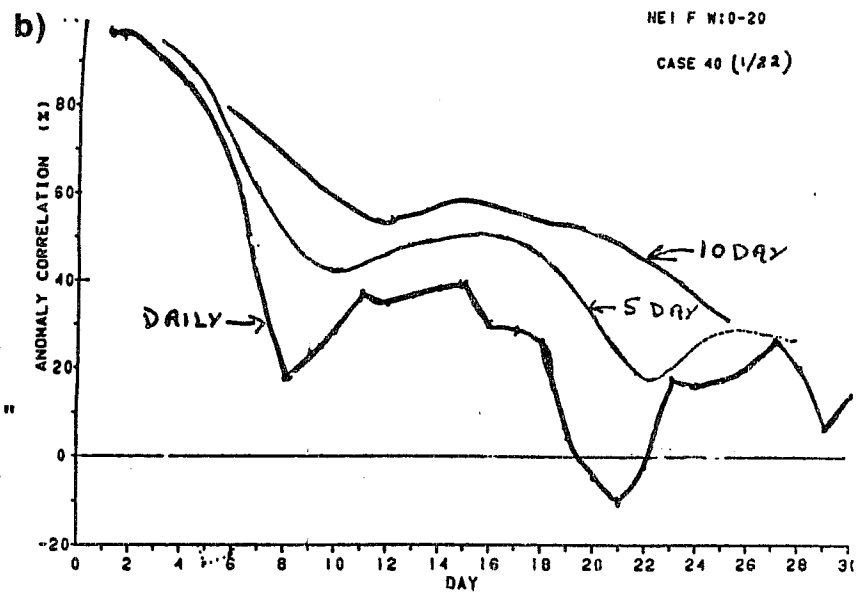


Fig. 3.1.2 AC of NH 500 mb height for daily (solid) and 10-day means (dashed) for a "poor" (a) and "good" (b) case. Initial conditions are for 21 Feb. and 22 Jan. respectively.

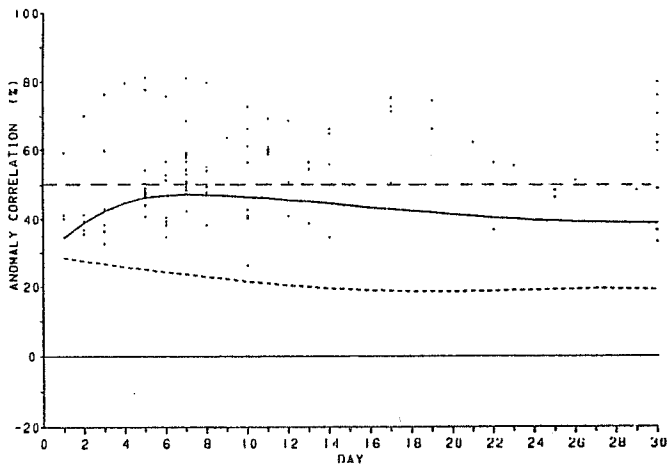


Fig. 3.1.3 AC of NH 500 mb height 1-N day means (N=1 to 30) verified as 30-day mean forecasts vs. N. Dots indicate the maximum AC and N for each case.

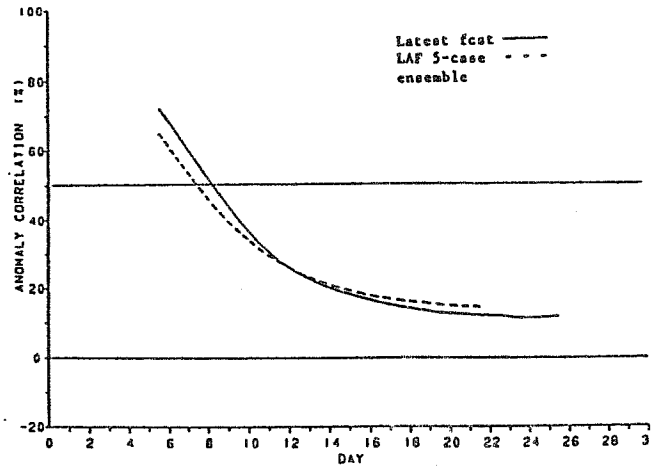


Fig. 3.2.1 AC (108 case average) of NH 500 mb height for LAF and latest available forecast vs. length of prediction (10-day means).

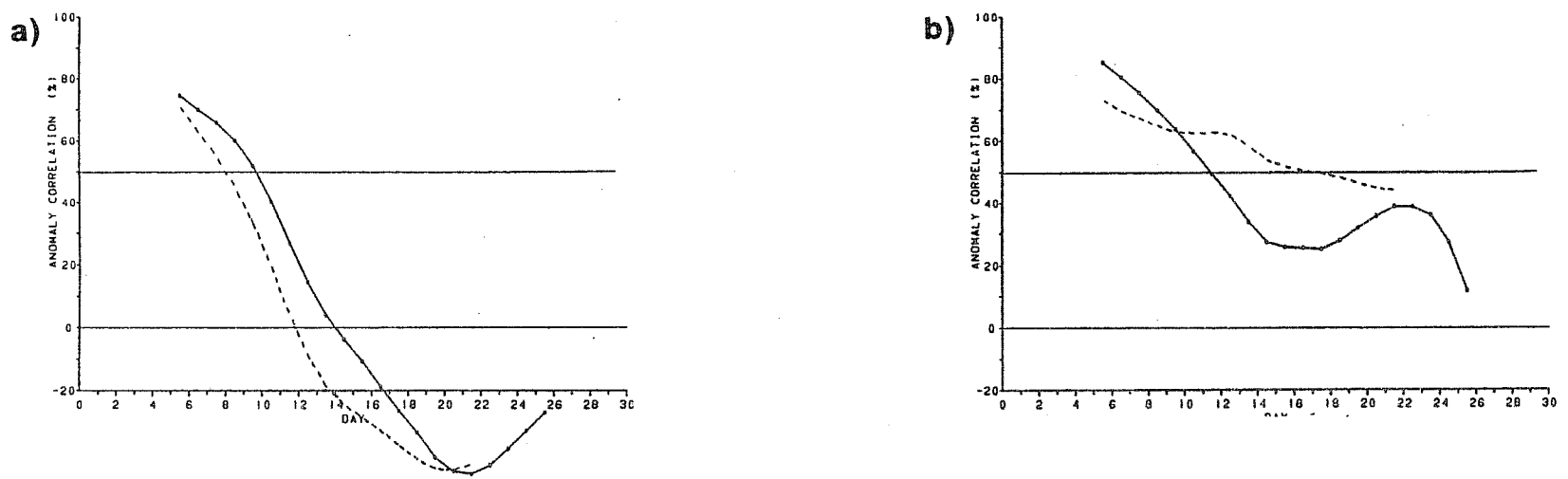


Fig. 3.2.2 AC of NH 500 mb height for latest (solid) and LAF (dashed) for a "poor" (a) and "good" (b) case (10-day means). Initial conditions are for 19 Feb. and 24 Jan., respectively.

mean forecasts. It should be noted, however, that using the average optimum value of N for all cases (N=7) already increases the AC to Ø.47. From this one can conclude that for the purpose of monthly forecasting it is not necessary to integrate beyond a week. However, more sophisticated procedures in constructing time averages, e.g., accounting for the dependence of predictability on scale (Dalcher et al,1988) and a priori indications of the expected accuracy of individual forecasts (See Sec.4), have the potential of extracting additional information from the integrations beyond a week.

Roads (personal communication) extended the above considerations to determining the period over which the DERF forecasts should be averaged to optimize representation of any given verification period. As an example, the best proxy overall for a 6-15 day mean turned out to be the average of forecast days 4 to 11, but the optimum averaging period varies considerably from case to case.

3.2 Lagged Average Forecasting (LAF)

As integrations proceed from the medium to extended ranges the intransitive nature of atmospheric circulations, coupled with uncertainties in initial conditions, renders the forecast process increasingly more probabilistic in nature (Lorenz, 1977). This suggests that an ensemble mean of forecasts starting from somewhat different initial states should be on average closer to the truth than a single deterministic prediction. As an alternative to the Monte Carlo (MC) approach of Leith (1974) for constructing ensembles from sets of randomly perturbed initial conditions, Hoffman and Kalnay (1983) proposed lagged average forecasting (LAF) where ensemble means are generated by combining the latest available prediction with forecasts for the same verification time from a sequence of earlier analyses. Dalcher et al (1988) obtained encouraging results with LAF in medium range forecasts with weights based upon horizontal wavenumbers. Our assessment of the advantages of LAF versus MC (or combined LAF/MC) is underway (Phase III). Whether or not either approach enhances skill, the estimate provided by the spread within ensembles of the probabilistic distribution of possible atmospheric evolutions can be a viable tool for predicting forecast skill (Section 4.1).

As a preliminary assessment of LAF, ensembles were constructed by averaging

with equal weights up to 5 forecasts starting from analyses 24 hours apart. The results for an ensemble size of 5, which encompasses initial conditions over a period of 4 days, are illustrated in Fig. 3.2.1. Through about day 10 (i.e., 10-day mean centered on day 10) the LAF predictions averaged over 108 cases are less skillful than the latest forecast. This is not surprising given the ensemble members are weighted equally without accounting for the larger errors of the earlier forecasts. Use of weighting schemes based on EOF's that account for the growth of error with time (see Section 3.4), however, produce only a marginal improvement in the DERF predictions through about 10 days. At more extended ranges, where individual ensemble members become equally likely realizations (weights tend towards equality), LAF has only a small positive effect. At all time ranges the effects of LAF for ensembles smaller than five are the same as in Fig. 3.2.1, but the magnitude is proportionately less. As with time averaging the effect of LAF is case dependent. Good forecasts are improved considerably, but poor forecasts are not improved or made worse (e.g.,Fig. 3.2.2).

For the period 1-10 January, 1987 additional 30-day forecasts were performed from 1200 GMT analyses to study the advantages of LAF with 12 versus 24-hour spacing of ensemble members, including the larger size ensembles possible with the 12-hour interval. As shown in Fig. 3.2.3 there is a slight advantage of 5-member LAF ensembles with 12-hour separation (initial conditions including a 48-hour period) over 24 hour spacing through day 8, presumably because the 12-hour LAF contains "younger" forecasts. This improvement would probably disappear with optimal weights instead of the the equal weights used here (Dalcher et al,1988). Beyond 8 days the 12-hour LAF does not appear to have any advantage. Combined 10-case ensembles, with predictions 12-hours apart encompassing 4 days of initial conditions, show some additional improvement beyond day 15. Note this is coincident with the "return of skill" that characterizes the cases of this limited sample, and the result may not be representative.

3.3 Correction of Systematic Errors

Miyakoda (1986) found that removing the systematic errors a posteriori substantially improved forecast skill scores. To assess the impact of this "empirical adjustment" on the DERF Phase II data the arithmetic mean error over the experimental period was subtracted from the predictions, and the

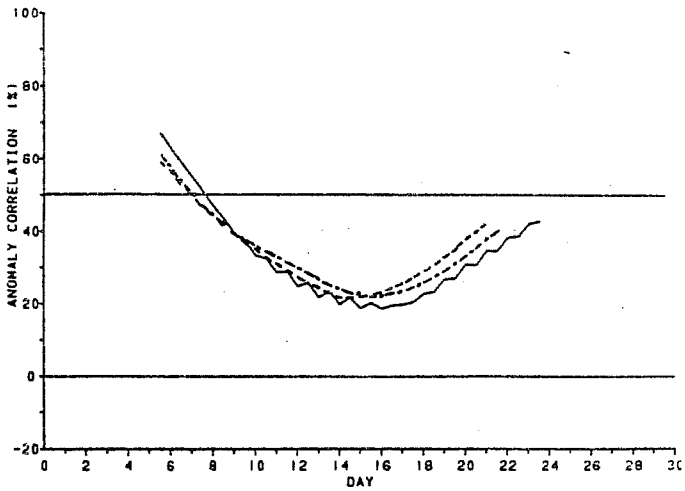


Fig. 3.2.3 Comparison of NH 500 mb height AC of 5 member LAF forecasts with 12 (solid) and and 24 (dash-dot) hour spacing, and 10 member LAF predictions with 12 hour spacing (dash) averaged over cases with initial conditions from 1-10 Jan. (10-day means).

