OPERATIONAL ENSEMBLE PREDICTION AT THE NATIONAL METEOROLOGICAL CENTER: PRACTICAL ASPECTS

M. Steven Tracton and Eugenia Kalnay

National Meteorological Center
Washington, DC 20233

April 1993

NMC Office Note 392

Submitted to *Weather and Forecasting*

1. Introduction

A new era in operational numerical weather prediction (NWP), particularly as it relates to medium- and extended- range forecasting, began on 7 December 1992 when the National Meteorological Center (NMC) began to perform daily ensemble predictions. Instead of a single prediction through day 10, the new system provides forecasters with 14 Medium-Range Forecast (MRF) model predictions valid for the same 10-day period. The construction of the forecast ensembles is based on a combination of time lagging (Lagged-Average Forecasting or LAF, Hoffman and Kalnay, 1983), and a new method denoted "Breeding of Growing Modes" or BGM (Toth and Kalnay, 1993).

The operational implementation of ensemble forecasting at NMC represents a fundamental change in its approach to NWP. Before 7 December 1992, a single global forecast to 10 days was computed once each day with the highest resolution model possible (triangular truncation T126, equivalent to 105 km). This was a "deterministic" approach, in which computer resources were devoted to constructing the most accurate possible single forecast. Now, NMC has taken a more "probabilistic" approach, in which the same computer resources are used to provide sets of multiple integrations with lower-resolution (T62, equivalent to 210 km) versions of the MRF model.

In adopting the ensemble approach we explicitly recognize that forecasts, especially for the the medium range (and beyond), should be considered stochastic, not deterministic in nature. That is, because of the inevitable growth of differences between forecasts started from even very slightly different initial conditions (Lorenz, 1963), there is no single valid solution, but rather a range of possible solutions. This is illustrated schematically in Fig. 1, where forecast trajectories starting from slightly perturbed initial conditions span a much larger range of possibilities after a few days of integration. As the figure suggests, the forecasts can be considered deterministic for the shorter ranges, since the solutions are close to each other; but, beyond a somewhat arbitrary time, one cannot ignore the fact that there is a range of plausible outcomes.[1]

In general, the ensemble mean should provide a better forecast than most individual members, since some errors in the individual forecasts should cancel when averaged. More importantly, as suggested by Fig. 1, the divergence of forecasts within an ensemble can provide a means to estimate the range and likelihood of possible outcomes. For example, the forecasts may cluster into just a few solutions, and it may be possible to assign

---

[1]Even at short ranges, differences between forecasts in some parameters or fields can be operationally significant, and the stochastic nature of numerical predictions should be considered. This is discussed further in secton 5.

probabilities to the differing scenarios based on the number of forecasts in each cluster (denoted A and B in the schematic figure). By adopting the ensemble approach and allowing for the possibility of providing reliable information on forecast uncertainties, the ultimate goal of NMC shifts away from just maximizing the *skill* of model forecasts toward enhancing the *total utility* of NWP products.

In practice, operational ensemble forecasting is not a new concept. For years forecasters at NMC and other operational centers have successfully utilized several numerical predictions originating from different centers and/or different initial conditions which encompass the same verification period (Tracton, 1993). However, this approach generally has been subjective, and based on the use of very few different forecasts (e.g., only NMC and ECMWF forecasts are currently available for the 6 to 10 day forecast period). The new ensemble forecasting scheme implemented at NMC provides an ensemble of 14 members. We can envision that with more computer resources, ensembles of hundreds may be possible.

In Section 2 we discuss experiments that justified the decision to trade off model resolution beyond six days for multiple runs. Details of the new operational configuration are discussed in section 3. A major challenge for realizing the potential of ensemble prediction is to condense the vast amounts of information from ensembles into a coherent and "user-friendly" form. Section 4 presents some examples of displays of ensemble products that forecasters have found useful. Finally, even though the present implementation of ensemble forecasting is modest in number, due to the computational constraints, it allows us to start acquiring valuable operational experience, and to perform further research and development of the methodology and applications. A brief discussion of the methods that we have used compared with those implemented at other centers, and a discussion of the possible evolution of NMC systems and further potential applications are presented in Section 5.

2. Model resolution forecast experiments

Because of near saturation of the NMC Cray YMP-832 supercomputer during late 1992, it was not possible to begin ensemble forecasting by simply increasing the number of MRF predictions at its operational T126 resolution. Operational experience has shown that the use of higher horizontal resolution in NWP generally leads to improved forecasts. Indeed, high resolution is probably one of the most important reasons present day forecast guidance is vastly superior to that attainable only a decade ago. One might expect, therefore, that simple truncation of horizontal resolution would lead to some loss of skill. However, it is important to remember that short synoptic waves are intrinsically predictable for periods shorter than longer waves.

Thus, it is reasonable to believe that the largest benefit from very high model resolution, which improves mostly the quality or representation of shorter waves, will be attained primarily at short ranges.

To test this hypothesis a series of experiments were performed in which 39 consecutive 10-day forecasts were generated at the operational T126 resolution (equivalent to 105 km) during the first five days. Beyond day five, the resolution was truncated to T62 (210 km), a resolution at which the model runs about nine times faster. In comparison to the actual operational forecasts, the experimental runs were identical for the first five days, and differed beyond then only in the reduced horizontal resolution. In addition to comparing the operational (MRFS) and truncated (MRFW) models with one another, each was compared to a set of corresponding "pure T62" forecasts (MRFZ), which were generated from a parallel T62 data assimilation cycle and ran at T62 resolution throughout the 10-day prediction. (The parallel T62 cycle runs daily for model test and development purposes.) Note that MRFZ forecasts differ from MRFS and MRFW both due to the differences between operational and parallel data assimilations (i.e., the initial conditions) and to the model resolution used to produce the predictions.

As illustrated by the daily sequence of Northern Hemisphere anomaly correlation (AC) scores in Fig. 2a, the skill of the truncated model (MRFW) for the 6-10 day mean (D+8) 500-mb height field, which is the key product used in the Climate Analysis Center's (CAC) medium-range forecasts (Wagner, 1989), was virtually indistinguishable from that of the operational T126 (MRFS) model. Both were generally better than the pure T62 (MRFZ) model. The rate of divergence between MRFS and MRFW predictions (forecast/forecast AC) beyond day five is small (Fig. 2b) and net differences insignificant in the 6 to 10 day mean. Indeed, the MRFS and MRFW D+8 charts are pattern correlated at greater than .98, whereas the corresponding correlation between MRFZ and MRFS is only .69. Note from Fig. 2b that the rate at which MRFZ diverges from MRFS beginning at day zero is much larger than the divergence of MRFW from MRFS beginning at day five. This suggests that difference between T62 (MRFZ) and T126 (MRFS) D+8 forecasts is due more to differences in the data assimilation (perturbations of the initial conditions) than to the effects of model truncation.

An important question relevant to ensemble prediction is the variability of the forecasts. If the lower resolution had less variability, it would result in a diminished ability to sample properly the range of possible outcomes. To address this question, hemispheric charts were produced (not shown) of the standardized variance in the MRFS, MRFW, and MRFZ 500-mb daily height fields about their respective 6-10 day means. (Standardized here refers to normalizing values by the climatological variability.) The hemispheric means (0.352, 0.349, and 0.353, respectively) are

virtually the same with little difference apparent in the spatial distribution and magnitude of centers. Thus, overall, the lower resolution in either mode does not appear to have a significant impact on the variability of forecasts in the 6 to 10 day range.

In summary, these results indicated that it would be possible to truncate the T126 MRF operational model to T62 beyond five days without jeopardizing the quality of the operational MRF run from 0000 UTC and, therefore, it would be also possible to use the computer savings for ensemble forecasting. Moreover, the results showed that the pure T62 forecasts are only slightly less skillful than the T126/T62 truncated model and possess comparable variability. Hence, the T62 model is a viable choice for generating the ensemble members to complement the operational runs. How this translates to the specifics of the new operational configuration is the subject of the next section.

3. The new operational scheme

Figure 3a shows the configuration of the global forecasts available at NMC before 7 December 1992. Two independent data assimilation cycles were routinely performed, one at T126 resolution (operational system) and the other at T62 (used for model development and parallel tests). T126 and T62 forecasts through 10 days were generated from the respective 0000 UTC initial conditions. Also, the "aviation" (AVN) forecasts were produced at high resolution (T126) from 1200 UTC analyses, but only through 3 days. Thus, even if one includes the experimental parallel cycle, only two global NMC forecasts were available for the 6-10 day forecast period.

Given no additional computer resources, it was obvious that ensemble forecasting with more members was possible only by reducing the resolution of the operational 10-day forecasts. As demonstrated above, it was possible to do so with virtually no loss in skill by truncating the model to T62 after five days of high (T126) resolution. For a reasonably sized ensemble, additional predictions would have to be made with T62 resolution. Section 2 indicates this was a viable proposition, especially with the presumption that the relatively small loss of skill at medium ranges would be offset by the increased information content of the ensembles.

The other major consideration was the strategy for providing initial state perturbations. The Breeding of Growing Modes (BGM) method developed at NMC (Toth and Kalnay, 1993) provided an efficient means for generating perturbations that reflect the fast growing modes present in the analysis cycle. As compared, for example, to random perturbations, perturbations generated by BGM considerably enhance prospects for reasonable sampling of divergent solutions.

With the above ideas in mind we chose the ensemble configuration shown in Fig. 3b, which is "computer resources neutral" when compared to the previous operational system (Fig. 3a). It combines BGM perturbed forecasts with overlapping predictions from time lagging (LAF), where initial differences are the model short range-forecast errors. All predictions are 12 days long, so that with time lagging of up to two days the net result is 14 forecasts available through 10 days every day. The nominal MRF predictions from 0000 UTC are now truncated from T126 to T62 beyond day six, and the 3-day AVN forecasts started from 1200 UTC are extended beyond 3 days with the T62 model. In addition, we use the T62 analysis cycle in the following way: a control forecast is run daily (as before), and the BGM perturbations are added and subtracted from the analysis to provide initial conditions for two additional forecasts.

A disadvantage of LAF is that the perturbations are not similar in size. The "older" forecasts have larger perturbations and, therefore, tend to be less skillful than the later or "younger" forecasts. This problem can be alleviated by weighting the different members as a function of the relative skill of the newer versus older predictions (Dalcher et al, 1988). The key advantage of LAF is that runs which complement those produced from the latest 0000 UTC initial conditions ("today") take advantage of earlier operational runs and don't have to wait for "today's" analysis to begin. Thus, while in theory BGM or alternative methods (e.g., Scaled Lagged Average Forecasting or SLAF, Ebisuzaki and Kalnay, 1991) could be used to generate additional sets of perturbed analyses valid "today", it is computationally unfeasible now to substitute them for the older time-lagged predictions.

For the present, the ensemble forecasts are only available for use and evaluation within NMC[2]. They are considered routinely in the CAC's thrice weekly medium-range 6-10 day predictions. Although this application was the primary motivation for implementing ensemble prediction, the potential use in preparing the 3-, 4-, and 5- day products of NMC's Meteorological Operations Division (MOD) is also being evaluated. Finally, it should be pointed out that twice per month the ensemble members are extended to encompass forecast periods through 15 days for use as guidance in the preparation of CAC's operational monthly outlooks.

Examples of output from ensembles relevant to operational forecast applications are presented in the next section. Most of these products were developed in conjunction with an earlier NMC experiment in ensemble prediction. In that experiment, the ensembles consisted of nine members produced solely by time lagging with six hour spacing of forecasts over a two-day period. Although

---

[2]NMC plans in the near future to include these forecasts within a file server accessible by interested users.

some results of that study are presented in Section 4, we expect to document it more completely in the near future. Suffice it to say that the results of the experiment were encouraging in demonstrating the utility of the products described below, especially with regard to the potential reliability of probability estimates.

4. Operational forecast applications

A major challenge of ensemble prediction is to condense large amounts of information into a coherent "user-friendly" form. The goal is to provide forecasters with succinct and operationally meaningful forecast output which displays the essential aspects of the range and likelihood of alternative scenarios. The output products can range from display of each individual prediction and measure of ensemble "spread", through ensemble means and "clustering" of similar members, to explicit estimates of the probability of some event or distribution of a given parameter. This output could be applied to virtually any product derived from the model forecast. Examples are the 500-mb height and 1000/500-mb thickness anomalies, occurrence of major cyclogenesis and the envelope of possible storm tracks, measures of blocking versus zonal flow regimes, the occurrence of precipitation, and the position of the 540 dam "rain/snow" line.

Illustrations of some products are presented below with the understanding that there are many additional and/or alternative outputs, and that continued interaction among forecasters within and outside NMC is essential for optimizing their development and utility. Also, we acknowledge that much of what is presented has roots in the seminal paper by Epstein (1971), who showed several graphical results of stochastic predictions to demonstrate how information on uncertainty can be depicted and to illustrate the value of specific information on uncertainty. In addition these products also reflect the recommendations of the ECMWF Workshop on New Developments in Predictability (1992) and the ECMWF Third Workshop on Meteorological Operational Systems (1992).

a. Individual forecasts, ensemble mean and "spread"

Figure 4 presents the individual D+8 (6-10 day mean) 500-mb height/anomaly charts from 0000 UTC 3 February 1993 ("today"). They are identified as 1 through 14 in accord with the nomenclature of Fig. 3b. The set of charts is clearly cumbersome but provides forecasters with the complete picture for reference when considering condensed output. The first of the condensed output is the ensemble mean prediction (Fig. 5), which here is simply the arithmetic average of the 14 predictions. A weighted averaging scheme with weights that reflect the relative skill of individual members is under development (van den Dool and Rukhovets, 1993). Ideally, i.e., with a perfect model and a large set of initial perturbations that adequately sample uncertainties in the analysis,

the ensemble mean should provide on average the best single forecast by filtering the unpredictable components of the individual predictions[3]. In the real world of imperfect models and limited samples, the filtering effect of ensemble averaging has proven generally to be rather small (e.g., Murphy, 1990, Brankovic et al, 1990, Tracton et al, 1989). Toth and Kalnay (1993) report that ensemble means, constructed simply as the average of two BGM perturbed forecasts, scored 2% higher than the control (T62) forecasts in the day five 500-mb AC over the Northern Hemisphere, and 3% over the Southern Hemisphere. Preliminary verifications show that the skill of the complete 14- member (unweighted) ensemble mean D+8 predictions, as was also true in the 9-member ensemble experiment referred to above, is just about the same as "today's" (i.e., the latest) MRF forecast. However, giving most weight to the five forecasts made from the two latest analysis times (i.e., the 12-hour old AVN forecast, the latest MRF forecast, and the control and 2 BGM forecasts at lower resolution) increases the skill somewhat over the unweighted mean.

Whether or not weigthing significantly improves the skill of the ensemble mean prediction, we expect that the larger gain from ensemble forecasting will come from information on uncertainties. The first example of output addressing forecast uncertainty is the ensemble "spread" (Fig. 6), which is expressed here as the standard deviation of the D+8 500-mb height predictions about the ensemble mean normalized by the climatological standard deviation. The spread chart displays the geographical distribution of the variability within the ensemble and, therefore, provides a regional measure of the forecast uncertainty. The operative hypothesis is that the larger the spread, the wider the range of possibilities and, therefore, the smaller the level of confidence in any particular outcome. The shading indicates regions where the spread is approaching or exceeds 1.0, i.e., where the degree of uncertainty is comparable to the climatological variability. These are areas where there is a low signal-to-noise ratio and, therefore, little or no confidence in the forecasts. For example, with reference to the individual predictions shown in Fig. 4, the small ensemble spread in Fig. 6 implies relatively high confidence in the predictions of the largely zonal circulation over the US. By contrast, large spread indicates essentially no confidence in forecasts of the cyclonic circulation south of Kamchatka and of the potential for blocking just to the north.