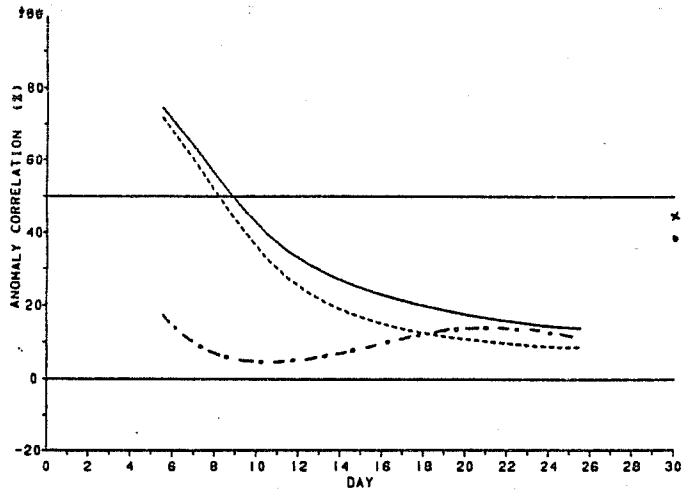


Fig. 3.3.1 AC (108 case average) of NH 500 mb height for with (solid) and without (dashed) "empirical adjustment" (10-day means), and persistence score dash-dot). Adjusted (X) and unadjusted (O) scores for 30-day means plotted at day 30.

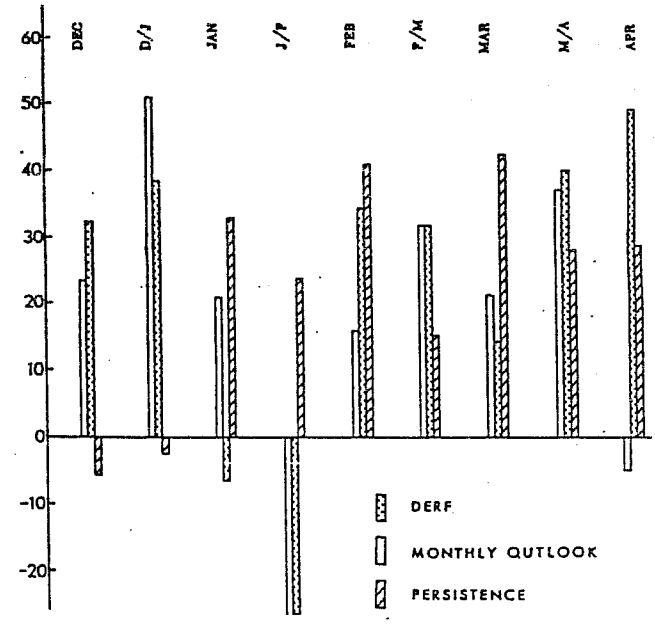


Fig. 3.5.1
Skill scores for monthly mean 3-class temperature forecasts verified against 100 U.S. stations (1.0 = perfection, 0 = no better than chance). Forecasts are outlooks for given month or mid-month to mid-month period.

forecasts reverified. Comparisons of verification scores with and without the systematic error correction are shown in Fig. 3.3.1.

Over the 108 case sample empirical adjustment improves the 500 mb height AC of 10-day means approximately 1/2 day relative to the 0.50 to 0.60 criteria for useful skill, and the adjusted predictions gain about 3 days of predictability relative to persistence. In the 30-day means the overall improvement in the AC is 6 points (0.39 to 0.45). The results here indicate gains in skill, but much smaller than those of Miyakoda. In part this may be due to a smaller contribution of climate drift to the total error. Also, as shown by Epstein (personal communication), the "systematic error," defined as the average error over some period (e.g., the DERF Phase II experiment) is to a large extent non systematic. A significant component of the average error depends upon transient aspects of low-frequency circulations and, hence, upon such factors as season, location, length of averaging period, and weather regime. Estimates of systematic error corrections based upon independent and possibly more representative "training" periods are discussed below.

3.4 Empirical Orthogonal Function (EOF) Filtering

The EOF analysis was based upon the 700 mb heights from the period 1949 - 1980 with the 30-year mean and seasonal cycle removed and anomalies in the time series normalized. Both forecasts and verifying analyses were expanded into the first 16 EOFs, which accounted for 81% of the total variance.

The effects of EOF filtering were evaluated for the time means of 1 to 7 days considered as proxies for the 1 to 30 day average forecasts. As shown by Mo (1988) the spatial filtering of EOF's can be used to remove higher frequency components and some of the systematic error. Also, because the first few EOF's represent the most persistent anomalies and dominant low-frequency patterns, this filter tends to work better for monthly means than for daily data. On average, the EOF filter increases the AC from 0.45 to 0.51, presumably because the low-frequency part of the spectrum is more predictable.

We also attempted to enhance the skill in predicting the 30-day means by time weighting the individual forecasts of the day 1 to 7 period. A simple candidate weight is the daily AC score of day 1-7 forecasts. Because of the

long record of forecasts available, we used 1982/1983 to 1986/1987 verifications of MRF predictions. An alternative approach is to determine weighting via linear regression of each EOF component of day 1 to 7 forecasts with the corresponding components of the 7 day mean of verifying analyses. The training period consisted of the four winters immediately preceding (and not including) the 1986/1987 cold season of the Phase II experiment. On average both schemes had adverse effects relative to the unweighted EOF scores. The simple weighting produced an 8 point decline in AC, on average, while the regression method degraded the result by 6 points (computation of weights by regression on 30 day, rather than 7 day means decreases the AC by another 6 points). The negative effect likely reflects too little influence given to the less accurate but nevertheless skillful last few days of each 7 day set of forecasts.

We have also assessed systematic error correction and combined error correction and LAF in the context of EOF filtered fields and 1-7 day proxies for the 30-day means. Unlike the "empirical adjustment" method of Section 3, an operationally relevant procedure for correcting systematic errors must estimate them a priori from independent data. The approach used here is similar to that of Saha and Alpert (1988) except the error correction was in EOF's rather than grid space. The model bias is determined from the mean error of the most recent 30 verifiable forecasts (30-day training period). For this study the biases of 1-7 day mean forecasts were based upon comparisons with the corresponding 7-day, rather than 30-day average of verifying analyses. Otherwise, the most recent verifiable forecast would be 30-days rather than 7-days old, and the corrections less likely to reflect the flow dependent component of the mean error. The "corrected" forecasts verify somewhat less well than uncorrected predictions in terms of AC (7 point difference), although the RMSE is reduced by a small amount (6 m). This is probably due to the fact that EOF filtering alone is very effective in reducing systematic errors (Mo, 1988), whereas Livezey and Schemm (1988) showed that for unfiltered fields correcting for systematic errors does improve AC scores. Finally, using longer training periods of 40 and 50 days made little difference on the results. This is consistent with Saha and Alpert, who showed the effect of bias corrections was insensitive to training periods as long as they are greater than about 20 days and include the most recent forecasts possible.

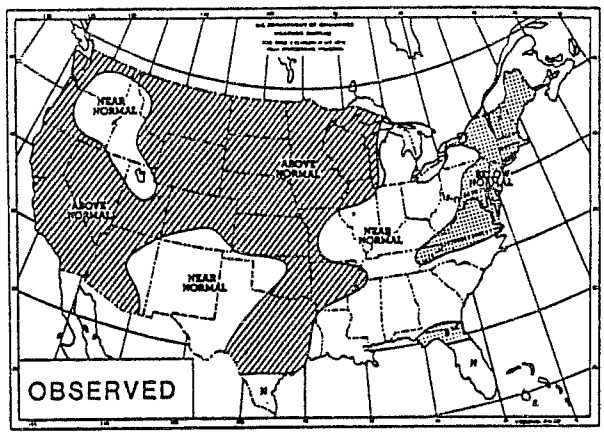
With respect to LAF, ensembles consisted of 4 sets of predictions

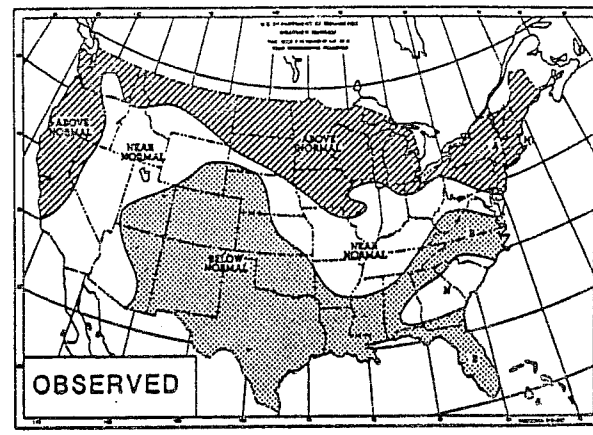
spaced 24 hours apart (1 to 7 to 3 to 10- day forecast averages) with weights following the approach of Dalcher et al (1988). Overall, LAF degrades the forecasts with respect to the corresponding (corrected) deterministic predictions (3 point difference). Similar results were obtained using only 3 elements in LAF, but scores were degraded even more when 2 ensemble members were used.

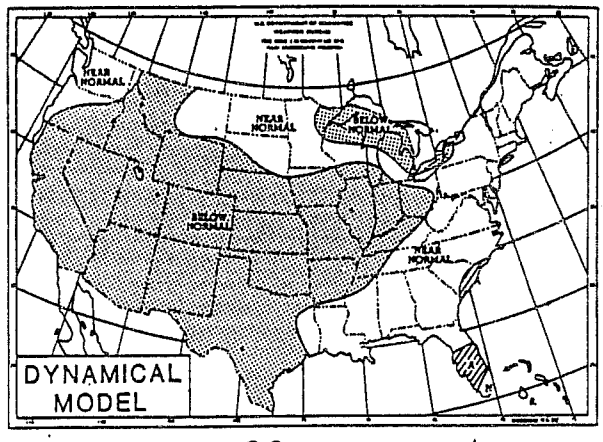
Thus, the EOF filter alone produces the largest improvement in 1-7 day mean proxy forecasts of 30-day means. In combination with EOF filtering the weighting and error correction procedures and LAF, on average, have only a small impact relative to the EOF filtering alone. As shown explicitly with time averaging (Sec. 3.2) and unweighted LAF (Sec. 3.2), the effects of postprocessing varies from case to case. More skillful predictions are generally improved, whereas poorer forecasts are not significantly affected. The basic character of the variability in skill over the DERF period remains essentially unchanged by statistical postprocessing. Significant gains in skill are likely only through improvements in the model and not from postprocessing alone.

3.5 Objective/Subjective Specification

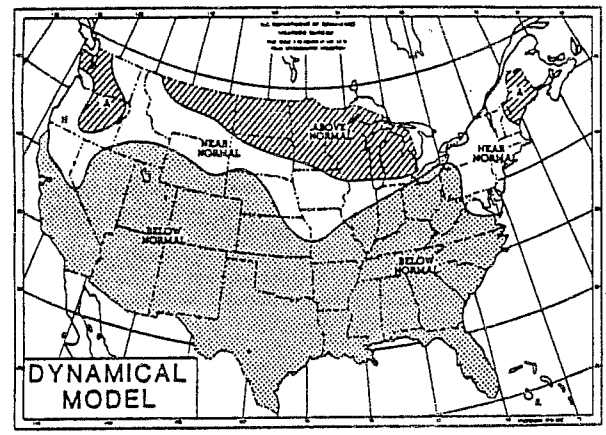
Objective/subjective specification refers here to obtaining from the model output parameters and quantities directly relevant to operational forecasting. One example is the surface temperature field derived from 30-day mean 700 mb height fields via Klein (1985) specification equations. The current operational procedure at the Climate Analysis Center (CAC) is to construct the 700 mb prognostic chart subjectively from combination of the latest 10-day medium range forecasts, the appropriate 700 mb one-month lag autocorrelation field, teleconnection and EOF patterns, analogs, and synoptic experience (Kalnay and Livezey, 1985). As a preliminary assessment of the skill of the DERF forecasts we compared for the limited number of matched comparisons possible the surface temperature derived from the model 1 to 30 day mean 700 mb height with the CAC operational 30-day Outlooks (Fig. 3.5.1). Skill scores are for 3-class predictions (above, below, normal) verified at 100 U.S. stations with values of 100 and 0 indicating perfection and no better than chance, respectively, given the climatological expectations of each class. The association between scores and actual forecasts is illustrated for two cases in Fig. 3.5.2. Overall, results are mixed with no obvious advantage of the DERF-based temperatures.



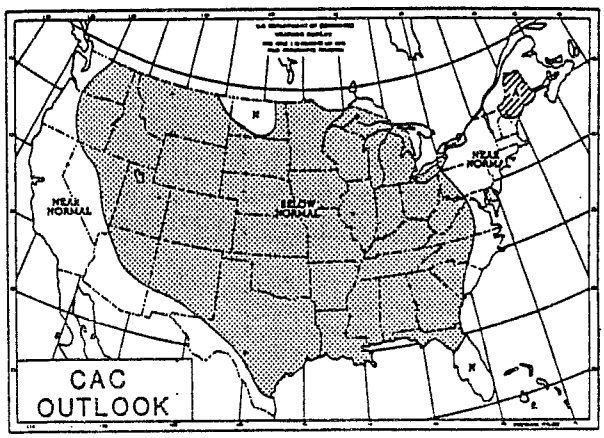




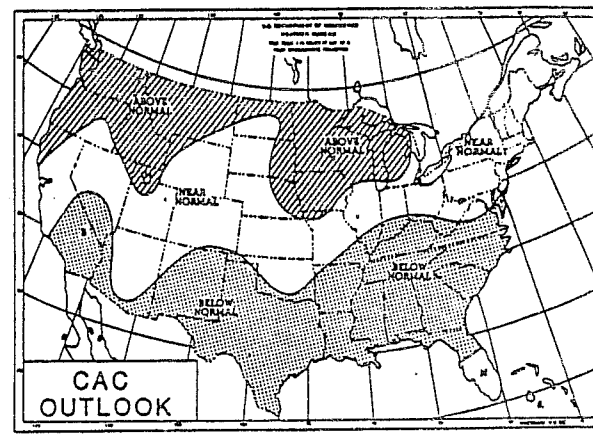
SKILL SCORE -26



SKILL SCORE 40



SKILL SCORE -29



SKILL SCORE 37

Fig. 3.5.2 DERF, CAC official outlooks, and verifying 3-class temperature fields for two cases, mid-Jan to mid-Feb, 1987 (left) and mid-Mar to mid-April, 1987 (right).

In appraising these results one should bear in mind the following; i) the closer relationship between DERF based and operational scores, compared to DERF and persistence results, reflects use of the first ten days of the model output as a tool in preparing the Outlooks, ii) the sample is limited with considerable variability of DERF results in absolute and relative terms, iii) dynamic predictions of 1-30 day means are generally not the best representation of a 30-day verification period, and iv) appropriate statistical postprocessing of model output, including ensemble averaging, has not yet been included in this comparison. A more rigorous, long term test is required before drawing conclusions about simply substituting DERF 30-day predictions for the present operational procedures. A large effort will be devoted to prescribe an optimal use of the dynamical output and to determine whether other parameters and/or levels (e.g., 850 mb temperature) and, most importantly, direct model forecasts of surface temperature can provide additional predictive information. The same applies to monthly mean precipitation, which also is obtained for operational purposes by specification from the 700 mb height field.

Another quantity relevant to operational forecasting that can be extracted from dynamic predictions is an estimate of storm activity. Fig. 3.5.3 shows that the "storm tracks" (inferred from bandpass filtered 500 mb height fields, Blackmon et al, 1977) in one 30-day integration are generally comparable to those in the verifying analyses; however, the predicted "storm tracks" are weaker and somewhat north of those observed, consistent with the model's systematic errors.

The storm tracks relate to the more general question of how much information DERF can provide on the variability within the 30-day time frame. Other examples include DERF's capabilities in specifying trends, extreme events, occurrence and length of spells, and regime transitions. To a large extent these items reflect the ability of the model to simulate the low-frequency behavior of atmospheric circulations, which can be represented in terms of teleconnection indices as defined by Wallace and Gutzler (1981). The overall performance of the model in predicting these indices, as calculated from 10-day means of 500 mb height fields, is displayed in Fig. 3.5.4 by the the correlaton between observed and forecast values as a function of forecast length. Changes in low-frequency circulations defined by the Western Atlantic (WA) and Western Pacific (WP) indices are the most predictable, and the Pacific North American (PNA) and

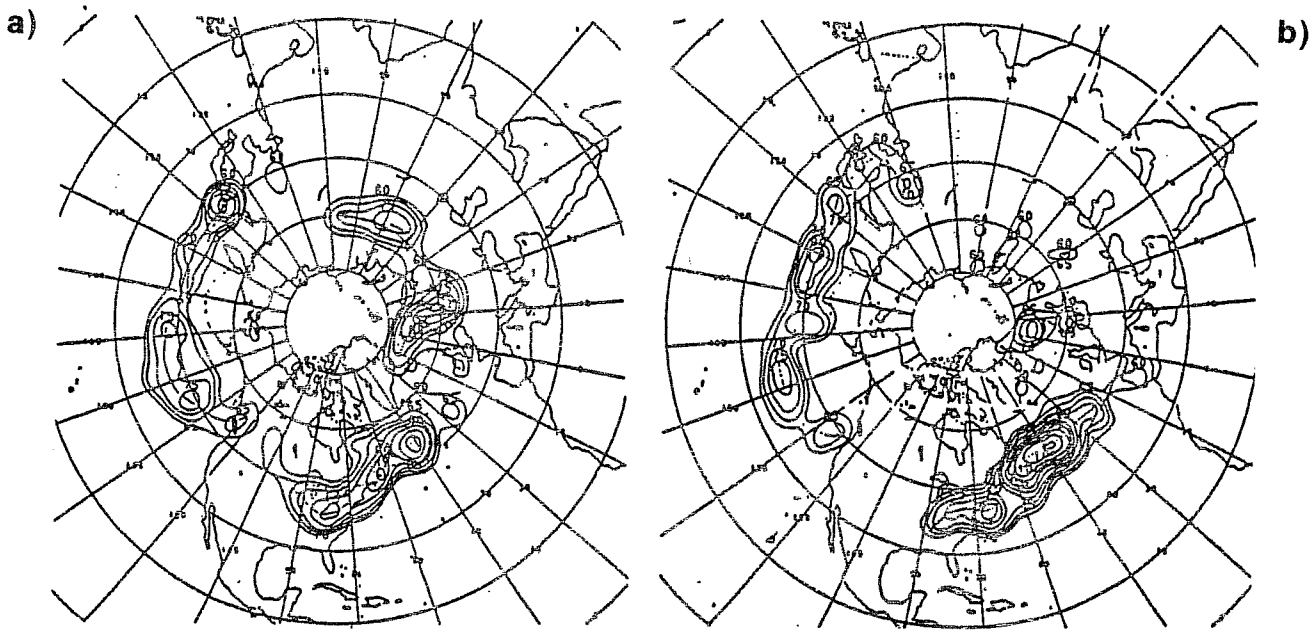


Fig. 3.5.3 Standard deviation of high-pss filtered 500 mb height data about the 30-day mean predicted (a) and observed (b) from 23 Jan 1987. Contour interval is 5 m.

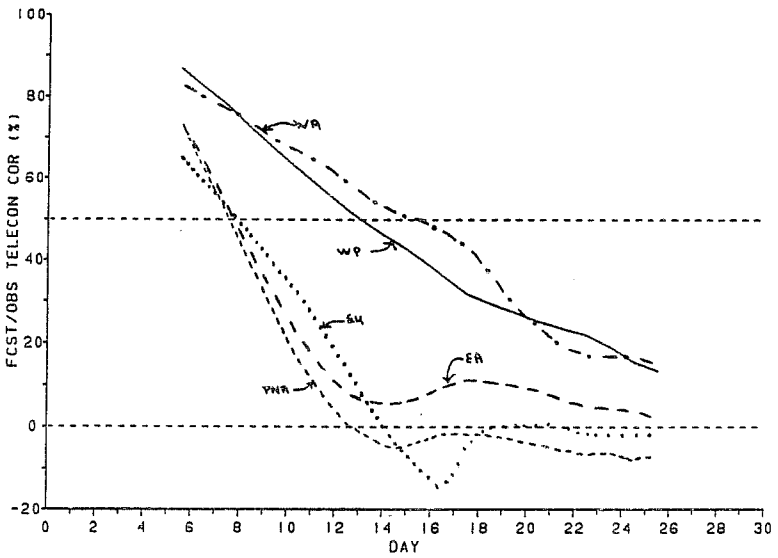


Fig. 3.5.4 108 case average correlation between observed and forecast teleconnection indices: WA, Western Atlantic, WP, Western Pacific, EA, Eastern Atlantic, WA, Western Atlantic, PNA, Pacific North American.

Eastern Atlantic patterns are handled most poorly by the model. It should be noted that the PNA index (by association) and EA pattern (by definition) are related to blocking activity, for which the model also has little skill by the 6 to 10 day range (see Sec 4.3).

4. ESTIMATION OF SKILL

It is clear from the results presented here and from other operational experience (e.g., Bengtsson and Simmmons, 1983) that the variability of skill at medium and extended ranges is very large and, therefore, the *a priori* estimation of this skill is a major outstanding problem in numerical weather prediction. Skill varies from day to day with longer term trends extending to weeks in length also frequently apparent (Fig. 2.2.1). Additionally, there is a marked regional dependence to error fields and verification scores. The ability to provide *a priori* guidance on the expected accuracy of predictions would enhance greatly the operational utility of forecasts. Indeed, at extended ranges, where the average skill is minimal, it is virtually essential to identify beforehand the relatively few good cases.

In this section we discuss preliminary results using four potential predictors of forecast skill. The first is the agreement between forecasts of LAF ensembles, as suggested by Kalnay and Dalcher (1987). The second is the persistence of the forecasts, as suggested by Chen (1988) and Palmer and Tibaldi (1987), and the third is the magnitude of the predicted anomaly following Branstator (1987). The fourth potential predictor, the PNA index following Palmer (1988), is discussed in context with the apparent dependence of skill on atmospheric regime, especially the occurrence of blocking.