It should be noted that that there is an element of subjectivity and user dependence when appraising the significance

---

[3] Even in this case, the ensemble forecast is only one outcome of a probability distribution. On average, however, it should be more skillful than any other arbitrarily selected member of the ensemble.

of differences among ensemble members. What constitutes a significant difference between forecasts for some purposes, may be irrelevant for other applications. For example, if Fig. 6 were the 1000/500-mb thickness rather than 500-mb height, the equivalent of 30m in spread might be considered small in general, but could be crucial to the level of confidence in rain-versus-snow situations. Considerations like these, of course, apply not just to the ensemble spread, but to most other alternative products representing divergence among the individual ensemble predictions.

The relationship between spread and confidence just described is often referred to as a relationship between spread and skill; however, because a priori any given forecast is only one arbitrary member of a probability distribution, correlations between spread and forecast skill are generally rather small (van den Dool, 1992). That is, spread reflects more the overall degree of uncertainty than the skill of any particular forecast, which by chance may be a skillful or unskillful member of the ensemble.

b. Objective clustering

The next level of output available to forecasters moves beyond an overall measure of confidence, the ensemble spread, to presentations that specifically address the range and likelihood of the differing scenarios encompassed by the ensembles. The first of these involve clustering, i.e., objectively grouping together ensemble members which are similar in some respect. In this preliminary effort, we use a simple and somewhat heuristic clustering algorithm: Based on 500-mb hemispheric anomaly correlations (AC), we find the two predictions that are least similar (smallest AC among all possible pairs). This pair defines the range of forecast possibilities. Next, we identify the ensemble members that are "similar" (defined as having an AC >0.60) to each of these extremes to form the first two clusters. Of the remaining forecasts, the two most similar are found, and members similar to them grouped to form the next cluster. The process continues until there is no longer any set of at least two forecasts that are similar. The clustering algorithm precludes any forecast from belonging to more than one cluster. The final step is to obtain the mean of the forecasts comprising each cluster.

Figure 7 presents the results of applying this procedure to the set of charts shown in Fig. 4. The individual members comprising each cluster mean are indicated by a "1" (yes) or "0" (no) in the appropriate position of the 14 slots along the top of each chart. The first two clusters are the means formed around the two predictions identified as least similar. Here, Cluster 1 contains seven forecasts, while Cluster 2 consists of only a single prediction (i.e., there is no forecast similar to it). Two additional clusters (Clusters 3 and 4), which consist of two forecasts each, are formed from the remaining ensemble members.

Ideally, the number of forecasts in each cluster should reflect the relative likelihood of which will verify most closely to actual events. From this perspective, Cluster 1 suggests the most probable scenario. For example, the broader scale trough associated with the low south of Greenland is more likely to extend southward over the eastern U.S. than southeastward over the mid Atlantic, as in Cluster 2. The model indicates the latter scenario, though, is possible and cannot be ruled out. Of course, whether this sort of reasoning provides consistently reliable estimates of possible outcomes must await a long period of operational experience with ensemble prediction. However, the results from the 9-member ensemble experiment were encouraging, and this is reflected in the objective verification of probabilities shown below.

It is evident that there is considerable latitude in clustering with regard to the field or quantity compared, the measure used to judge similarity, and criteria for binning. Similar procedures could be applied, for example, to 1000 mb height (or mean-sea-level-pressure, MSLP) with groupings based upon root-mean-squared differences over limited regions. Fig. 8 displays the dominant cluster of the 1000 mb height field predictions over the eastern U.S. from 10 March 1993 ("today") and verifying 14 March 1993. This cluster, which consists of a clear majority (10) of forecasts, including two-day old MRF and T62 runs (6-day forecasts), points toward the major east coast storm that since has become known as the "Blizzard of 93". At the other extreme (not shown) a single perturbation run showed little development and, thus, introduced some element of uncertainty.

Other examples of the clustering procedure (not shown) are to bin forecasts using some measure of circulation regime, such as the Pacific North American (PNA) index (Wallace and Gutzler, 1981) or a measure of blocking activity (e.g., Lejenas and Okland, 1983). Also, note that clustering can be done with respect to similarities in the temporal evolution of features and/or quantities, as well as to spatial comparisons.

c. Graphical clustering

Figures 9 and 10 illustrate additional charts which condense the information of Fig. 4 and provide quick visualization of the possible solutions. Fig. 9 displays on the same base map the location and magnitude of anomaly centers from all 14 predictions. Inspection immediately reveals the consistency or lack of it in the forecast placement and intensity of centers. One can view this as a type of clustering in that groupings of anomaly centers suggest by their numbers the relative degree of confidence in prediction of these centers. Thus, for example, the ensemble clearly suggests a greater chance for a positive anomaly over northern Europe than over the Balkans. Note, however, that verification of one does not preclude the other (i.e., both centers could verify). This differs

from the above where, for example, clusters that indicate zonal versus blocking flow over the same region (at the same time) are mutually exclusive.

Figure 10 displays a composite chart of the 564 dam contour from all 14 forecasts. Similar maps with contours that reflect higher and lower latitude circulations are available to the forecaster (not shown). Visual inspection of Fig. 10 reveals that the degree of consistency among ensemble members ranges from virtual unanimity in the circulation over the western Pacific to a wide envelope of possibilities over Europe. Even there, however, a clear majority of forecasts favor the northerly zonal flow as compared to the cutoff circulations further south suggested by some predictions.

Outputs like those in Figs. 9 and 10 can be applied readily to a sequence of daily charts and, thereby, display the evolution of systems and growth of differences among forecasts. By way of example Figs. 11 and 12 present the day 3, 4, 5, and 6 1000-mb height centers (equivalent to MSLP centers) and 558 dam 500-mb contours for the 3 February case. Note that many forecasts suggest development of a low center over the northern Gulf of Mexico and subsequent movement toward the northeast. These developments occur in response to the evolving 500-mb short wave over the southeastern U.S. There are clearly differing versions of these events (more so than in the 10 March case) as a function of the particular ensemble member. The significance of the differences depends upon the specific requirements of the user. Collectively, the ensemble predictions suggest a strong likelihood of a major weather event affecting the east coast of the U.S. There is much less agreement and, therefore, confidence in prediction of details in the timing and amplitude of the event. Of course, this just reflects the general limited predictability of smaller-scale features beyond a few days in advance. However, while this is true in general, it is not always so, and consistency in the details among ensemble members can indicate the exceptions (i.e., indicate a signal in the noise of generally unpredictable components).

Ideally, considerations like those discussed in this section should provide reliable estimates of uncertainties. As noted in regard to the objective clustering, however, whether they do so consistently in reality must await evaluation over an extended period of operational experience.

d. Probability forecasts

The products described above lend themselves primarily to qualitative statements about the relative likelihood of different possible outcomes. Ensemble forecasting can also provide the basis for quantitative estimates of probabilities. The first example (Fig. 13) displays the probability over the Northern

Hemisphere of the D+8 500-mb height anomaly exceeding 0.5 times the climatological standard deviation. Probability estimates here are defined simply as the percentage of predictions out of the total 14-member ensemble that satisfy this (somewhat arbitrary) criterion. Cross-hatched and hatched areas indicate probabilities of greater than 50% for positive and negative anomalies, respectively. Unshaded regions are those where there is no majority among ensemble members in the sign and specified magnitude of the predicted anomaly. For example in Figure 13 there is near unanimity in the outlook for large positive anomalies over the Northwest Territories, greater than a 70% chance for negative anomalies over Baja, and no consensus in the predictions between these areas. To the extent that sensible weather elements are linked to the height anomaly field, as via Klein specification equations (Klein, 1985) in CAC's medium-range forecast operations (Wagner, 1989), the probabilities convey the degree of confidence in those elements. Also, it is often possible to infer the height anomaly patterns in uncertain regions via teleconnections (O'Connor, 1969; Namias and Clapp, 1981) from the high confidence areas.

Probabilities like those above, which reflect the spread among ensemble members (e.g., Fig. 6), can be obtained for individual days and/or five-day means directly for any model field by simply counting the number of members in the ensemble falling within specified categories. Examples include the chances for temperature (or layer thickness) anomalies to be above or below given threshold values (see Fig. 14, for the March 1993 case) and the odds for precipitation to exceed some specified amount (not shown). Output can be presented in map form, such as in Figs. 13 and 14, or for specific point locations. Finally, the change in probabilities over time suggests trends in levels of confidence and degree of predictability. For a given ensemble, probabilities will generally decrease over the forecast period, and for some specified minimum level of confidence (e.g., climatological expectation) one can assess the usable length of the forecast. Conversely, for a fixed verifying date, the probabilities will generally increase as the initial time ("today") gets closer and the ensemble members tend to agree more with one another. For some lower threshold of uncertainty one can decide when it's reasonably safe to become a believer.

Unlike many products derived from ensembles, quantitative evaluation of probability forecasts is relatively straightforward via Brier (1950) and/or ranked probability scores (Epstein, 1969). However, our operational experience to date is not long enough to provide meaningful verifications. Therefore, for purposes of illustration, Fig. 15 presents from the 9-member ensemble experiment a "reliability" diagram of a 34-case average observed frequency versus forecast probability of D+8 500-mb height anomalies. The diagram reflects a major component of the Brier skill score (Murphy,1985) with the ideal being values lying along

the forecast equals observed diagonal. The results show that, even for this limited 9-member, LAF (no "breeding") ensemble strategy, the forecast probabilities provide fairly reliable estimates of the actual frequency of significant (>0.5 std. deviation) height anomalies. The correlation between predicted and observed frequencies, 0.77 and 0.73 for above and below normal categories, respectively, indicates a fairly linear relationship. However, the slope is less steep than the diagonal in both categories, which reflects a bias toward overconfidence in the predictions. Forecast probabilities are greater than observed frequencies for likely occurrences and less than observed when unlikely. To the extent biases of this sort are systematic, the probability estimates can be calibrated to enhance their credibility.

5. Discussion

With the operational implementation of ensemble prediction, NMC explicitly recognizes that forecasts are inherently stochastic, i.e., probabilistic in nature. There is no unique solution, only an array of possibilities, which ensemble forecasting attempts to sample. The objective is to provide reliable estimates of the range and likelihood of those possibilities. To the extent this objective is achieved, the utility of NWP is expected to increase by providing users information necessary to weigh uncertainties in making decisions. This represents a major shift in the primary goal of NWP research, with the focus now toward maximizing the utility of forecasts and not just the average level of model skill.

From the discussion in the previous sections it is clear that there are several major questions related to ensemble prediction that need to be addressed in an operational center. The first is the extent to which computer resources should be invested in higher-model resolution and more advanced physical parameterizations versus satisfying the need for multiple runs. Certainly, it would be desirable to use the most accurate model with the least systematic error in order to explore alternative forecast scenarios, given uncertainties in initial conditions. Thus, for example, a model that performs better in predicting the depth of cyclones and frequency of blocking is less apt to bias ensemble members toward unlikely outcomes in sampling the intensity of storms or blocking activity. The question is what level of accuracy is sufficient given that model improvements and multiple runs inevitably compete for computer resources.

Fortunately, in the implementation described here, we found a compromise that provided the necessary computer resources for ensemble forecasting without sacrificing the quality of the original operational predictions. The nominal 0000 UTC MRF run is truncated beyond day six from T126 to T62 resolution, based on experiments that showed no adverse impact upon forecasts. This is because the benefit of the higher resolution is obtained in the analysis cycle and in the first few days of the forecast, when the

shorter scales are still predictable. The computer savings are used to produce the complementary runs with T62 model resolution beginning at day zero or, in the case of the AVN extension, day 3 (Figs. 3a and 3b). Although these T62 runs are individually slightly less skillful on the average, they provide the potential for increased information from additional forecasts. Similar compromises may be required in the future.

A second major question is the choice of methodology for generating the perturbed set of initial conditions. As discussed by Toth and Kalnay (1993), the BGM approach has shown considerable promise for representing the fast growing errors present in the analysis and, therefore, for providing a representative sampling of forecast scenarios. The BGM scheme, the SLAF scheme, and the singular modes method implemented at ECMWF (Palmer et al, 1992, Lorenz, 1965) are clearly better than simple "Monte Carlo" forecasting, which cannot represent well the initial growing errors. All three methods have the advantage over time lagging (LAF) that, in principle, a large number of realistic initial perturbations of similar size can be created and used for ensemble forecasting. In practical terms, however, all of these methods require waiting until "today's" analysis is completed before beginning the multiple runs. LAF has the advantage that some of the complementary forecasts are created during the previous days' operational cycle, and therefore distribute the computer load more efficiently in time (Fig. 3b). Hence, coupling a limited number of BGM runs with time lagging was the only operationally feasible approach to produce a reasonably sized (14 member) sampling of initial conditions. Whether this ensemble size and configuration is adequate for the purpose of obtaining realistic estimates of possible outcomes remains to be determined.

A third major question addressed in this paper is how to condense the enormous amounts of information from forecast ensembles into displays that can be "digested" and interpreted easily by forecasters, and, therefore, are useful in an operational setting. We have provided several examples of such "user-friendly" displays, but they certainly do not exhaust all possibilities, nor are they necessarily the most effective ways to express the information available from ensembles. The products derived from ensembles obviously should be tailored suit the particular application. The key challenge is to transform what, given a single forecast, can be expressed only as categorical statements into meteorologically more useful estimates of the chances for realizing several possible scenarios.

Overall, the present NMC implementation of ensemble prediction is relatively modest: 14 member ensembles through the 10-day forecasts. However, it does provide the basis for the development of necessary operational experience with ensemble forecasting and for research directed toward maximizing the utility of NMC's numerical guidance. A key element for success in ensemble

prediction is the sustained interaction between NMC and outside users. To this end, NMC plans to make the raw ensemble data and some of its products routinely available as soon as communications permit. Evaluation of the present implementation, including feedback from outside users, will form the basis for development of future systems (after the acquisition of a Class 7 supercomputer), and for increased use of graphical displays in interactive workstations. We expect to document results of this evaluation and provide information on changes in the operationsal configuration in a timely fashion.

Finally, while the implementation described here was directed toward medium-range forecasting, it should be clear that the fundamental concepts apply equally well to short-range forecasting. Although the divergence among short-range predictions resulting from uncertainties in the initial state is smaller than at longer lead times, for some parameters (e.g., quantitative precipitation, stability indices), even small differences are operationally significant. In other words, there is an stochastic element even in the short-range numerical forecasting. We expect that in the future, with expanded computer resources, it will be possible to perform short-range ensemble forecasting by running a regional, high resolution model (either the ETA model, Mesinger et al, 1992, or the Regional Spectral Model, Juang and Kanamitsu, 1993) within each of the members of the ensemble of global forecasts.

Acknowledgements

We are most grateful to M. Iredell, W. Ebisuzaki, and R. Schechter for their essential help with the computational aspects of this work. Special thanks are expressed to Joseph Irwin of NMC's Automation Division for his many outstanding contributions to the operational implementation of ensemble forecasting. We profited considerably from many stimulating discussions with Edward Epstein, Huug van den Dool, Zoltan Toth and Dave Baumhefner. Reviews of the manuscript by Pete Caplan and Louis Uccellini were very helpful. The encouragement and support of Ron McPherson are also much appreciated.

References

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, 78, 1-3.