4.1 Forecast Agreement and Forecast Skill

Several researchers have suggested that the dispersion between the members of an ensemble of forecasts may be a useful *a priori* estimate of forecast skill. Kalnay and Dalcher (1987) showed that Monte Carlo (MC) forecasts were useful for this purpose and suggested that similar results could be derived from LAF ensembles. The reason for this is that the atmospheric instabilities that are a major cause of the intrinsic loss of predictability will also tend to increase the dispersion between members of the forecast

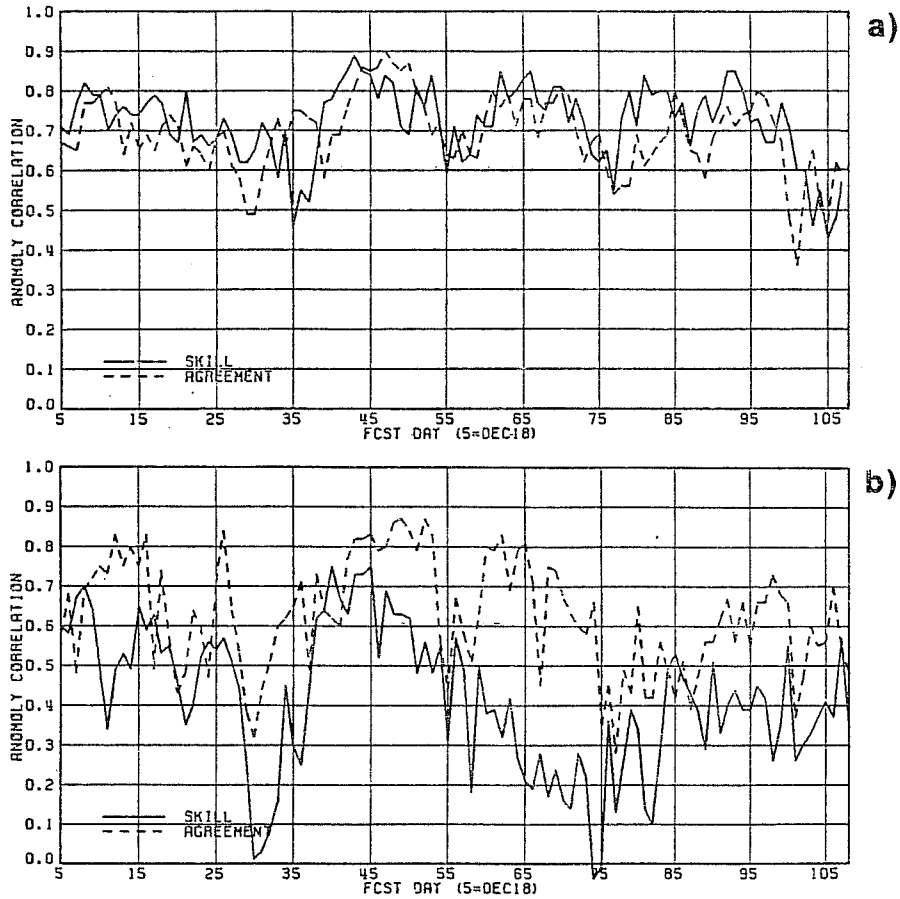


Fig. 4.1.1 Northern Hemisphere 500 mb skill (solid) and agreement (dashed) for 1-10 (a) and 1-25 (b) day means.

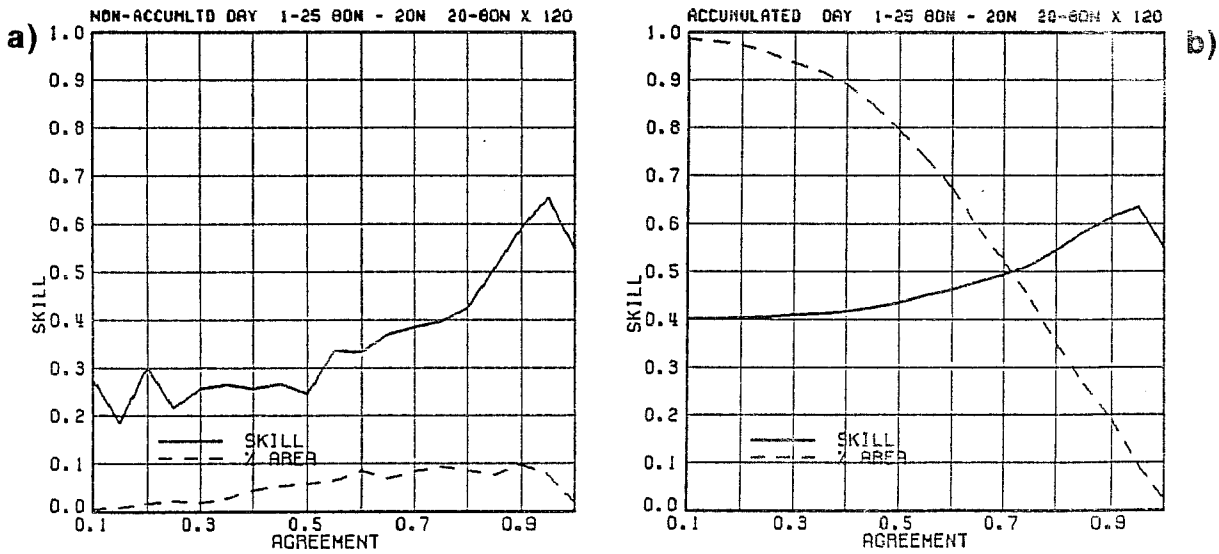


Fig. 4.1.2 Relationship between agreement and skill in all the regions from 20° to 80°N and 120° longitude wide in the 1-25 day forecast averages. The dashed line represents the percentage of the regions having the agreement indicated in the absicissa. See text for further discussion. (a) non-accumulated and (b) accumulated.

TABLE III. Northern Hemispheric Correlation Between Agreement and Skill (skill in parenthesis). Correlations significant at 95% are underlined.

Time Average (days)	Correlation
1-5	.22 (.90)
6-10	.36 (.44)
11-15	.12 (.15)
16-20	-.02 (.10)
1-10	.42 (.72)
6-15	.35 (.32)
11-20	.09 (.16)
1-15	.39 (.56)
6-20	.33 (.28)
11-25	.13 (.17)
1-20	.40 (.47)
6-25	.31 (.25)
1-25	.43 (.42)

ensemble. Low forecast agreement therefore should decrease the confidence in predictions, while high agreement generally should be associated with greater skill. This will be true, however, only to the extent that the model responds realistically to uncertainties in initial conditions, and error growth is not dominated by model deficiencies.

We define forecast agreement here as the average AC between the latest (base) prediction and the remaining members of LAF ensembles consisting of 5 consecutive forecasts 24 hours apart (Kistler et al, 1988). The measure of skill used is the AC of the latest forecast. Fig. 4.1.1a shows the day-to-day relationship between agreement and skill over the Northern Hemisphere for 1-10 day means. There is a clear relationship between the two with the correlation between them of 0.42, or about 17% of the explained variance. Fig.4.1.1b presents the same comparison but for the forecast days 1 to 25. Although now the forecast agreement is generally much higher than the skill, the variations between the two are still correlated at 0.43. Like the skill itself, however, the correspondence between agreement and skill is strongest in the early part of the forecasts. This is seen from Table III, which presents the skill/agreement correlations along with the skill for different averaging periods. Correlations above 0.26 are significant at 95% and are underlined. Not unexpectedly, the correspondence between agreement and skill decreases in the latter part of the forecasts and disappears when the average skill itself becomes insignificant.

Kalnay and Dalcher (1987) reported that the forecast of the medium range skill was more successful on a regional basis than on a hemispheric basis. We have attempted to estimate regional skill here by computing the AC over a moving window of typically 60^{o} latitude by 120^{o} degrees longitude. This simple method allows the estimate of regional skill as well as the generation of maps of expected and observed regional anomaly correlations which may be operationally useful. The optimal size of the window is probably dependent on the length of the forecast (R. Livezey, personal communication). For short forecasts, a smaller window (such as 30^{o} latitude by 60^{o} longitude, as used by McCalla and Kalnay, 1988) is more appropriate because it resolves synoptic scales well. For extended range forecasting a window much larger should be used, since slowly varying atmospheric modes have much larger horizontal scales.

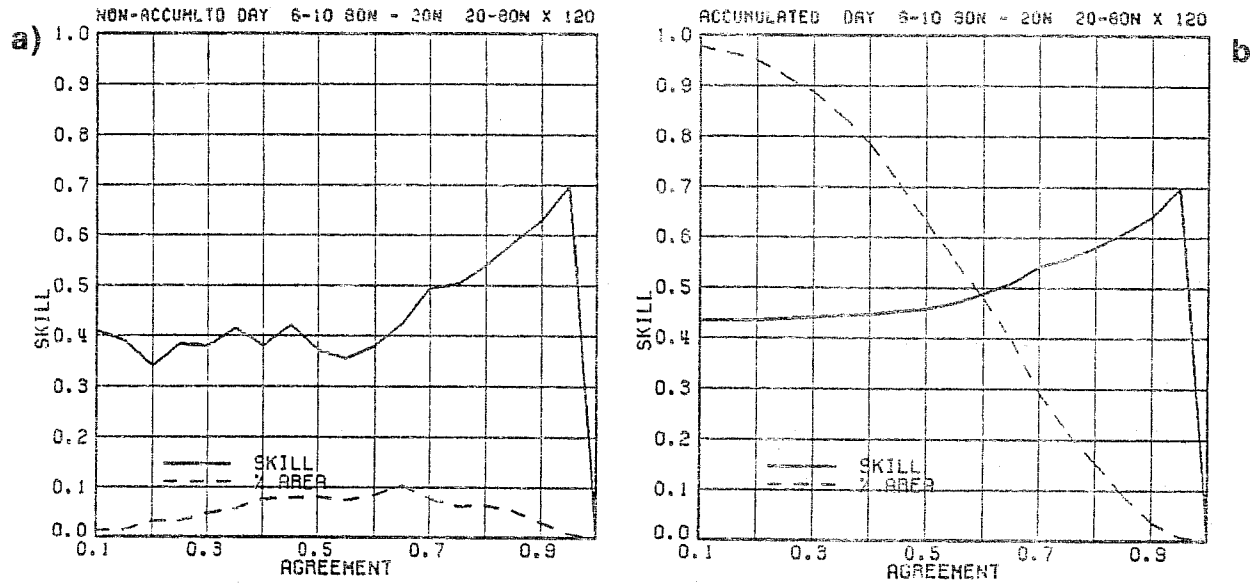


Fig. 4.1.3 As Fig. 4.1.2, except for 6-10 day forecasts.

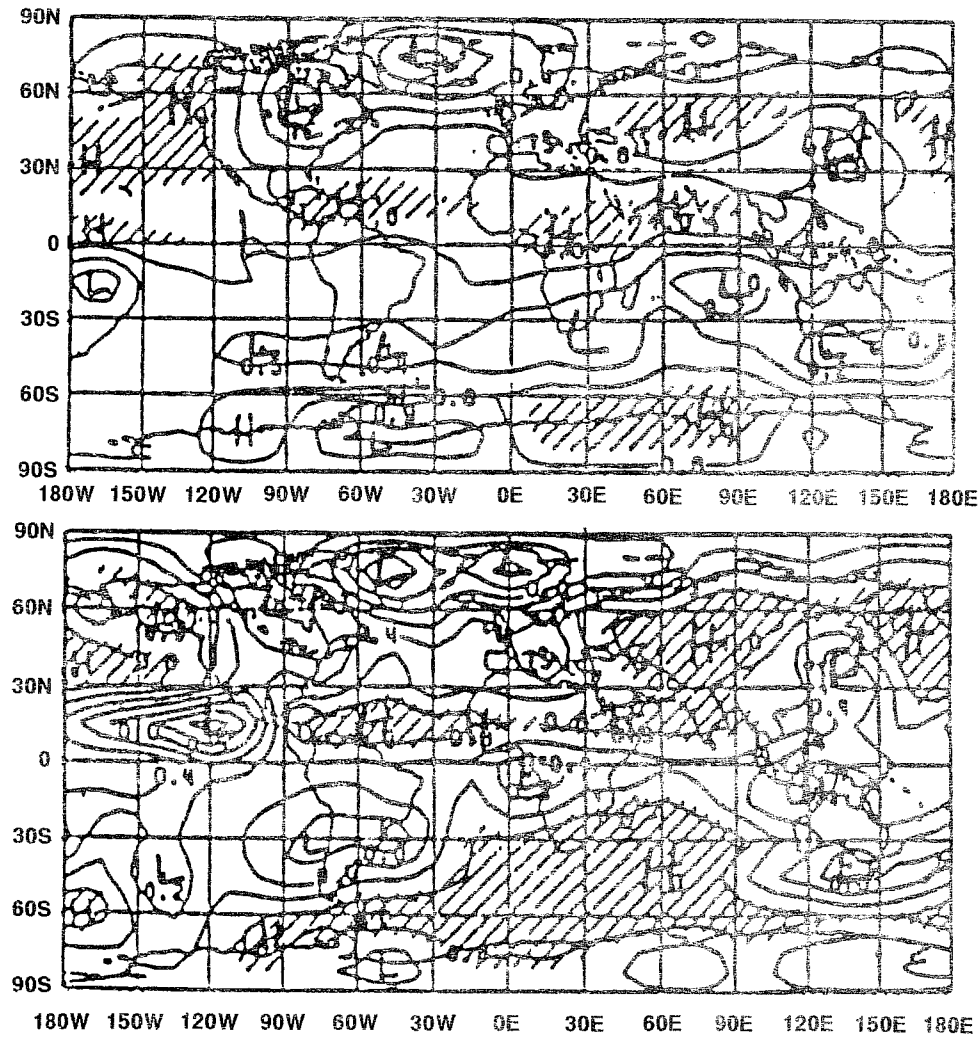


Fig. 4.1.4 Regional agreement (top) and observed regional skill (bottom) of 1-10 day mean using a moving window of 30° latitude by 60° longitude for forecast from 31 Jan (shaded greater than 70%).

Figure 4.1.2a presents the average 1-25 day forecast skill corresponding to all regions of 120° longitude from 20°N to 80°N, binned according to their forecast agreement. It shows that in regions where the agreement is low (less than about 0.50) the average forecast skill is approximately constant and very low (about 0.20 - 0.30); however, the average skill increases with agreement when the agreement is higher than 0.50. The dashed curve at the bottom shows the percentage of the areas having the indicated agreement. Fig. 4.1.2b shows a similar statistic now accumulated for all the areas that have at least the agreement indicated in the abscissa. It shows that in almost 100% of the regions the agreement is at least 0.10 and the average skill is 0.40. However, the 50% of the regions that have an agreement of at least 0.72 have an average skill of 0.50, and the top 20% of the regions that have an agreement of at least 0.89 have an average skill of 0.62.

It is interesting to make the same comparison for a shorter averaging period, say from days 6-10. Figure 4.1.3a shows again little dependence of the skill on agreement smaller than 0.60. For larger agreements the skill increases with agreement, so that in Fig. 4.1.3b we see that the top 20% of the areas which have an agreement of at least 0.75 have an average skill of about 0.43 for all regions.

The percentage of variance in skill explained by forecast agreement is only about 10% on a regional basis (including all regions). The skill in certain regions may be more predictable, and our results indicate that the skill of LAF forecasts is more predictable than that of the latest forecast presented here.

We are exploring the possibility of producing maps of expected regional forecast skill. Fig. 4.1.4 shows an example of an unusually successful prediction of the 10-day average forecast skill based on the agreement computed using smaller areas of 30° latitude by 60° longitude, and sliding them by 15° latitude and longitude.

Another way to depict the regional and time distribution of expected skill is the use of time-longitude (Hovmoller) diagrams. Fig. 4.1.5a presents the observed 10-day mean 500 mb height averaged over the band 55° to 65°N, and Fig.4.1.5b shows the sequence 1-10 day mean LAF ensemble predictions (5 members) verifying at the same analysis time. The absolute value of the

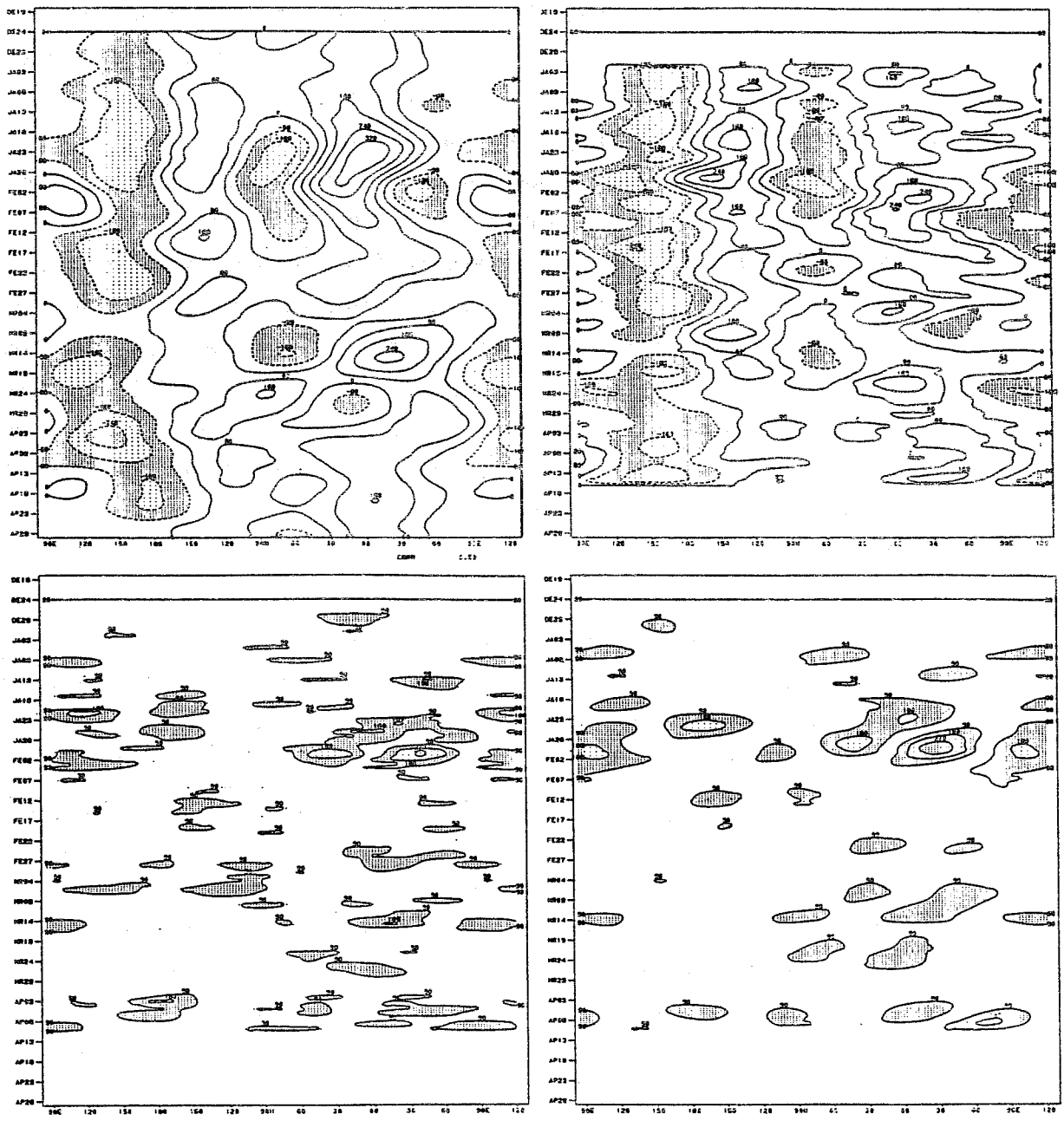


Fig. 4.1.5 (a) Moving 10-day average 500 mb height analysis between 55° and 65°N (zonal mean removed); contour interval 80m, (b) same for 1-10 day mean forecasts, (c) absolute error (b/c difference); contour interval 90m, and (e) forecast spread, i.e. average absolute difference between latest 10-day forecast and previous four predictions; contour interval 90m.

error (difference between these fields) is shown in Fig. 4.1.5c, and the "spread". i.e., the average absolute difference between the latest (base) forecast and the four previous forecasts verifying on the same 10 days, is diplayed in Fig.4.1.5d. The forecast spread is noisier than the error, so that the correlation between these two fields is only 0.28. Nevertheless, there is a fairly close correspondence overall between the errors and spread, especially during the latter half of January around the Greenwich Meridian where both errors and spread reach their maximum values in association with the evolution of a major blocking event. Finally, the degree of correspondence between errors and spread seen in the Hovmoller diagrams is also apparent in sequences of horizontal depictions of these same fields (not shown). Both sorts of display may be quite useful in an operational framework.

4.2. Forecast Persistence and Forecast Skill

Atmospheric instabilities are a basic cause of forecast error growth (Lorenz, 1982). This implies that persistent situations, in which the atmosphere is more stable, may also be more predictable. For this reason, Palmer and Tibaldi (1987) and Chen (1988) have suggested the degree of forecast persistence as an additional predictor of skill. We test this hypothesis in Fig. 4.2.1 , which presents the same statistics as Fig. 4.1.2, but now using as predictor, in the abscissa, the forecast persistence, defined as the anomaly correlation between the time average of the forecast for days 1 to 5 with the average of the forecast for days 6 to 10. Clearly other indices of forecast persistence also may be used (Chen, 1988). In this case our index provides some discrimination of forecast skill, although less so than the forecast agreement. For example, in Fig. 4.2.1 we see that the 50% of the areas that have at least a forecast persistence of 0.60 have an average skill of about 0.46, and the 20% that have forecast persistence of at least 0.75 have an average skill of about 0.48 (compared to 0.62 in the case of forecast agreement). For the shorter 6 to 10 day averaging period, forecast persistence provides overall less discrimination of skill than does forecast agreement (compare Figs. 4.2.2 and 4.1.3), although it is better at low levels of skill.

It is clear that these two possible predictors of skill, forecast agreement and forecast persistence, need not be independent, since the more stable areas with high predictability also may be those with highest

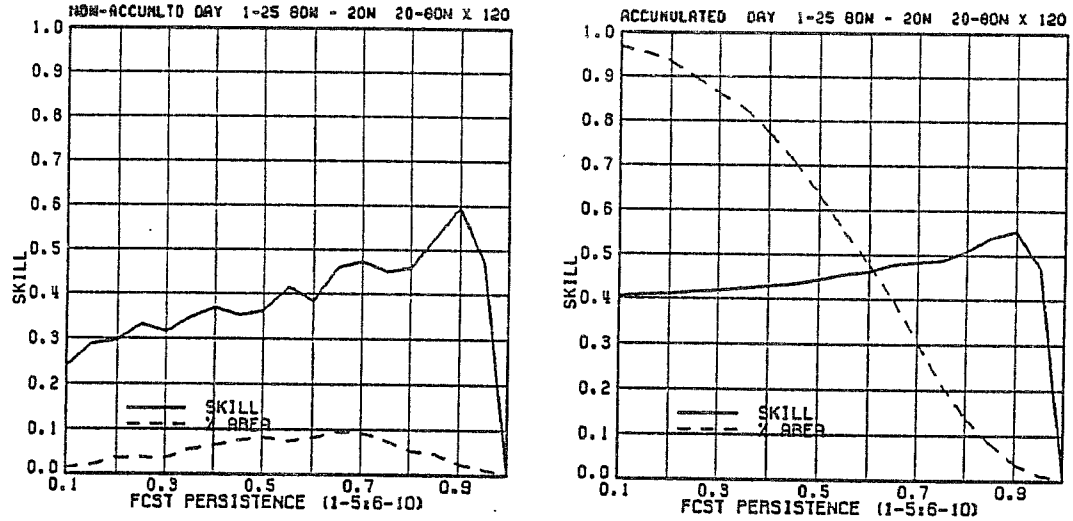


Fig. 4.2.1 As Fig. 4.1.2, except the relationship between forecast persistence and skill.