Brankovic, C., T. N. Palmer, F. Molteni, and U. Cubasch, 1990: Extended-range predictions with the ECMWF models: time lagged ensemble forecasting. *Quart. J. Roy. Meteorol. Soc.*, 116, 867-912.

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

ECMWF Workshop on New Developments in Predictability, 13-15 November 1991, 333 pp.

Ebisuzaki, W. and E. Kalnay, 1991: Ensemble experiments with a new Lagged Analysis Forecasting scheme. Research Activities in Atmospheric and Oceanic Modelling, Report No. 15, WMO.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, 8, 985-987.

Epstein, E. S., 1971: Depicting Stochastic Dynamic Forecasts. *J. Appl. Meteor.*, 4, 500-511.

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

Juang, H-M. and M. Kanamitsu, 1993: The NMC Regional Spectral. Submitted to *Mon. Wea. Rev.*

Kalnay, E. and Z. Toth, 1991: Estimation of the growing modes from short range forecast errors. Research Highlights of the NMC Development Division: 190-1991, pp. 160-15. Available from NMC Washington, D. C. 20233.

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

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.

Lorenz, E. N., 1963: Deterministic non-periodic flow. *J. Atmos. Sci.*, 20, 130-141.

Lorenz, E. N., 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus*, 17, 321-333.

Mesinger, F., T. L. Black, D. W. Plummer, and J. H. Ward, 1990: ETA model precipitation model forecasts for a period including tropical storm Allison. *Wea. and Forecasting*, 3, 483-493.

Murphy, A. H., 1985: Probabilistic weather forecasting. Probability, Statistics, and Decision Making in the Atmospheric Sciences, edited by A. Murphy and R. Katz, Westview Press, 337-377.

Murphy, J. M., 1990: Assessment of the practical utility of extended-range ensemble forecasts. *Quart. J. Roy. Meteorol. Soc.*, 116, 89-125.

Molteni, F. and T. N. Palmer, 1993: Predictability and finite-time instability of Northern Winter Circulation. Submitted to *Quart. J. Roy. Met. Soc.*

Namias, J. and P. F. Clapp, 1944: Teleconnections of 700 mb height anomalies for the Northern Hemisphere. *CALCOFI Atlas No.* 29, Scripps Institute of Oceanography, 265 pp.

O'Connor, J. F., 1969: Hemispheric teleconnections of mean circulation. Environmental Science Service Association Tech. Report WB 10, Silver Spring, MD, iii + 103 pp.

Palmer, T., F. Molteni, R. Mureau, R. Buizza, P. Chapelet, J. Tribbia, 1992: Ensemble Prediction, Research Department Technical Memorandum No. 188, 45 pp. (available from The Director, European Center for Medium Range Weather Forecasts, Shinfield Park, Reading, Berkshire, RG29AX)

Toth, Z., and E. Kalnay 1993: Ensemble forecasting at NMC: The generation of Perturbations. Submitted to *Bul. Amer. Met. Soc.*
Tracton, M. S., 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, 2230-2247.

Tracton, M. S., 1993: On the skill and utility of NMC's medium-range central guidance. *Wea. Forecasting,* 8, 147-153.

Van den Dool, H., M., 1992: Forecasting forecast skill, probability forecasting, and the plausibility of model produced flow. ECMWF Workshop on New Developments in Predictability, 13-15 November 1991, 87-92. (available from The Director, European Center for Medium Range Weather Forecasts, Shinfield Park, Reading, Berkshire, RG29AX)

Van den Dool, H, and L. Rukhovets, 1993: On the weights for a 14 member ensemble, 6-10 day forecast average. To be published as an NMC Office Note. Details will be provided wih "proof")

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

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

Figure Captions

Fig. 1: Schematic of ensemble prediction, with individual trajectories drawn for forecasts starting from a representative set of perturbed initial conditions within a circle representing the uncertainty in the initial conditions, and ending within the range of possible solutions, represented by the ellipse. For the shorter range, the forecasts are close to each other, and they may be considered deterministic, but beyond a certain time, the forecast is stochastic. Forecasts may cluster into groups of similar trajectories (denoted A and B in the figure), whose probability of verification may be related to the number of forecasts in each group.

Fig. 2: (a) Sequence of Northern Hemisphere AC scores for the operational T126 (MRFS), truncated T126/T62 (MRFW), and pure T62 (MRFZ) D+8 (6-10 day mean) 500-mb height fields (solid, dashed, dotted, respectively). Case 1 to 39 are from initial conditions for 18 February to 27 March 1992. **(b)** Decay curve for the 39 case average Northern Hemisphere agreement (AC) between MRFW (dashed) and MRFZ (dotted) with MRFS (solid).

Fig. 3: Operational configuration of global predictions before 7 December 1992 (a), and the new ensemble configuration (b). In (b), individual ensemble members are identified by numbers 1 to 14.

Fig. 4: Individual ensemble predictions of the D+8 500-mb height (solid) and anomaly fields (dashed) for 3 February 1993. Negative anomalies are shaded, and units are dam and m for the height and anomaly fields, respectively. Numbers in upper right identify individual members per the convention shown in Fig. 3.

Fig. 5: Ensemble mean D+8 500-mb height and anomaly forecast for 3 February 1993 with shading and units as in Fig. 4.

Fig. 6: Ensemble spread, defined as the standard deviation of the 14 D+8 500-mb height predictions (Fig. 4) about the ensemble mean (Fig. 5), normalized by the climatological standard deviation; shaded areas correspond to spread greater than 0.6.

Fig. 7: Cluster mean D+8 500-mb predictions (see text for details) with the individual forecasts from Fig. 4 that comprise each cluster identified by a 1 (yes) or 0 (no) in the 14 positions from left to right across the top.

Fig. 8: Cluster 1 1000-mb height forecast (with 1000/500-mb thickness superimposed) from 10 March 1990 ("today") and verifying 14 March 1993 ("blizzard of 93 case"). Units are m and dam for the 1000 mb height and 1000/500-mb thickness, respectively.

Fig. 9: Fourteen member composite of 3 February 1993 positive (+) and negative (-) D+8 500-mb anomaly centers; amplitude of centers

in m.

Fig. 10: Fourteen member composite of the 564 dam D+8 500-mb contour from 3 February 1993.

Fig. 11: Fourteen member composite of 1000 mb high (H) and low (L) centers for days 3, 4, 5, and 6 from 3 February 1993; amplitude of centers m.

Fig. 12: Fourteen member composite of the 558 dam contour for days 3, 4, 5, and 6 from 3 February 1993.

Fig. 13: Ensemble forecast probabilities exceeding 50% for the D+8 500-mb standardized height anomalies from 3 February 1993 greater than +0.5 (cross hatched) and less than -0.5 (hatched).

Fig. 14: Forecast probability from 10 March 1993 for the 1000/500-mb thickness to exceed 540 dam on 14 March 1993. Operationally, the 50% line is viewed generally as the line of equal chance for rain versus snow.

Fig. 15: Forecast probability versus observed frequency of D+8 standardized 500-mb height anomalies greater (a) and less (b) than .5 in magnitude. Results are the 34- case mean for 9 member time-lagged ensembles over the period February 1991 through August 1992.