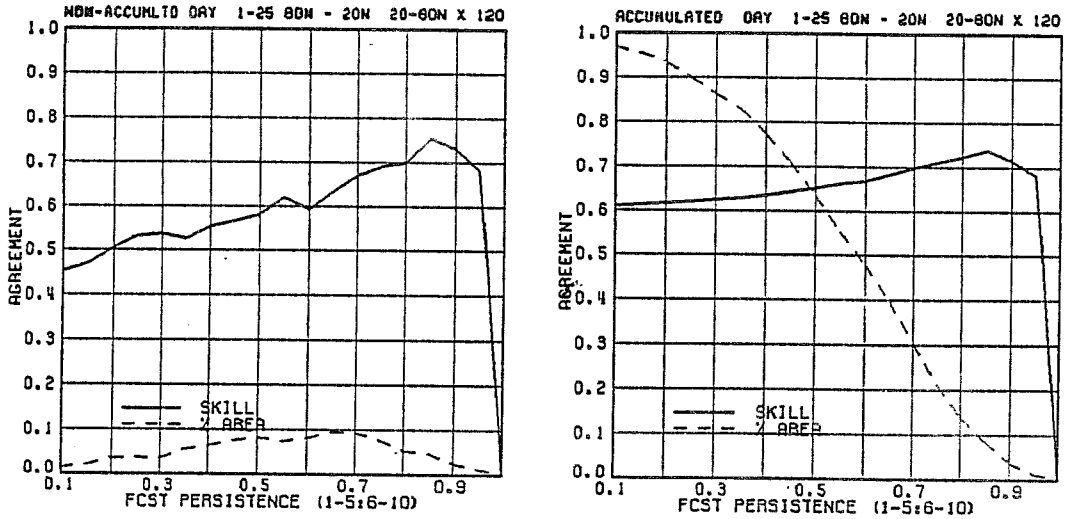


Fig. 4.2.2 As Fig. 4.2.1, except for 6-10 day forecasts.

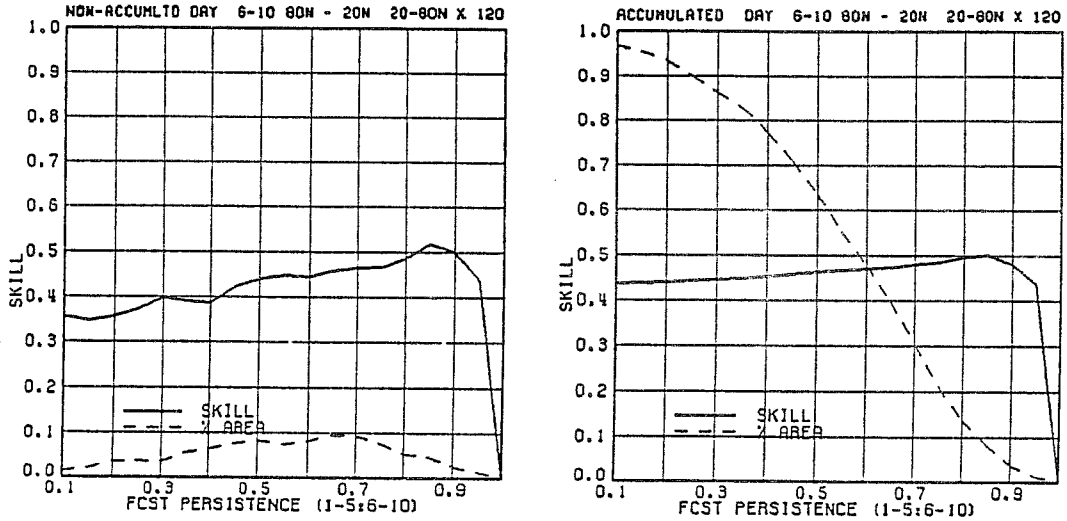


Fig. 4.2.3 As Fig. 4.1.2, except for the relationship between forecast persistence and agreement.

persistence. Fig. 4.2.3 shows the agreement as a function of persistence, indicating that there is indeed some relationship between the two.

Another possible predictor of skill is the amplitude of the anomaly (Branststor, 1987). A similar analysis using standardized anomaly amplitude as predictor (not shown) indicates virtually no relationship between skill and anomaly amplitude except for small anomalies, for which the skill (measured by anomaly correlation) also becomes very small. This is not a true indication of loss of skill but a result of the fact that for small anomalies the signal-to-noise ratio in the amomaly correlation becomes very poor.

4.3 Circulation Regime and Forecast Skill

Recently Palmer (1988) and O'Lenic and Livezey (1988) have sought to determine relationships between circulation regime and errors in medium range forecasts over several years. Each of these studies employs EOF characterizations of low-frequency anomalies in the forecasts and/or initial analyses to stratify the distribution and magnitude of prediction errors. As a first step, we have adopted a more subjective, synoptic approach (e.g., Gronaas, 1983) to determine associations that might exist between skill and atmospheric flow in the DERF Phase II data.

The dominant relationship we have found so far is that between forecast skill and evolution of the the planetary wave structure associated with blocking events. This is probably not surprising given the many empirical and theoretical investigations of the predictability of blocking (e.g., Bengtsson, 1981, Legras and Ghil, 1985), but the nature and degree of the association is remarkable. To illustrate, Fig. 4.3.1 displays the sequence of hemispheric AC scores of 6-15 day means oriented to permit direct comparison with a time-longitude plot of blocking activity. To filter the higher frequency component of variations in skill, the scores plotted are for 5-case running means. The blocking index is the difference between the analyzed (10-day mean) 500 mb heights at 60° and 40° N (Lejenas and Okland, 1983). Positive values, i.e., higher heights to the north (or, equivalently, an easterly component of the geostrophic wind), indicate an anticyclone at northerly latitudes and provide a generally reliable measure of the blocking events discerned from subjective appraisal of charts.

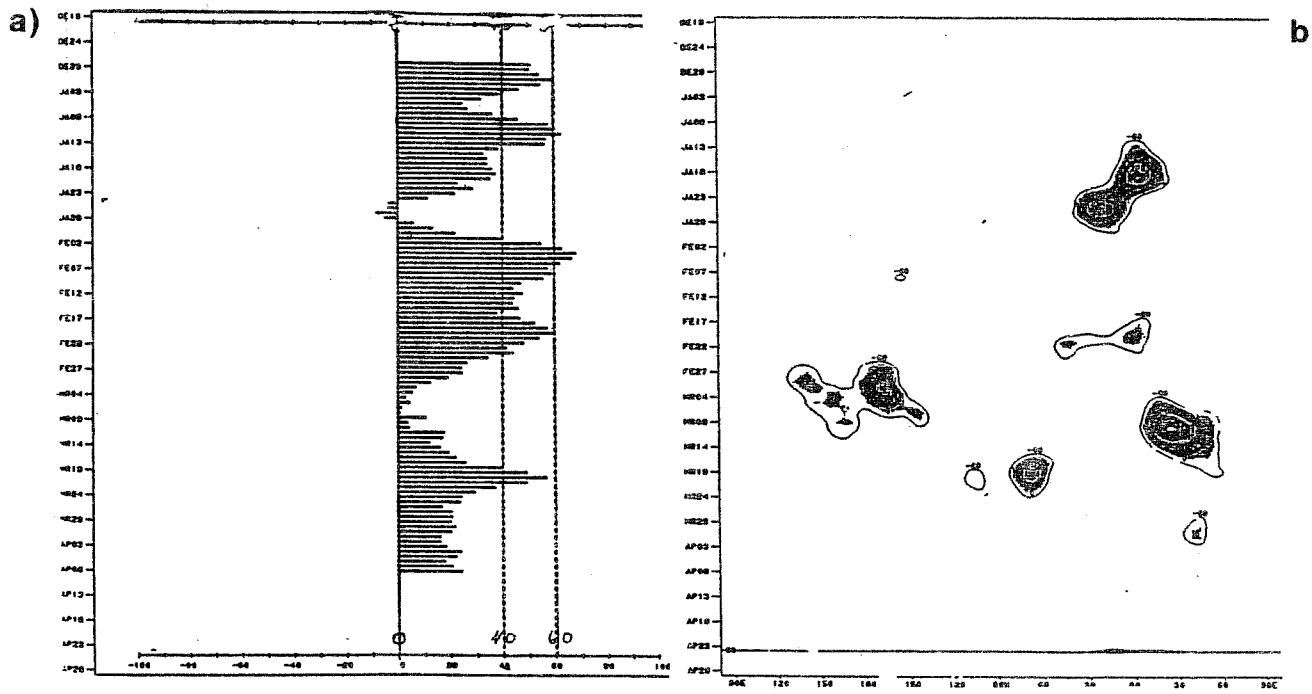


Fig. 4.3.1 (a) Contiguous series of 500 mb height AC for predictions of 6-15 day means plotted on the centre of the verifying period, (b) time-longitude plot of blocking index derived from 10-day mean 500 mb height analyses (shaded greater than -30), contour interval 60m.

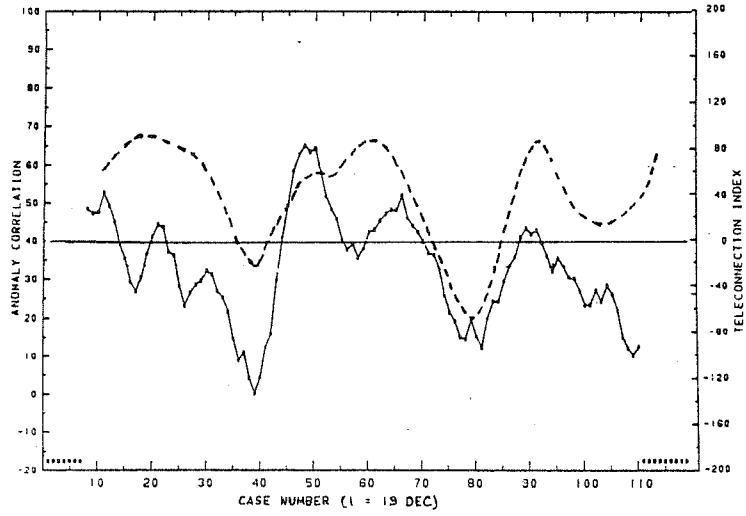


Fig. 4.3.2 AC of NH 500 mb height (solid) of 6-15 day means and PNA teleconnection index (dashed).

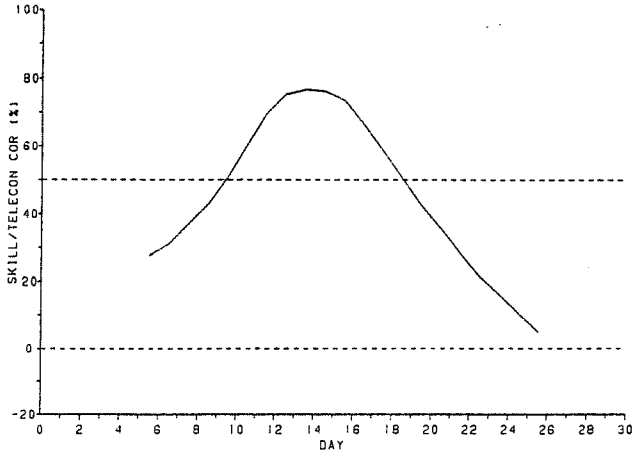


Fig. 4.3.3 Correlation between NH 500 mb height forecasts (10-day means) and PNA teleconnection index vs. length of forecast.