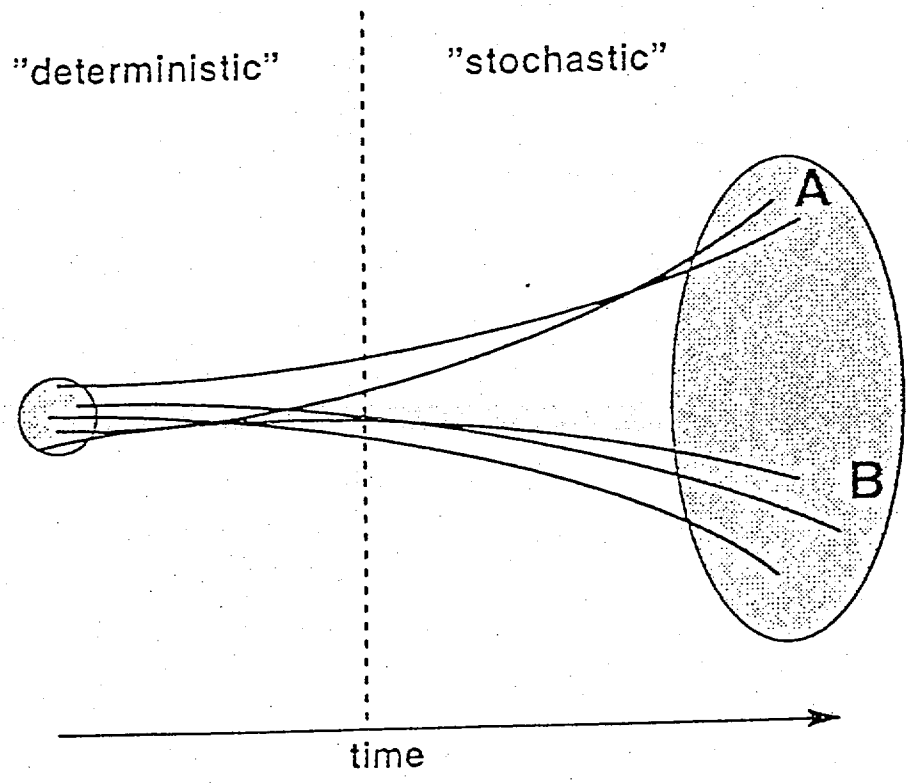


FIG 1

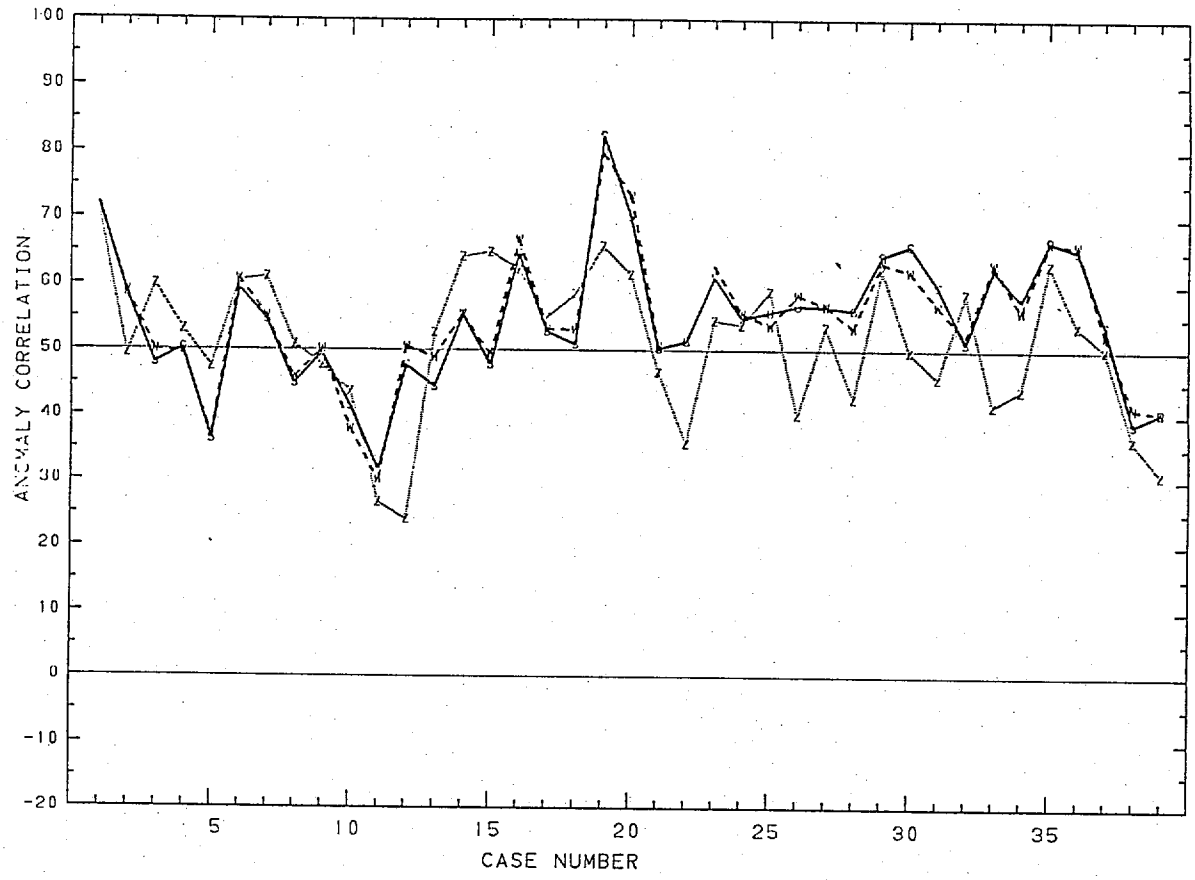

ANOMALY CORRELATION
100
90
80
70
60
50
40
30
20
10
0
-10
-20
5
10
15
20
25
30
35
CASE NUMBER

FIG 2a

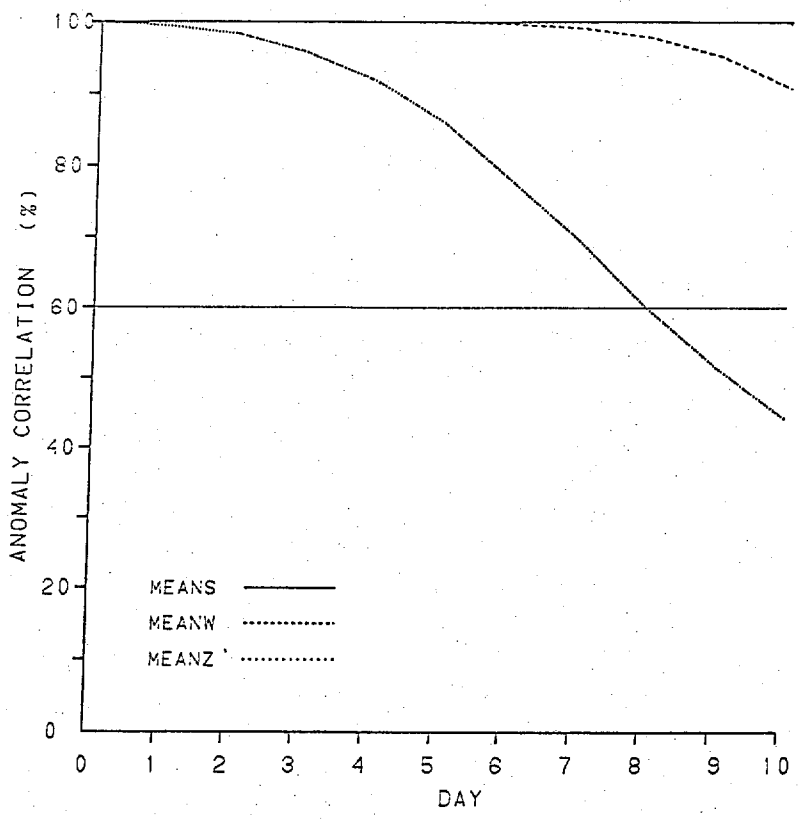

ANOMALY CORRELATION (%)
100
80
60
40
20
0
MEANS
MEANW
MEANZ
0
1
2
3
4
5
6
7
8
9
10
DAY

FIG 2b

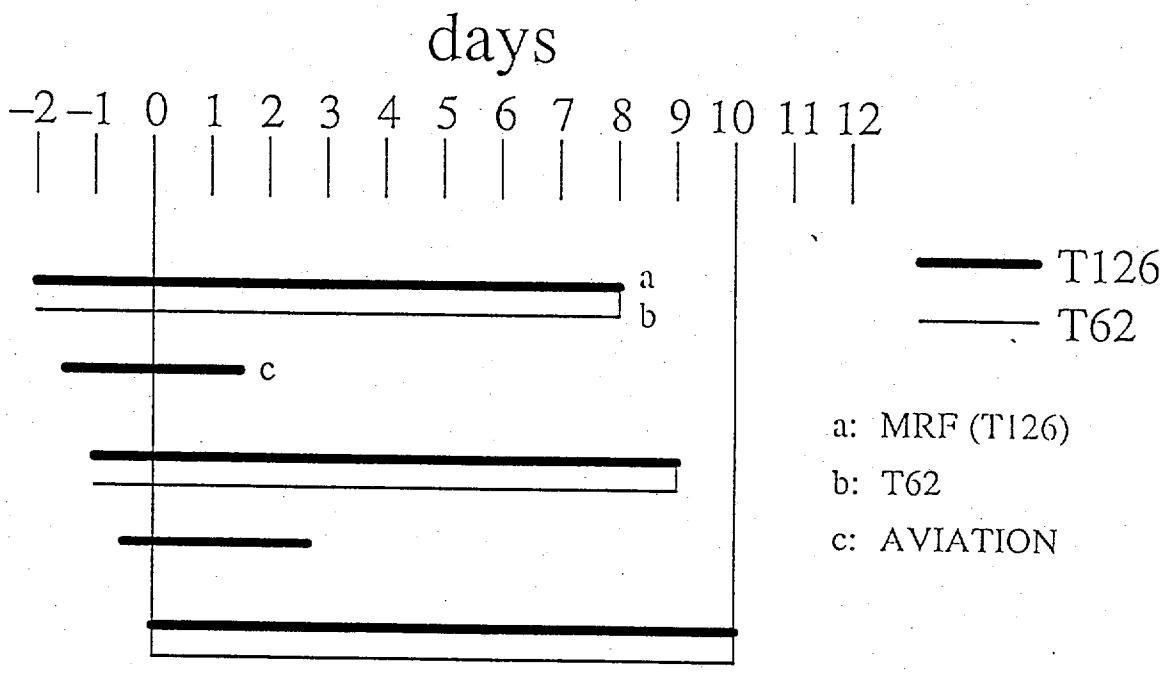


Fig. 3a

Previous Operational Configuration

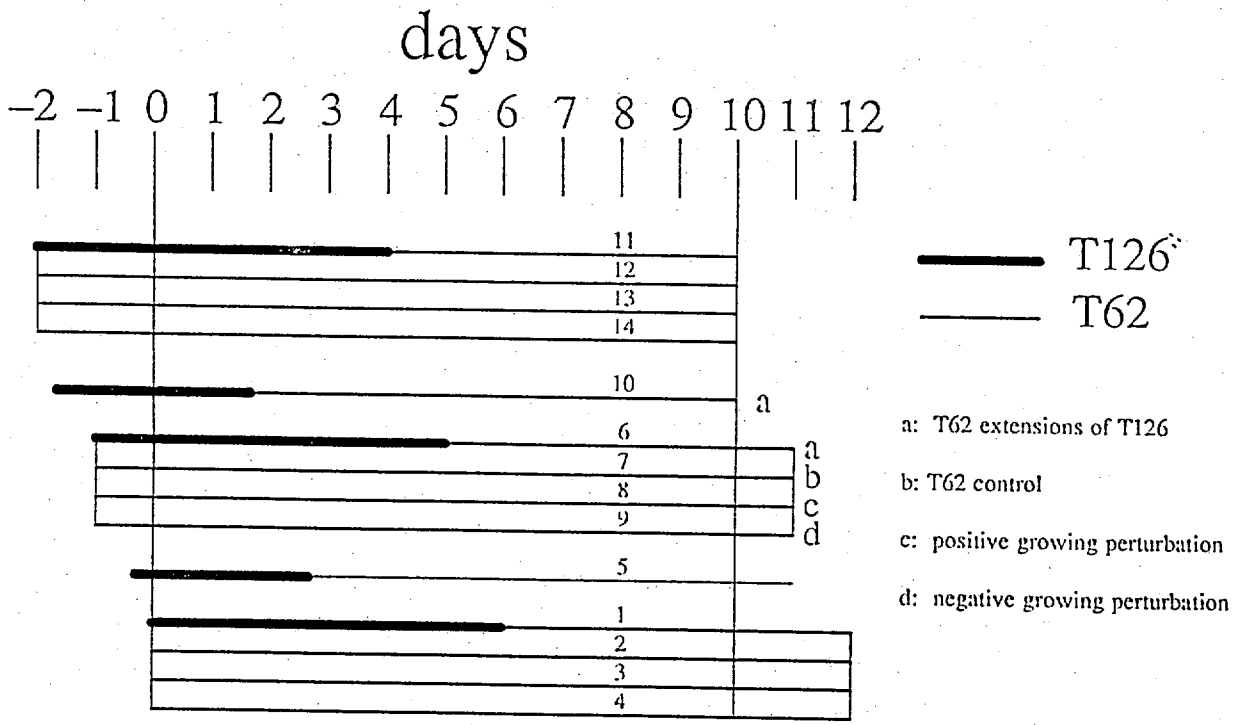
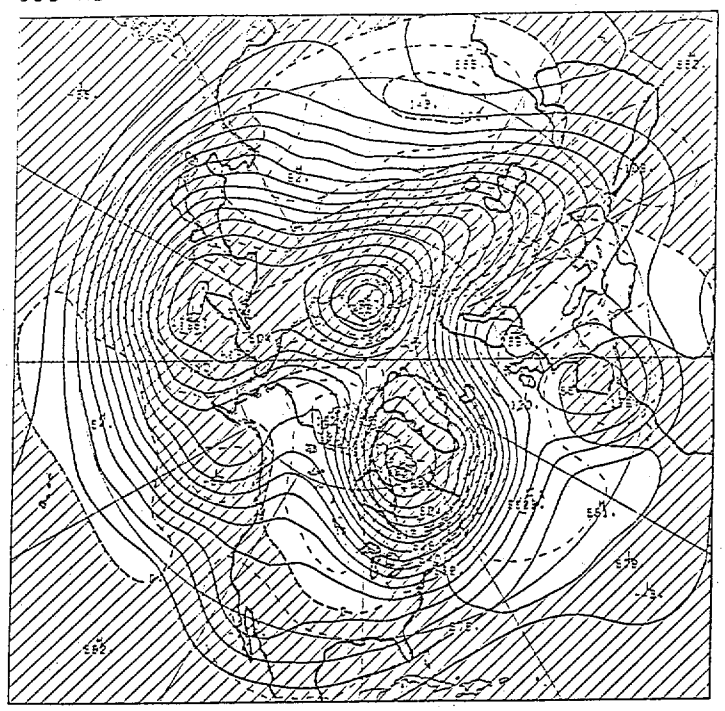
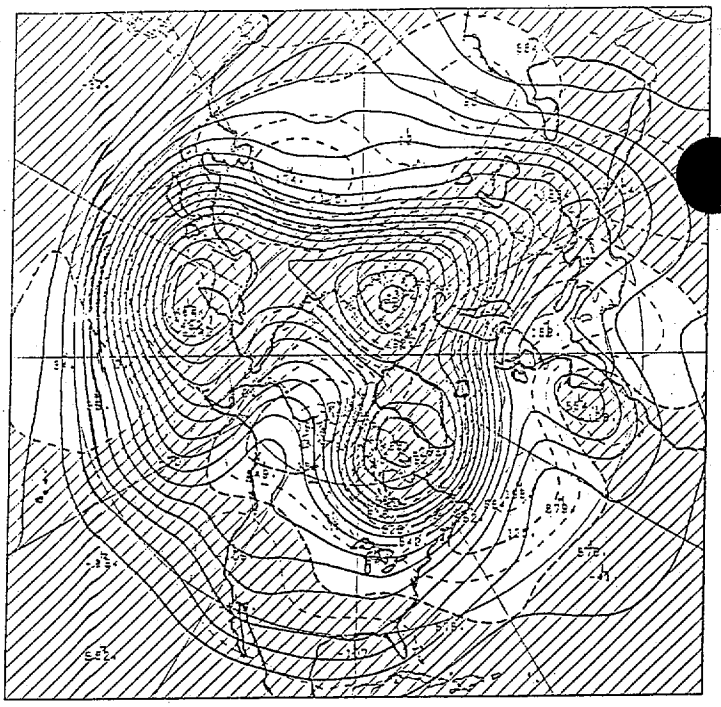
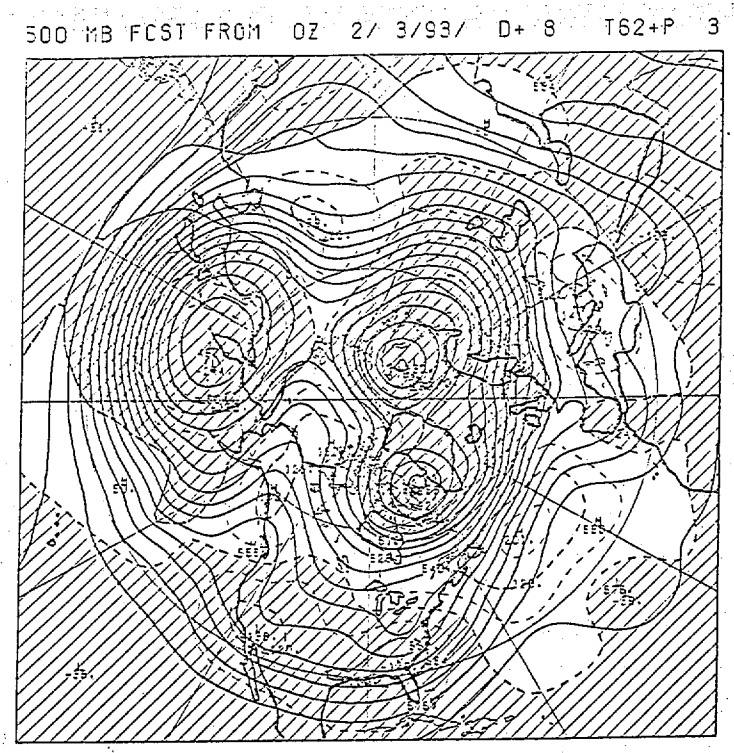


Fig. 3b

New Ensemble Configuration


500 MB FCST FROM 0Z 2/ 3/93/ D+ 8 MRF 1


500 MB FCST FROM 0Z 2/ 3/93/ D+ 8 T62 C 2


500 MB FCST FROM 0Z 2/ 3/93/ D+ 8 T62+P 3

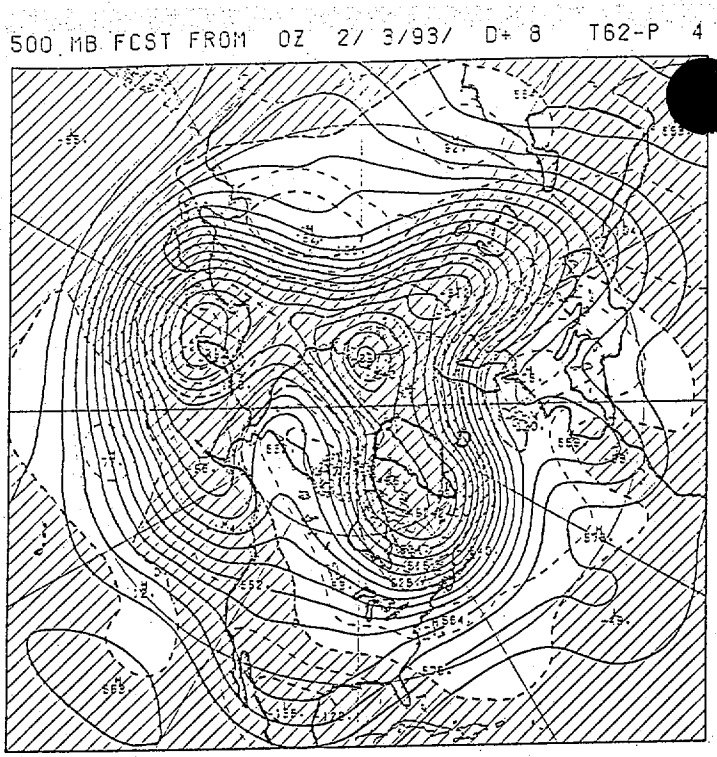
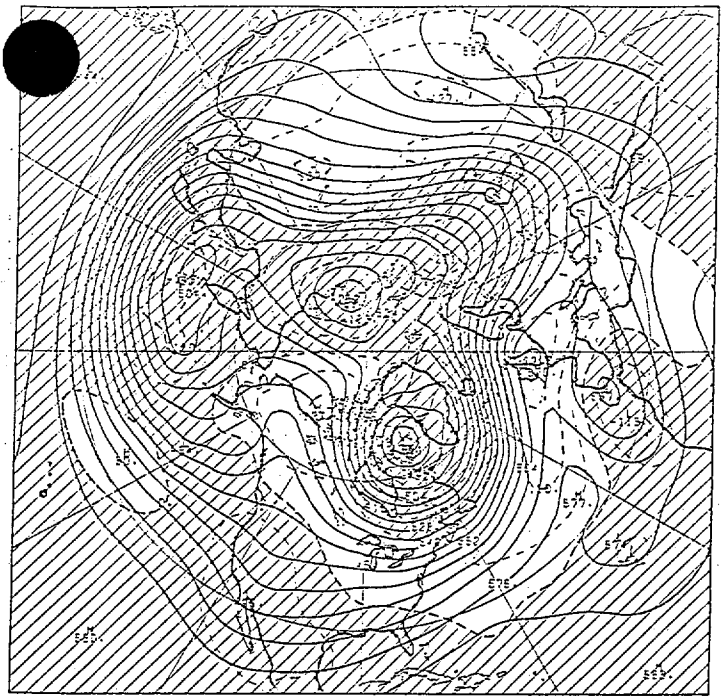
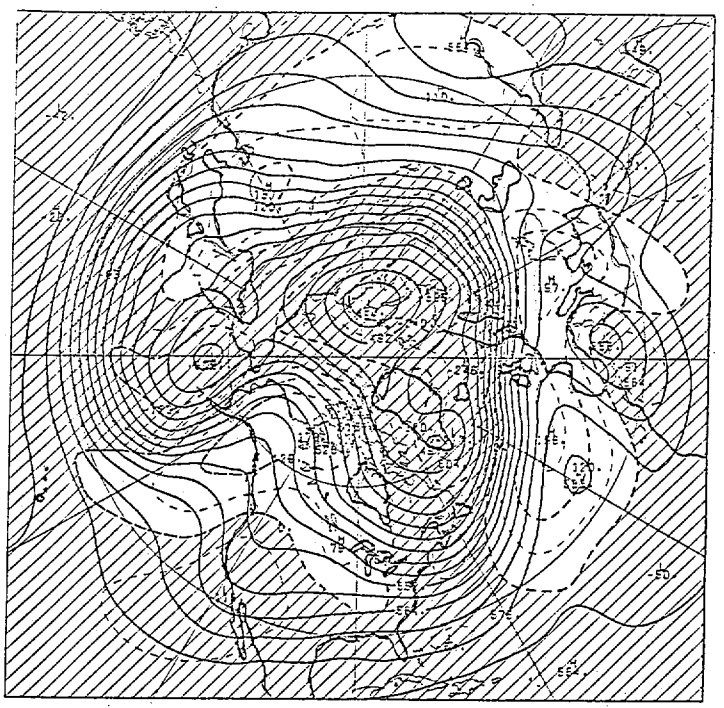