The first major blocking event occurs over the eastern Atlantic and western Europe during the period from 13 to 27 January, 1987. The AC clearly shows a dramatic decline in skill with the onset of the block and more dramatic recovery coincident with its demise. The especially poor cases, with the AC dropping to values less than zero, occur as the block retrogresses in association with what 500 mb charts show are major cyclonic developments immediately up and down stream. The next major blocking episode occurs in the Pacific beginning around 22 February, and it too is accompanied by a marked decrease in the AC. The lowest scores correspond to a period of overlap between this event and renewed blocking activity in the Atlantic during the first week of March. The verifications recover sharply as the Atlantic block wanes, albeit in the face of a relatively minor event over North America just after mid-month.

Certainly not all the variations in skill are related directly to blocking, as evident from the the case to case variability within and between blocking occurrences, especially in the unsmoothed scores (Fig. 2.1.1c). A large component, however, clearly is.

It should be noted that the correspondence discussed above is between blocking over what might be considered a relatively limited region and the hemispherically averaged AC. An explanation for this is that errors associated with the evolution of the blocks are large and typically extend well beyond their immediate vicinity. This is especially true for the blocking in the Atlantic sector. Another consideration is that from a wider perspective large changes occur in the planetary scale circulation over the Northern Hemisphere coincident with, if not directly linked to, the incidence of blocking. In particular we observe that in each instance the onset of blocking coincides with a transition from positive to negative values of the PNA index. Thus, the blocking apparently is only one aspect of much broader changes in circulation regime which, in turn, are related to variations in forecast skill. Moreover, there is a distinct suggestion this relationship is a potential predictor of forecast skill, since the beginning in the decline of the PNA index preceeds the blocking which we have shown is contemporary with variations in predictive skill. This is clearly shown in Fig. 4.3.2, which displays the time trace of AC for 6 to 15 day mean predictions together with the PNA index about the initial (not verifying) conditions. The correlation between these two curves is 0.62. The correlation is actually largest (0.77) for the AC of the ten day means

centered on day 14 (Fig.4.3.3). Thus, a large part of the variability in forecast skill, especially at mid ranges, can be anticipated by the values of the PNA index of the initial circulation regime. The more predictable cases generally coincide with positive values, while the poorer forecasts occur when the PNA becomes negative. This result is consistent with the studies of Palmer (1988) and O'Lenic and Livezey (1988). Whether the additional association found here between the PNA and blocking activity is more than coincidental needs further exploration.

The component of skill variability related to blocking in the 6-15 day means shown above quite clearly results from the inability of the model to simulate properly the evolution of blocking circulations. To examine further model capabilities at other time ranges comparisons were made of time-longitude plots of the observed and predicted blocking indices of 5-day means as a function of forecast length (Fig. 4.3.4). Prospects for capturing blocking episodes are very good in the first 5 days but decrease to virtually nil by the 6-10 day period. This result is similar to that of Tibaldi and Molteni (1987) with regard to the ECMWF model. As suggested above this decreasing ability to capture blocks is related to the large spread (low agreement) amongst members of LAF ensembles in regions of blocking. Additionally, even though blocking is simulated reasonably well when it develops within the first few days of integration, the blocks usually are not maintained properly as the forecasts proceed. This is probably related to the non-persistent character and almost discontinuous evolution of the blocking circulations during this particular winter. In more persistent and less complex situations other investigations (e.g., Miyakoda, 1983) have provided examples of remarkably successful predictions at extended ranges.

There are very good forecasts at extended ranges in our data set associated with blocking, but the association is rather curious in that it relates to the phenomenon referred to as the "return of skill". Fig. 4.3.5 shows the AC curve of 10-day means for the forecast from 7 January. The AC decreases steadily to zero by the middle of the forecast, but then shows a remarkable recovery to levels at the end of the integration comparable to those at the beginning. The Hovmoller diagrams and AC trace shown in Fig. 4.3.6 demonstrate what seems to account for this behavior and illustrate further the considerabe influence of blocking on verification scores. Fig. 4.3.6a is for the observed 10-day mean 500 mb height fields averaged over the 55-

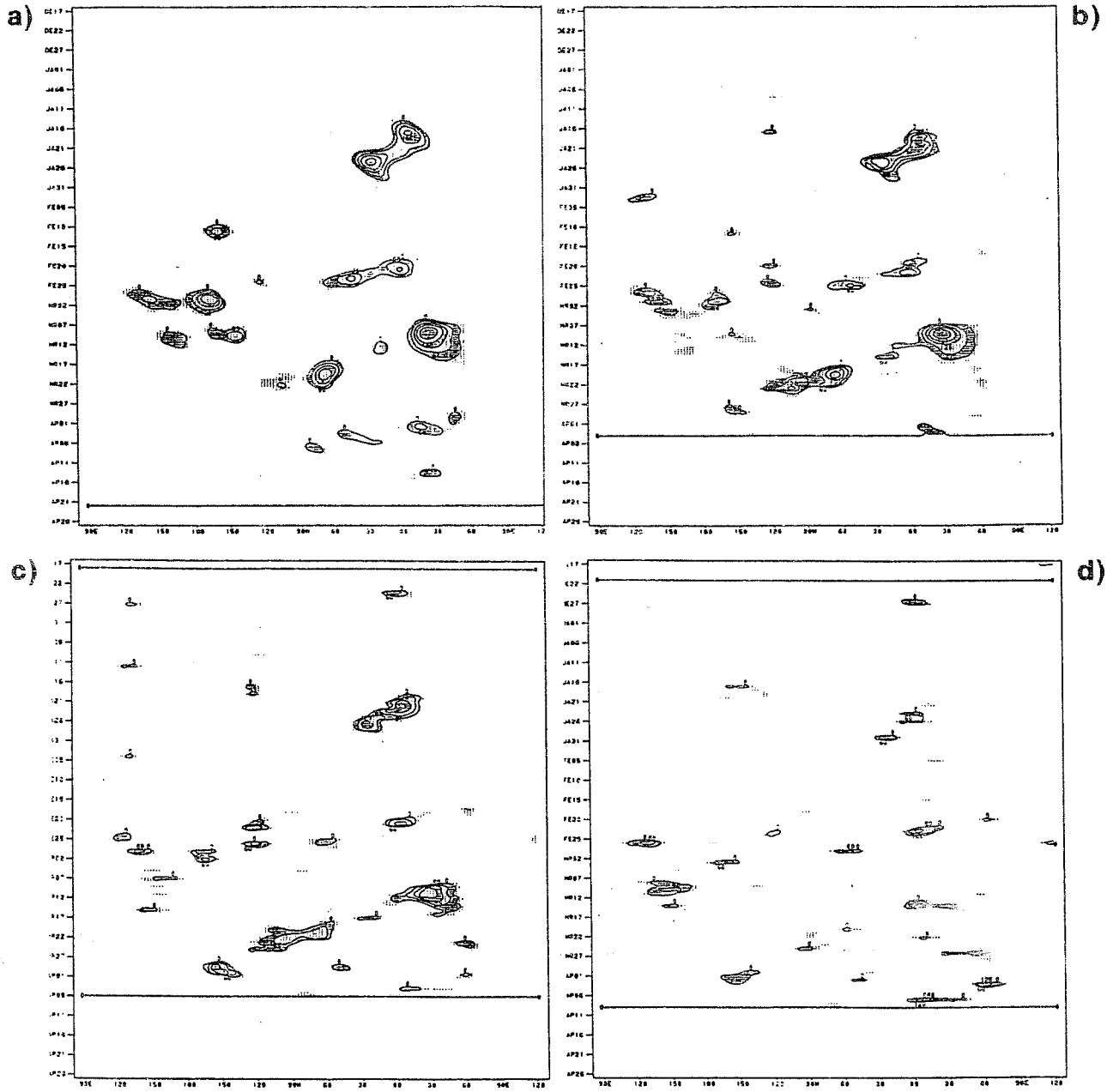


Fig. 4.3.4 Time-longitude of blocking index derived from 5-day mean 500 mb height analyses (a), and 1-5 (b), 3-7 (c) and 6-10 (d) day forecasts.

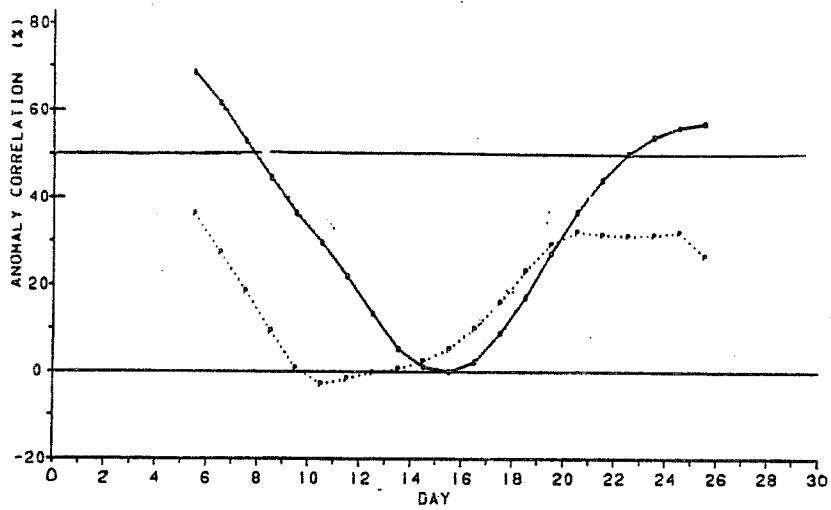


Fig. 4.3.5 AC of NH 500 mb height of forecast from 7 Jan (10-day means). Forecast, solid: persistence score, dot-P.

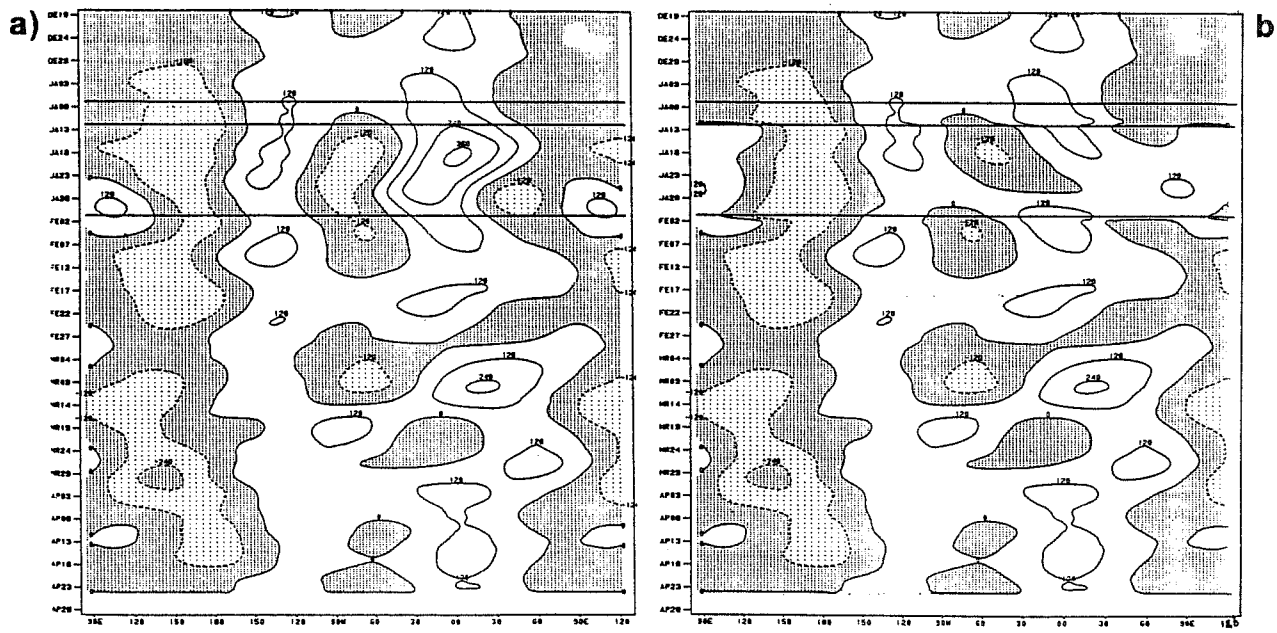


Fig. 4.3.6 (a) Time-longitude plot of observed 10-day mean 500 mb height (zonal mean removed); (b) as (a), except forecast from 7 Jan inserted, with 3 horizontal lines indicating time of initial conditions, and centres of first and last 10-day verification periods.