500 MB FCST FROM 0Z 2/ 3/93/ D+ 8 T62-P 4

FIG. 4

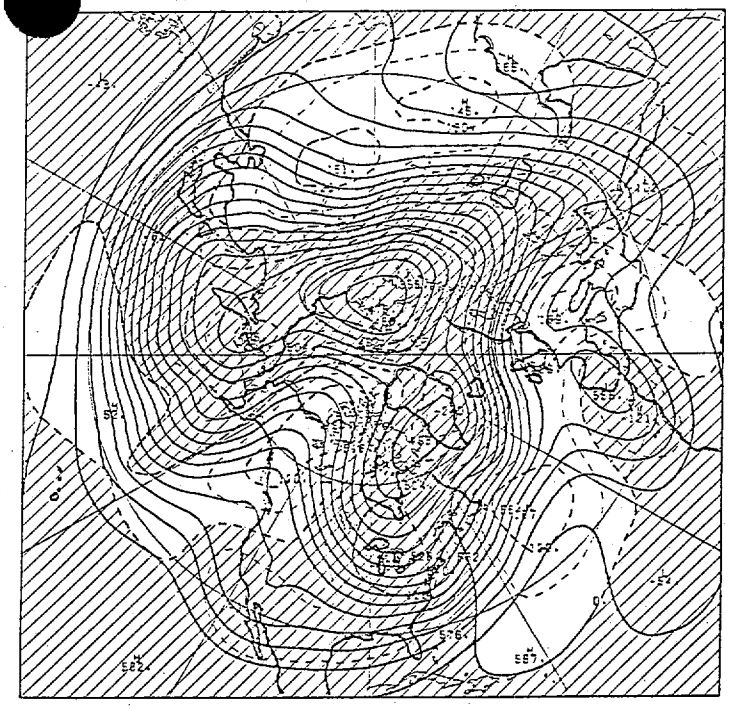
500 MB FCST FROM 12Z 2/ 2/93/ D+ 8 AVN 5



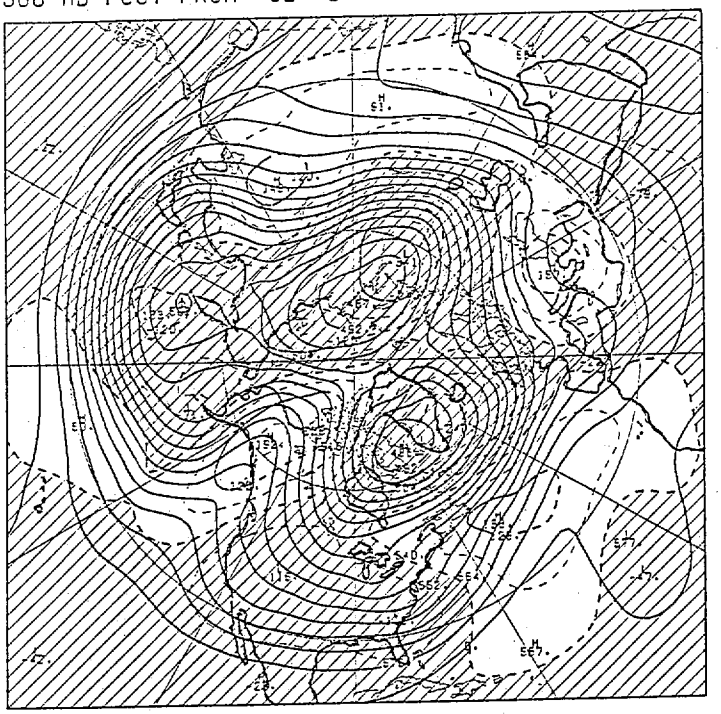
500 MB FCST FROM 0Z 2/ 2/93/ D+ 8 MRF 6



MB FCST FROM 0Z 2/ 2/93/ D+ 8 T62 C 7

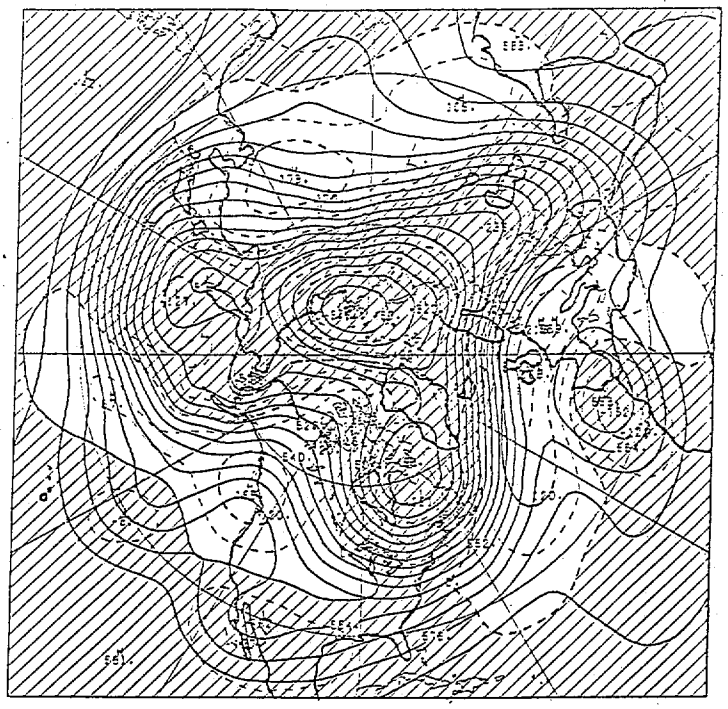


500 MB FCST FROM 0Z 2/ 2/93/ D+ 8 T62+P 8

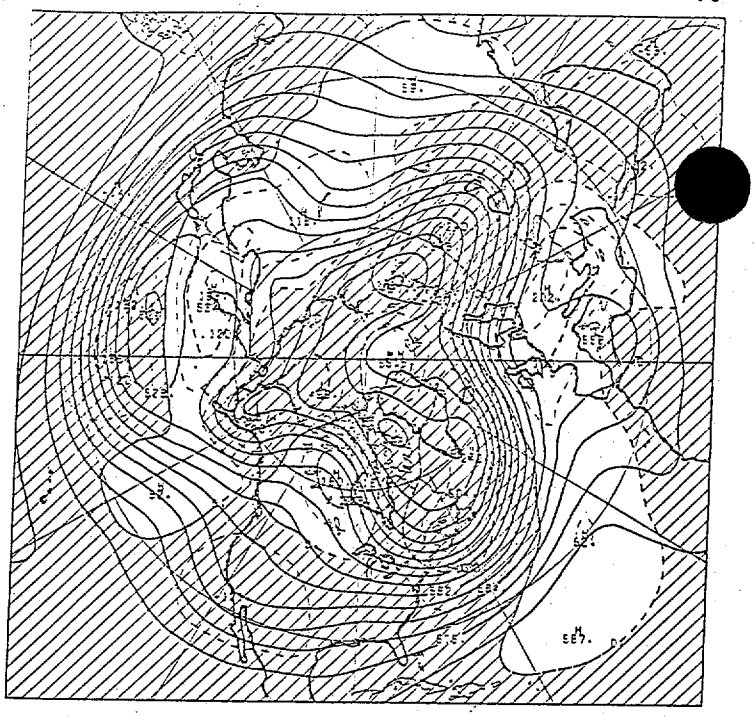


4 (CONT.)

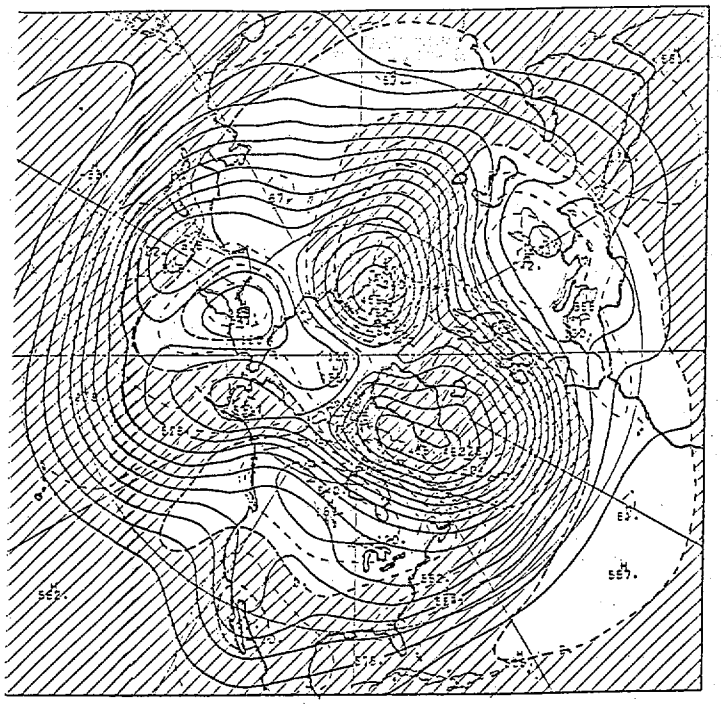
500 MB FCST FROM 0Z 2/ 2/93/ D+ 8 T62-P 9



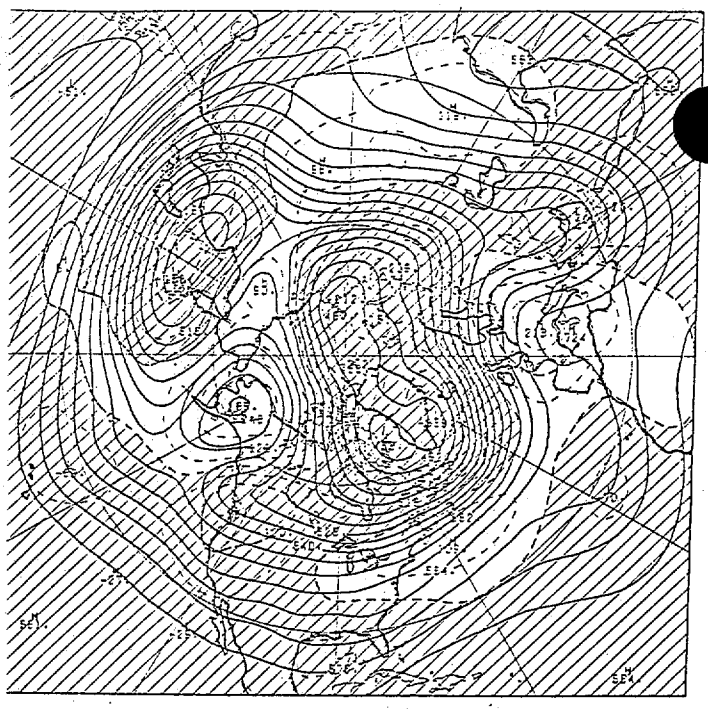
500 MB FCST FROM 12Z 2/ 1/93/ D+ 8 AVN 10



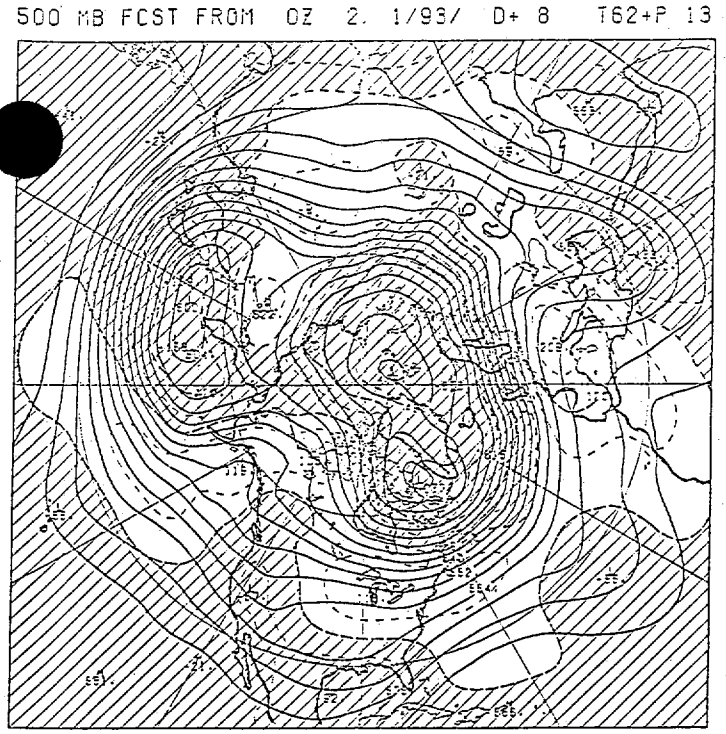
500 MB FCST FROM 0Z 2/ 1/93/ D+ 8 MRF 11

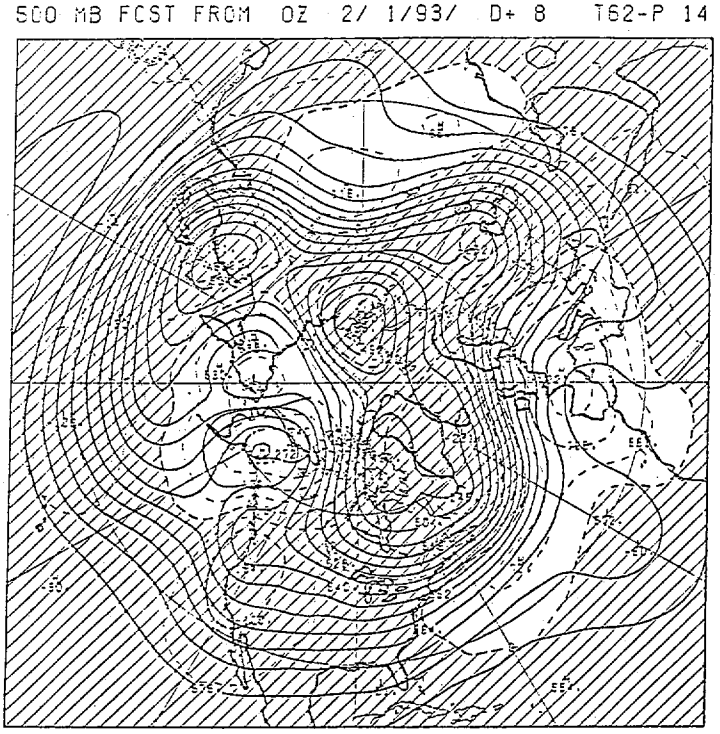


500 MB FCST FROM 0Z 2/ 1/93/ D+ 8 T62 C 12

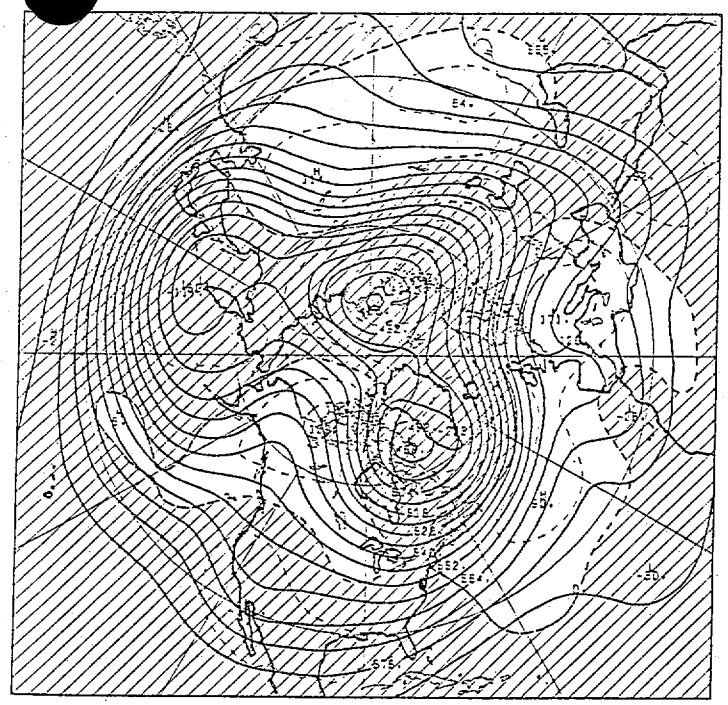


4 (CONT.)





4 (CONT.)



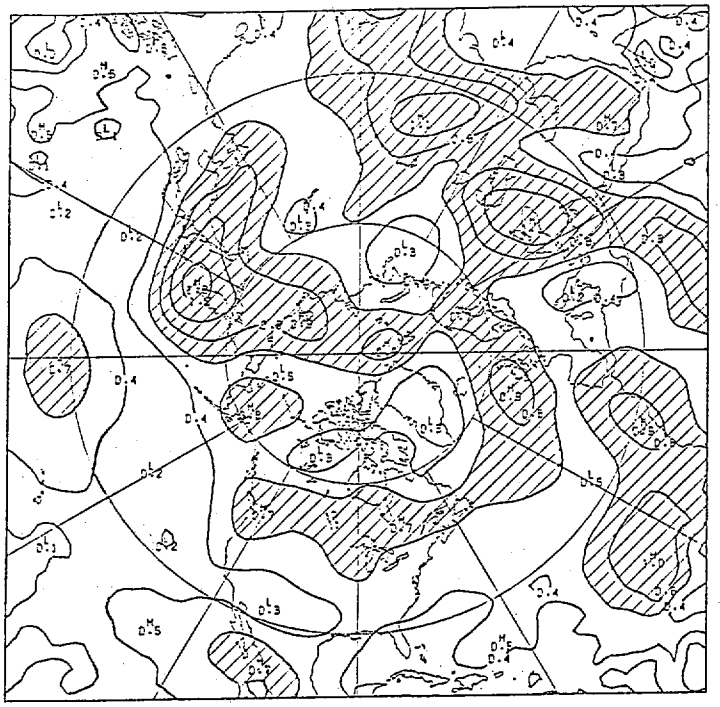
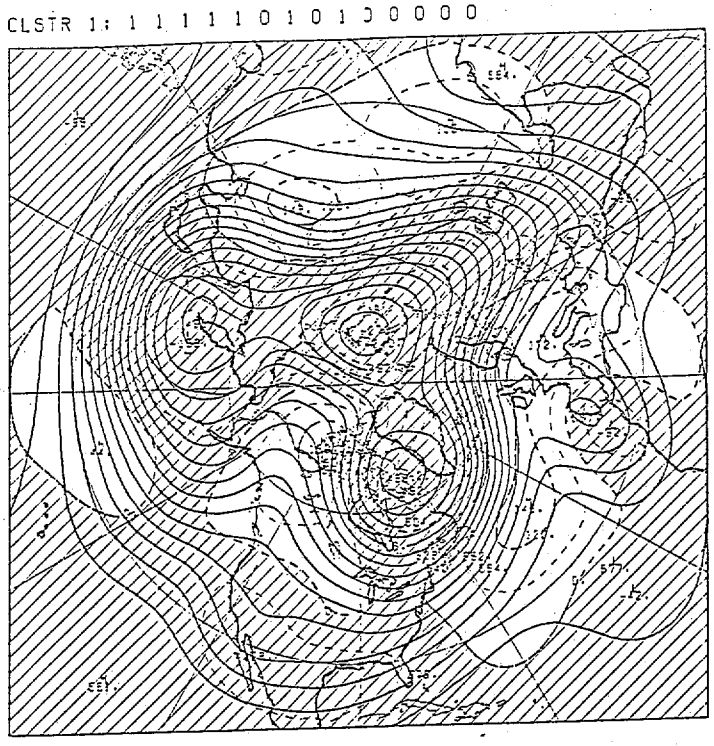