65oN. Fig. 4.3.6b is the same, except the forecast is substituted for the analyses at the appropriate verification times. The decline of skill is closely related to failure of the model to simulate the evolution of the ridge and trough system which can be identified with the first of the previously mentioned blocking events. Towards the middle of the forecast period the block begins to decay, and the skill score recovers. The recovery occurs as the atmosphere moves towards a state more like that predicted and which also is closer to the initial conditions. This is evident both from the Hovmoller diagrams and Fig. 4.3.5, which shows that the persistence forecast, as well as the dynamical prediction, display a return of skill. The term skill here, however, is probably not appropriate, for the apparent success of the forecast reflects a seemingly fortuitous set of circumstances and not the model's ability to simulate the relevant atmospheric circulations. Any hope of a priori recognition of situations like this probably will require combining information on blocking occurrence from the early part of the dynamical forecasts with the climatolgy of relatively transient blocking episodes and/or other aspects of the predicted circulation.

5. SUMMARY AND CONCLUSIONS

We have presented a summary of early results from the NMC DERF experiments, particularly the so called Phase II effort in which 108 contiguous 30-day forecasts were generated for the period 14 December 1986 to 30 March 1987 using the then operational Medium Range Forecast (MRF) model. This rich data set has now been condensed from 216 to 5 tapes through selection of fields and T20 truncation, and it is available to the research community (Schubert et al, 1988).

The extended integrations allow us for the first time to assess the climate drift of the model and, thereby, the question of how systematic are the "systematic" errors. Already the results have demonstrated some serious model deficiencies, such as the tendency to zonalization (later alleviated by the introduction of gravity wave drag) and the error drift in the stratosphere, which does not saturate even by day 30.

One of the most striking results is that the skill at medium and extended ranges varies strongly from case to case, and this is reflected in the predictions of 1-30 day means. In the 1-30 day means the dynamical model is

almost always more skillful than persistence, an encouraging outcome since persistence is very competitive with the CAC's operational Monthly Outlooks. However, on average, the best estimate of the 30-day mean circulation is not the forecast 30-day mean, but the average of only the first 7 to 10 days. Beyond 10 days there are many cases of skillful predictions, but, since the average skill is low, the operational utility of extended runs requires a means for a priori discrimination of the good from the poor cases.

We have begun to consider the problem of enhancing the forecast skill by statistical postprocessing, including time averaging, LAF, correction of systematic errors and EOF expansion. Results indicate that on average; i) 10-day means are more predictable than daily fields, ii) simple LAF with equal weights results in degradation of skill early in the forecasts (when "older" predictions are given too much influence -less so with 12 than 24-hour spacing), but a small improvement later, iii) the correction of systematic errors determined a posterori ("empirical correction") produces far less spectacular improvements than those reported by Miyakoda (1986), and iv) in the context of 1-7 day means as proxies for 1-30 day average forecasts, EOF filtering alone produces the largest improvement - no additional improvement was obtained by correction for systematic errors based upon independent data or from LAF with weights based upon the approach of Dalcher et al (1988). On an individual case basis, a main finding is that these postprocessing procedures can significantly enhance skill, but only of those predictions which are already skillful. Postprocessing does not have a significant effect on poor forecasts, and gains here are likely only through model improvements.

We have examined four potential predictors of skill. Forecast agreement can explain about 17% of the skill variance both in the 1-10 and 1-25 day means. As true with the skill itself, the relationship between skill and agreement is strongest in the early part of the forecasts. Regionally, on average, forecast agreement explains about 10% of the variance, and forecast persistence an additional 5%. The magnitude of the forecast anomaly only indicates the apparent loss of skill at small anomalies where the AC is unreliable. The PNA index of the initial circulation regime is a remarkably good indicator of forecast skill at mid ranges, where the correlation between the PNA index and skill reaches 0.77 for the 10-day mean centered on day 14. This predictor has much less skill earlier in the

in the predictions (in agreement with Palmer's (1988) hypothesis that this relationship is associated with barotropic instability errors which develop slowly), and therefore it complements well the previous predictors.

A major finding is that the inability to predict the evolution of blocking events beyond a few days into the forecasts has a major influence upon the variability in forecast skill. Diagnostic studies are in progress to investigate predictability and its relationship to cyclone/planetary scale interactions associated with occurrence of blocking. Also, it is important to determine whether the relationship we found between the antecedent PNA index and later blocking holds for longer data sets.

The research program in DERF at NMC has only begun. The DERF II data offers continued opportunities (to the research community, as well as NMC) for addressing many important problems in prediction and predictability. The lower resolution LAF and MC experiments (DERF Phase III) now being run, together with continuing efforts to develop schemes for predicting skill and to optimize the information content of model output, are intended to complement DERF Phase II and lead to a multi-year experiment beginning in 1989 whose ultimate objective is incorporating DERF into the operational environment.

References

Bengtsson, L., 1981: Numerical prediction of atmospheric blocking - A case study. Tellus, 33, 19-42.

_____, L. and A. Simmons, 1983: Medium-range weather prediction-operational experience at ECMWF. Large Scale dynamical processes in the atmosphere. Eds. B. Hoskins and R. Pearce, Academic Press, 337-363.

Blackmon, M., J. Wallace, N. Lau and S. Mullen, 1977: An observational study of the Northern Hemisphere wintertime circulation. J. Atmos. Sci., 34, 1040-1053.

Branstator, G., 1987: The variability in skill of 72-hour global-scale NMC forecasts. Mon. Wea. Rev., 114, 2628-2639.

------, T., 1988: Medium and extended range predictability and stability of the Pacific/North American mode. Quart. J. Roy. Met. Soc., 114, 691-173.

Chen, W., 1988: A new method in forecasting forecast skill. Eighth Conference on Numerical Weather Prediction, Baltimore Maryland, February 22-26, 1988, Amerişan Meteorological Society, Boston, Mass., 647-652.

Dalcher, A., E. Kalnay and R. Hoffman, 1988: Medium range lagged average forecasts. Mon. Wea. Rev., 116, 402-416.

Gronaas, S.,1983: A pilot study on the prediction of meduim range forecast quality. ECMWF Tech. Memo. No. 119.

Hoffman, R. and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. Tellus, 35A, 100-118.

Hollingsworth, A., K. Arpe, M. Tiedtke, M. Capaldo and H. Savijari, 1980: The performance of a medium range forecast model in winter - impact of physical parameterizations. Mon. Wea. Rev., 108, 1736-1773.

Kalnay, E. and R. Livezey, 1985: Weather predictability beyond a week - an introductory review. Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics, LXXXVIII Corso, Soc. Italiana di Fisica, Bologna, Italy, 311-346.

_____, E. and A. Dalcher, 1987. Forecasting forecast skill. Mon. Wea. Rev., 115, 249-356.

Kistler, R., E. Kalnay and S. Tracton, 1988: Forecast agreement, persistence and forecast skill. Eighth Conference on Numerical Weather Prediction, Baltimore, Maryland, February 22-26, 1988, American Meteorological Society, Boston, Mass., 641-646.

Klein, W., 1985: Space and time variations in specifying monthly mean surface temperature for the 700 mb height field. Mon. Wea. Rev., 113, 277-290.

Legras, B. and M. Ghil, 1985: Persistent anomalies, blocking and variations in atmospheric predictability. J. Atmos. Sci., 42, 433-471.

Leith, C., 1974: Theoretical skill of Monte Carlo forecasts. Mon. Wea. Rev., 102, 409-418.

Lejenas, H., and H. Okland, 1983: Characteristics of Northern Hemisphere blocking as determined from a long time series of observational data. Tellus, 35A, 350-362.

Livezey, R. and J. Schemm, 1988: The relative utility of persistence and medium-range dynamical forecasts. Eighth Conference on Numerical Weather Prediction, Baltimore Maryland, February 22-26,1988, American Meteorological Society, Boston, Mass., 478-479.

Lorenz, E., 1982: Atmospheric predictability experiments with a large numerical model. Tellus, 34, 505-513.

------, E., 1977: An experiment in nonlinear statistical weather forecasting, Mon. Wea. Rev., 105, 590-602.

McCalla, C. and E. Kalnay, 1988: Short and medium range skill and the agreement between operational models. Eighth Conference on Numerical Weather Prediction, Baltimore Maryland, February 22-26, 1988, American Meteorological Society, Boston, Mass., 634-640.

Mansfield, D., 1986: The skill of dynamical long-range forecasts, including the effect of sea surface temperature anomalies. Quart. J. Roy. Met. Soc., 112, 1145-1176.

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

------, K., T. Gordon, R. Caverly, W. Stern, J. Sirutis, and W. Bourke, 1983: Simulation of a blocking event in January 1977. Mon. Wea. Rev., 111, 846-869.

-------, K., and J. Sirutis, 1985: Extended Range Forecasting, Advances in Geophysics, Vol. 28, Part B, S. Manabe, Ed., Academic Press, 55-85.

Mo, K., 1988: The predictability of low-frequency patterns in the NMC MRF. Eighth Conference on Numerical Weather Prediction, Baltimore Maryland, February 22-26, 1988, American Meteorological Society, Boston, Mass., 628-633.

Molteni, F., U. Cubasch and S. Tibaldi, 1986: 30- and 60- day forecast experiments with the ECMWF Spectral Models. Workshop on Predictability in the Medium and Extended Range, 17-19 March, 1986, ECMWF, Reading, U.K., 51-107.

O'Lenic, E. and R. Livezey, 1988: Relatinships between low-frequency anomalies and errors in meduim- and extended- range forecasts. Eighth Conference on Numerical Weather Prediction, Baltimore Maryland, February 22-26, 1988, American Meteorological Society, Boston, Mass., 624-627.

Palmer, T. and S. Tibaldi, 1987: Predictability studies in the medium and extended range. ECMWF Tech. Memo. 139.

_____, T., 1988: Medium and extended range predictability and stability of the Pacific/North American mode. Quart. J. Roy. Met Soc., 114, 691-713.

Saha, S. and J. Alpert 1988: Systematic errors in NMC medium range forecasts and their correction. Eighth Conference on Numerical Weather Prediction, Baltimore Maryland, February 22-26, 1988, American Meteorologoical Society, Boston, Mass.,472-477.

Schubert, S., F-C Chang, W. Lau and R. Kistler, 1988: A reduced version of the NMC DERF II Data Set. Experimental Climate Forecast Center, NASA/Goddard Laboratory for Atmospheres (in press).

Shukla, J., 1981: Dynamical predictability of monthly means. J. Atmos. Sci., 38, 2547-2572.

Tibaldi, S. and F. Molteni, 1987: On the operational predictability of blocking. ECMWF seminar on the Nature and Prediction of extratropical weather systems, Vol. II, 7-11 Sept. 1987, Reading, U.K., 329-371.

van den Dool, H., 1985: Prediction of daily and time-averaged temperature for lead times of 1-30 days. Ninth Conference on Probability and Statistics in Atmospheric Sciences, October 9-11, 1985, Virginia Beach, Va. American Meteorological Society, Boston, Mass, 144-153.

Wallace, J. and D. Gutzler, 1981: Teleconnections in the geopotential height field during the Norhtern Hemisphere winter. Mon. Wea. Rev., 109, 784-812.

White, G., 1988: Systematic Performance of NMC Medium-Range Forecasts 1985-88. Eighth Conference on Numerical Weather Prediction, Baltimore, Maryland. American Meteorological Society, Boston Mass.,466-471.