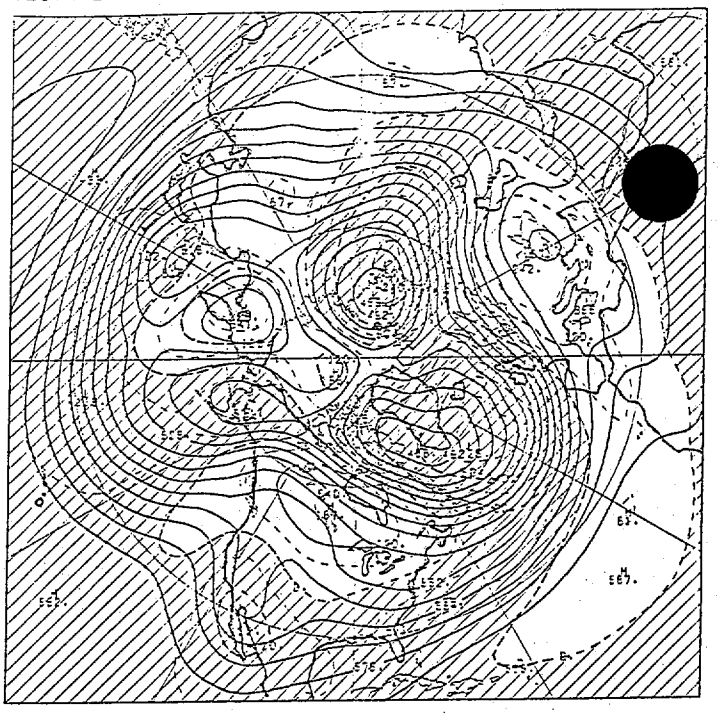
FIG. 5

FIG. 6

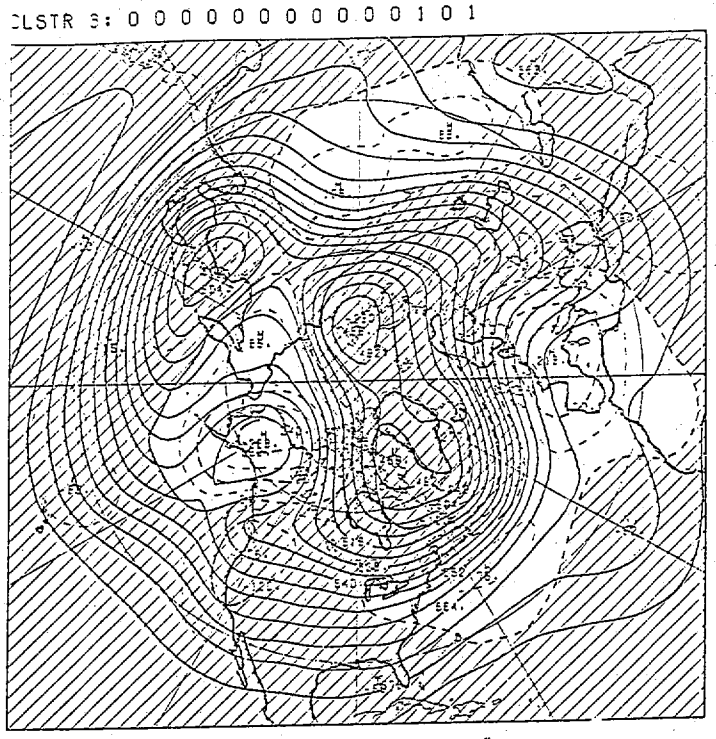
CLSTR 1: 1 1 1 1 1 0 1 0 1 0 0 0 0 0



CLSTR 2: 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0



CLSTR 3: 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1



CLSTR 4: 0 0 0 0 0 1 0 1 0 0 0 0 0 0

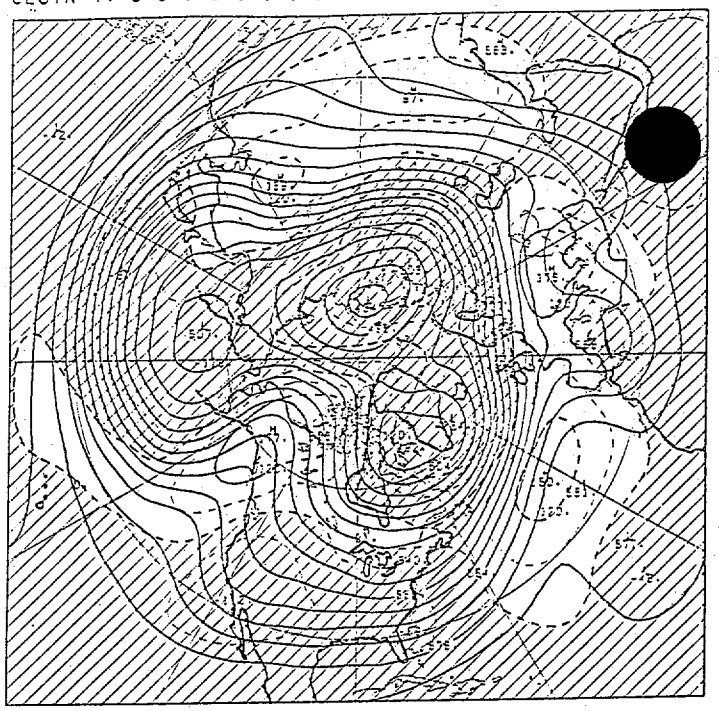


Fig. 7

CLSTR 1: 1 1 1 1 0 1 1 1 0 1 1 1 0 0; DAY 4

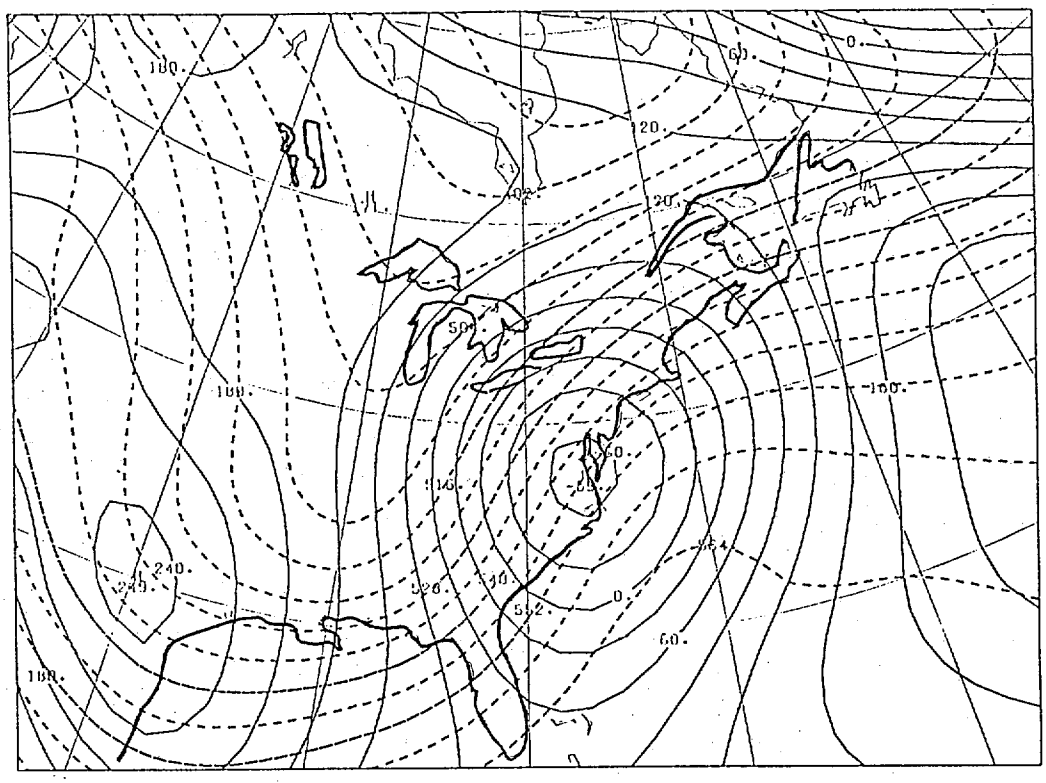


Fig 8

POS/NEG CNTRS: 0Z 2/ 3/93/ D+ 8

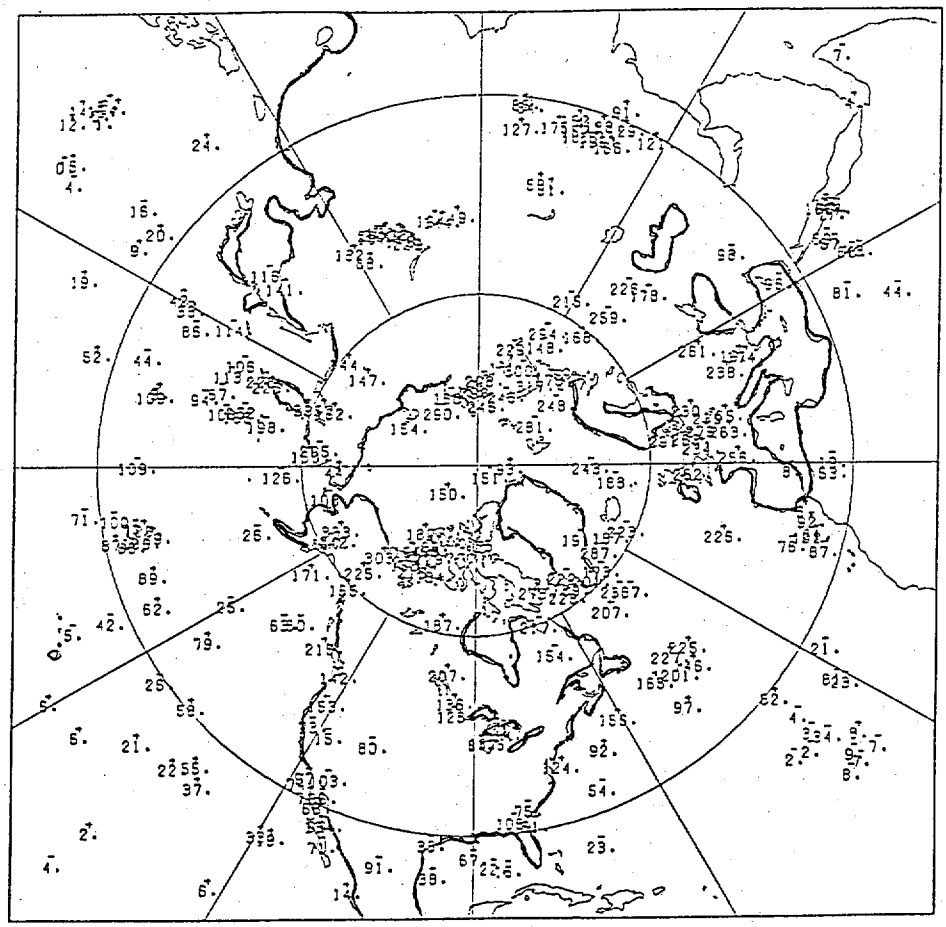


Fig 9

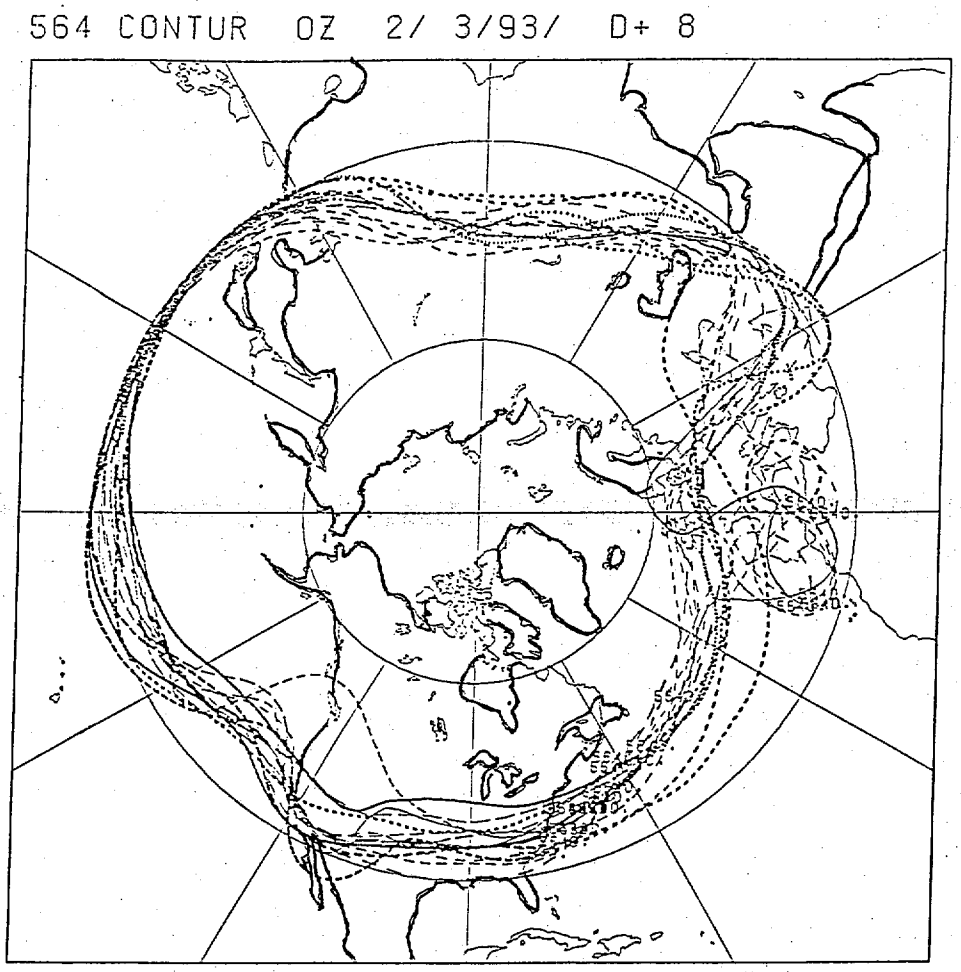
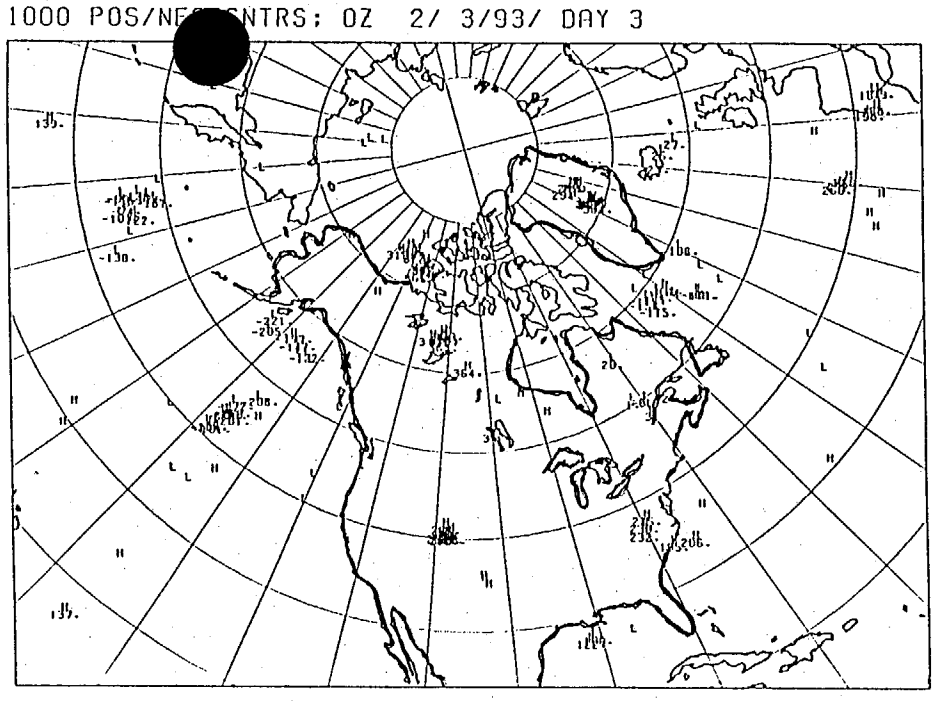
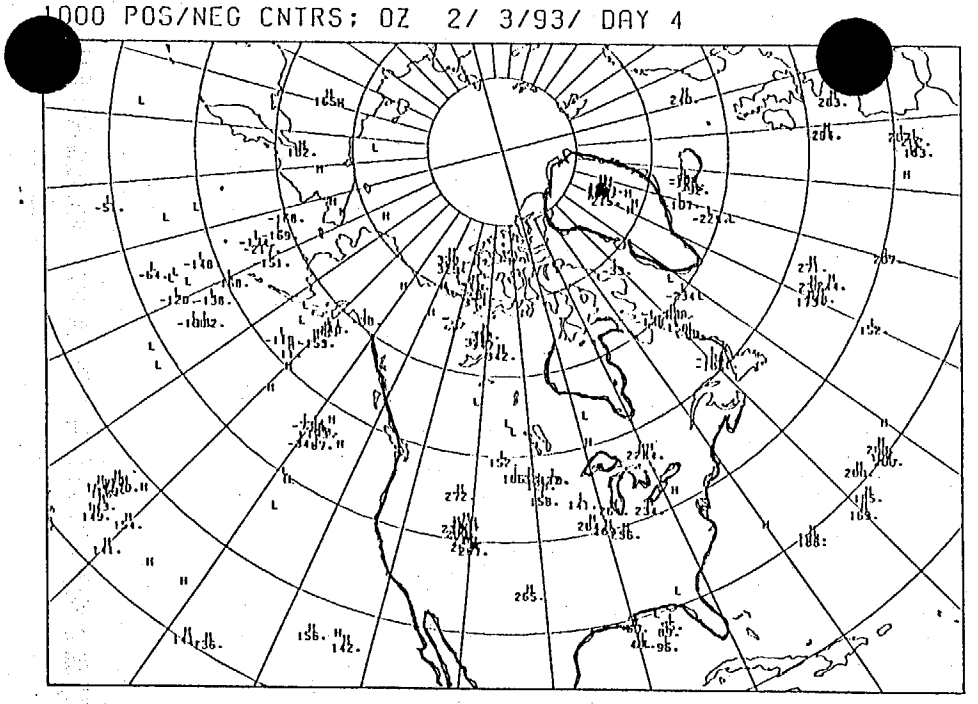
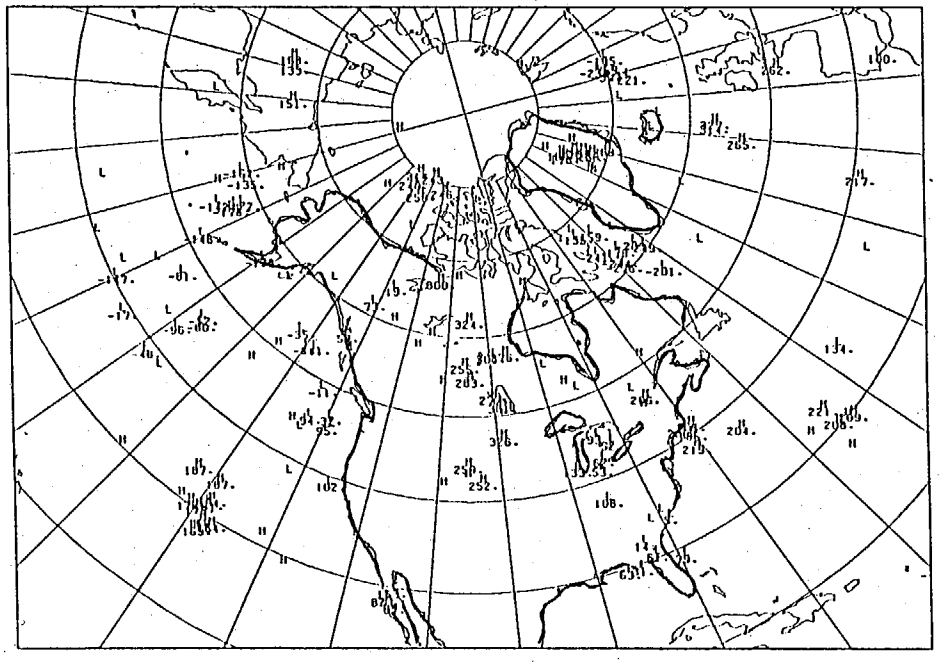


FIG. 10


1000 POS/NEG CNTRS; OZ 2/ 3/93/ DAY 3


1000 POS/NEG CNTRS; OZ 2/ 3/93/ DAY 4


1000 POS/NEG CNTRS; OZ 2/ 3/93/ DAY 5

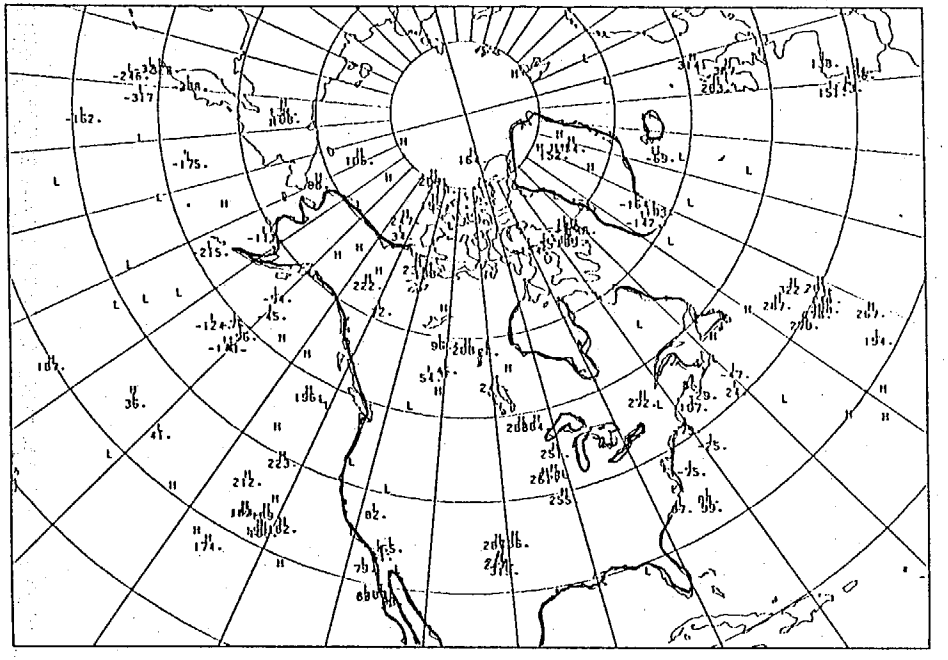
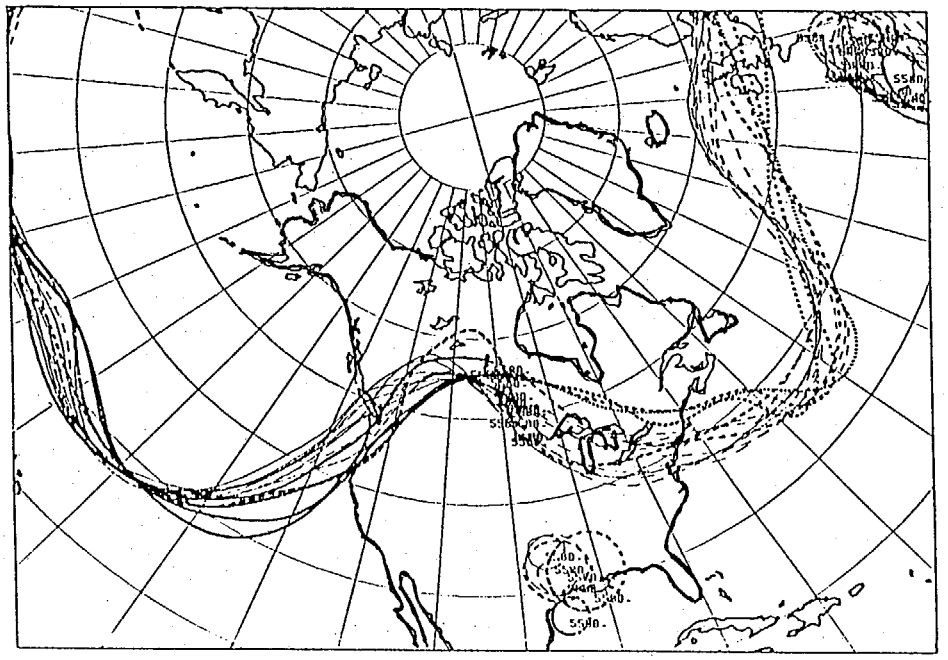
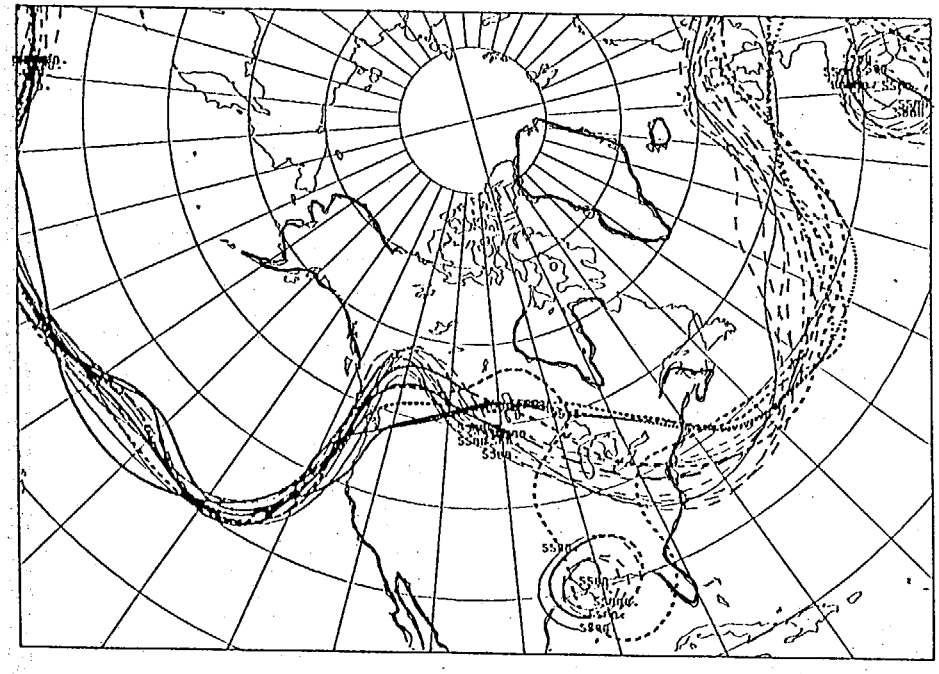

1000 POS/NEG CNTRS; OZ 2/ 3/93/ DAY 6

FIG. 11

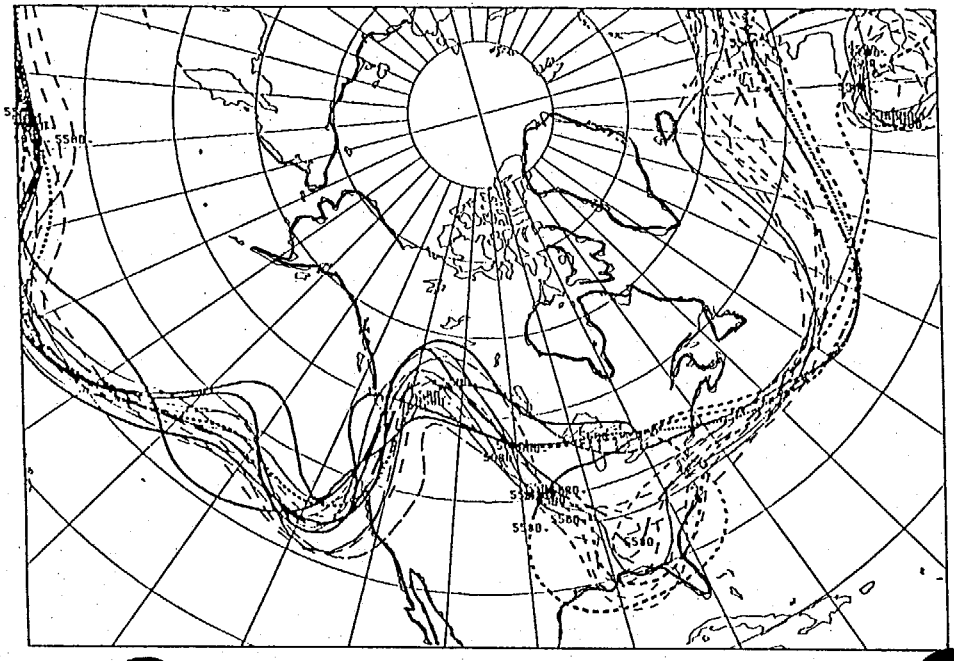
558 CONTUR OZ 2/ 3/93/ DAY 3



558 CONTUR OZ 2/ 3/93/ DAY 4



558 CONTUR OZ 2/ 3/93/ DAY 5



558 CONTUR OZ 2/ 3/93/ DAY 6

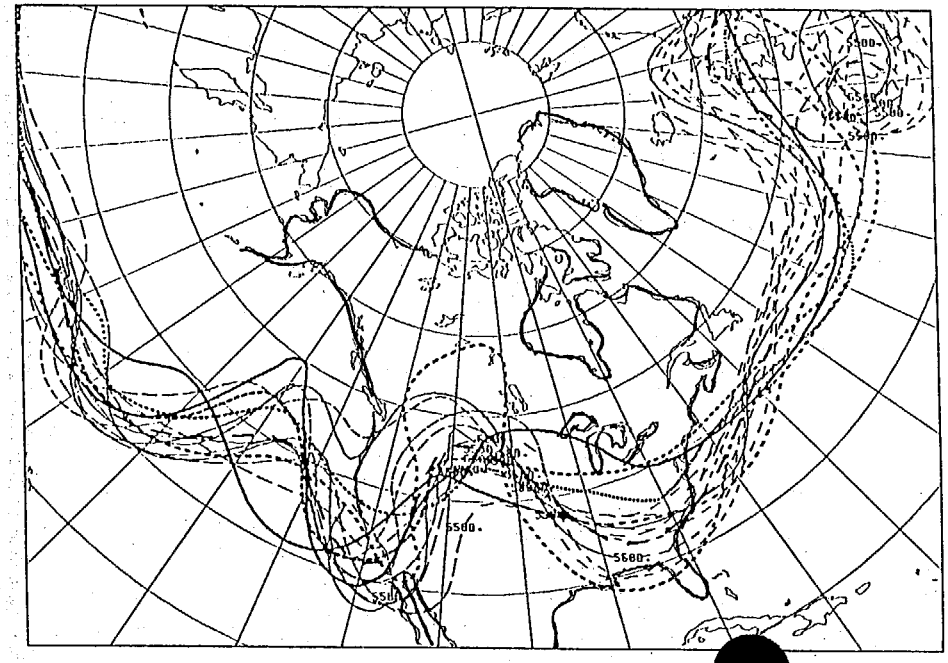


FIG. 12

% STD ANOM .GT. .5 OZ 2/ 3/93/ D+ 8

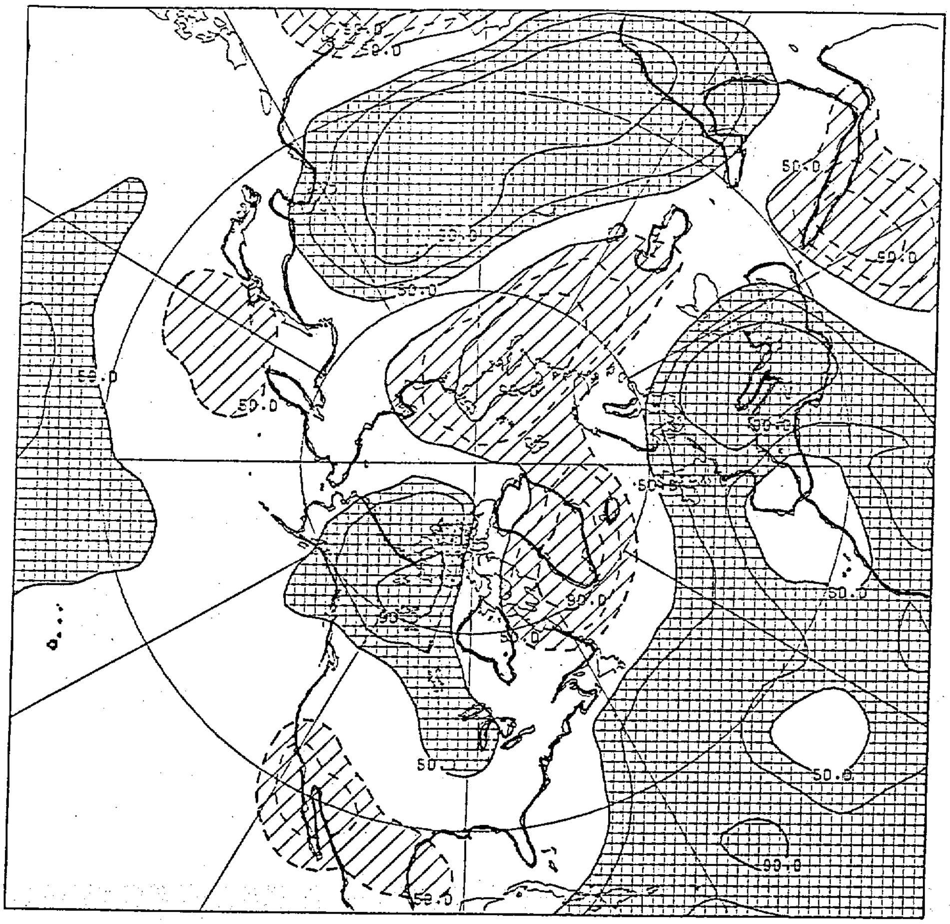


Fig 13

% TCK LESS THAN 540 OZ 3/10/93/ DAY 4

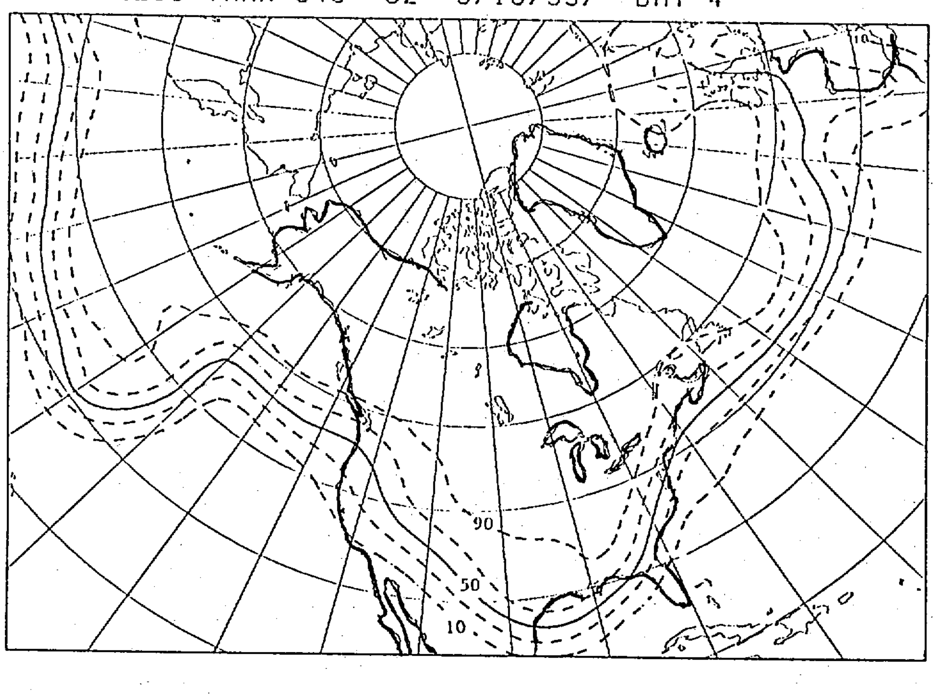
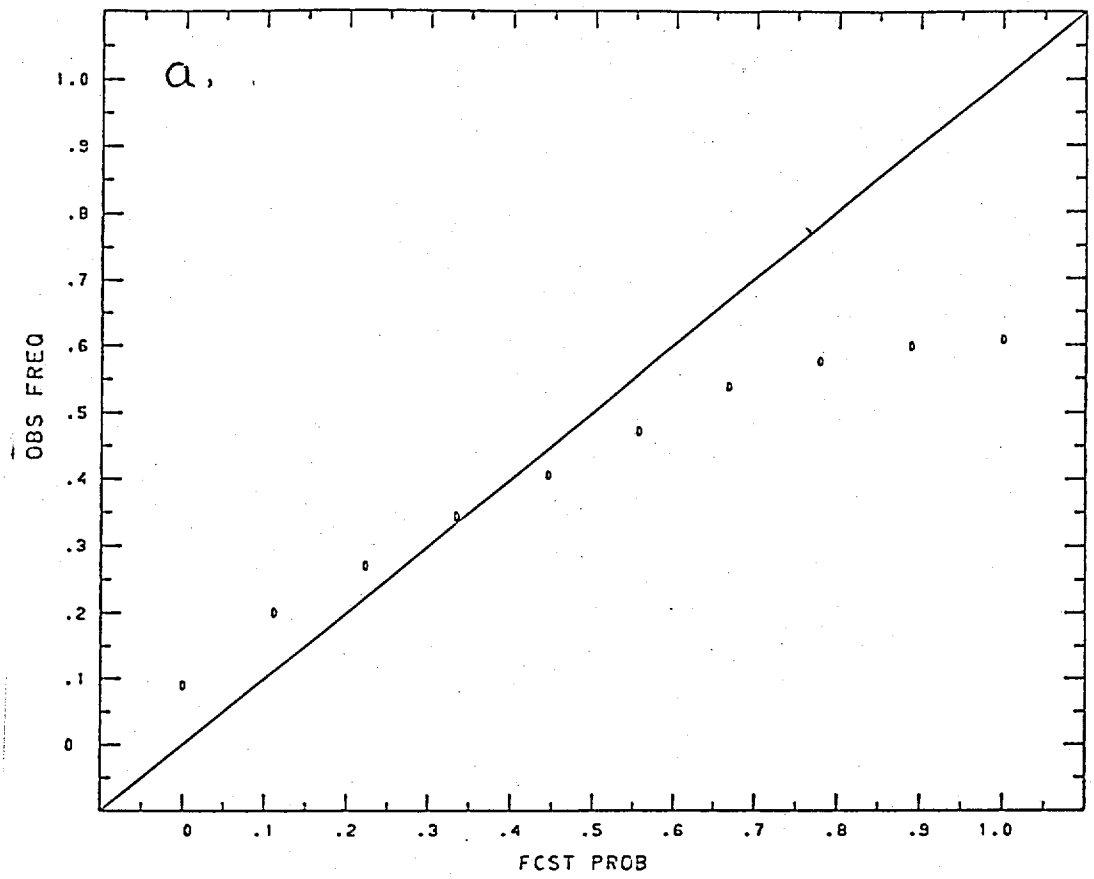


Fig 14


a.
OBS FREQ
FCST PROB
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0

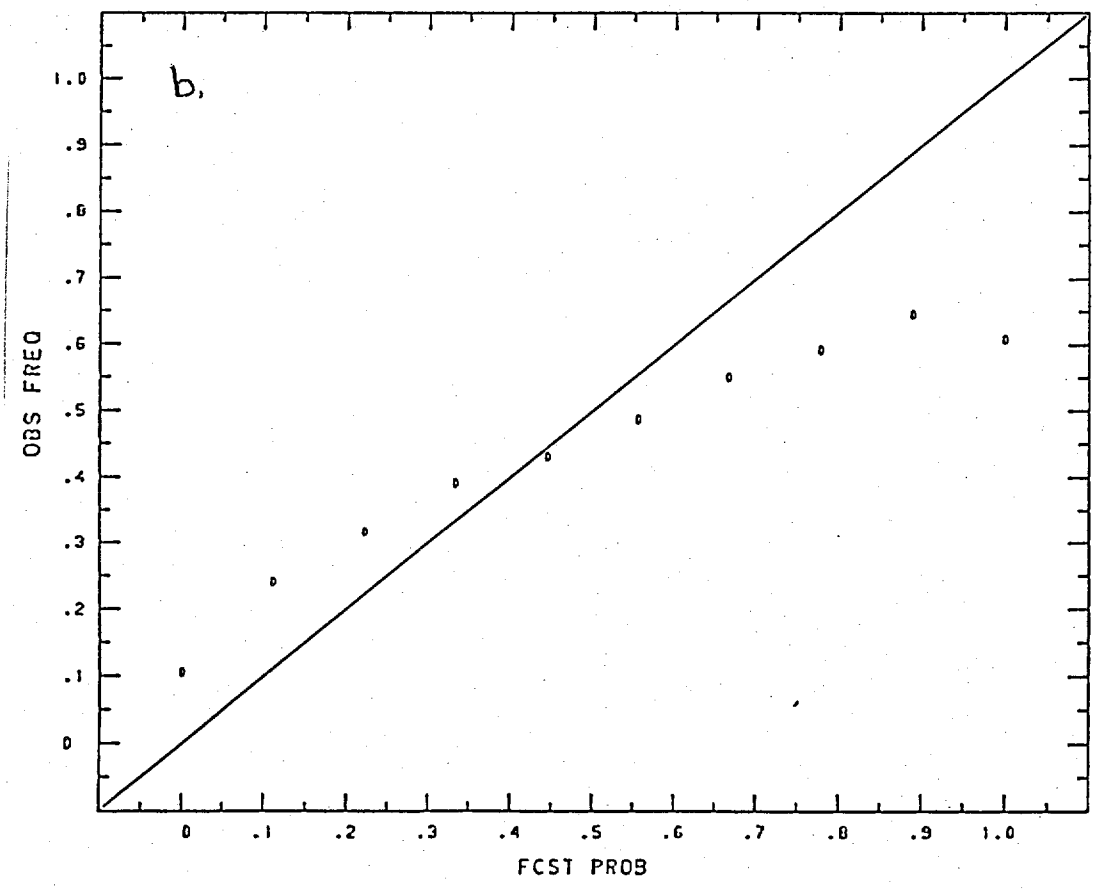

b.
OBS FREQ
FCST PROB
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0

Fig. 15