ACTIVITIES AND PLANS IN DYNAMIC EXTENDED RANGE FORECASTING (DERF) AT NMC

M. Steven Tracton
Climate Analysis Center
National Meteorological Center
Washington, U.S.A.

1. INTRODUCTION

The National Meteorological Center (NMC) is actively engaged in a program to assess the feasibility of operationally useful monthly predictions by extension of the basic model used for medium range forecasts (MRF). To assist us orient our efforts, NMC, in collaboration with the Center for Ocean Land Atmosphere Interaction (COLAI) of the University of Maryland, hosted a workshop on Dynamical Extended Range Forecasting (DERF) during October 2-4, 1985. Some of the problems considered were i) rationale and strategy for DERF, ii) experimental design, iii) model requirements, iv) diagnosis of model performance and behavior, v) evaluation and interpretation of experimental results, vi) archiving and sharing experimental data, and vii) coordination and collaboration amongst interested parties. A summary of the Workshop is presented in Section 4.

One of the major outcomes of the Workshop was a clearer picture of the overall direction of DERF activities at NMC. A plan which reflects that direction is outlined in Section 2.It provides the backdrop against which specifics can be refined through experience and continued interaction with other groups.

In the spirit of continuing collaboration and coordination a meeting of some of the"key players"in DERF was held at NCAR during the week of December 9, 1985. A key focus of this meeting was the best approach for NMC over the next year (PHASE I - see Section 2). A brief summary of the

NCAR meeting and the specifics of NMC's PHASE I plan that evolved from it are presented in Section 3.

Progress and results to date relevant to NMC's DERF activities are presented in Section 5.

2. NMC DERF PLAN

The following is the latest iteration of the NMC DERF plan which reflects input from the recent DERF Workshop, including the latest thinking vis-a-vis taking maximum advantage of the window of opportunity likely to exist soon after the expected arrival of a second Cyber 205 (or equivalent) in early 1987. This plan is intended to provide the backdrop against which specifics can be later refined.

FUNDAMENTAL OBJECTIVE: Assess the feasibility of operationally useful monthly predictions by GCM's that could be run by current or next generation computers. In NMC's case, the GCM envisioned is a logical extension of, or the same as, the model used for operational medium range predictions. To achieve this objective, a four-phase program is planned.

Phase I (Now to January 1987): This is a buildup period leading to a comprehensive performance test of DERF. Particular goals include:

1) Improvements to NMC medium range prediction model (MRF), e.g., physics parameterizations, inclusion of internal diagnostics, and refinements to lower boundary conditions.

2) Production, archiving, dissemination, and analysis of at least 10 experimental, 30-day integrations with the NMC model.

3) Collaboration with other research groups, including COLAI, GFDL, NCAR, ECMWF, the UK Meteorological Office and GLA, in assessment of model performance at the extended range and coordination with them on such matters as case selection, parallel runs with other models, and

identification of critical questions, issues, and strategies for assessing the feasibility of DERF.

4) Refinement of the design of the DERF performance test.

Phase II (January 1987 to January 1988): Controlled offline runs on the second Cyber 205 using the operational model as of January 1987. The goal here is to produce and make available for research within and outside NMC a large, relatively homogeneous sample of extended range forecasts for assessing feasiblity of operationally useful monthly predictions. The precise strategy of the performance test, e.g., case selection, frequency of runs, size of ensembles, monitoring and evaluation efforts, collaboration and coordination with other groups, etc., will be resolved during Phase I. It is premature to define the specifics at this point. The currrent thinking revolves around sets of biweekly ensembles on current cases supplemented as computer time permits with selected "canned" cases. (Note: the selection and canning must begin during Phase I).

Additionally, we will explore the possibility of complementing the full blown model runs with a lower order model. Parallel runs with this model can be generated to establish the correlation, if any, between it and the sophisticated model relative to error characteristics, climate drift, and overall performance. To the extent there is a correlation, the lower order, presumably more efficient, model could then be used to enhance the sample of the experiment. (If feasible within resource limitations, this possibility will be explored preliminarily during Phase I).

Phase III (January 1988 - January 1990): The principle focus here will be analysis and evaluation of runs made during Phase II. (This, of course, must begin at a lower level during Phase II in order to provide feedback to the experimental program.) Additionally, DERF integrations

will continue during Phase III as strategy warrants and resources permit.)

Evaluation will include: 1) Extensive diagnostics (internal and after the fact) of model behavior (systematic errors, variance, etc.), 2) application and assessment of approaches for extracting the maximum possible useful information, e.g., lagged average forecasting, applying statistical corrections, most appropriate averaging period, 3) assessment of the reliability and variability of skill and its prediction, including stratification of errors/skill/utility by regime, etc., and 4) comparison of results with other groups.

The final aspect of Phase III evaluation will be a summary of where we stand relative to the fundamental objective and recommendations of where to proceed from that point on.

Phase IV (January 1990-?): The approach here obviously depends upon the conclusions drawn from earlier efforts. The possibilities range from operational implementation of extended 30 day integrations to abandoning DERF. The middle ground is continued research, development, and experimentation. The availability or non-availability of a Class VII computer will set the framework for the sort and level of activity pursued.

3. PHASE I PLANS

3.1 NCAR DERF Meeting

A step in the direction of continued cooperation following the DERF workshop was a meeting of some of the key players in DERF at NCAR the week of 9 December. A list of participants appears in Appendix A. The meeting was arranged to take advantage of the collective presence of individuals concerned with DERF at the NCAR Model Intercomparison Workshop. A key focus of the meeting was the best approach for NMC over the next year ("Phase I") given relatively limited resources, and how NMC's efforts and

those elsewhere could most productively complement each other. The clear concensus was that NMC must concentrate on improving the Medium Range Forecast model (MRF), which eventually will be used for DERF. This should receive priority attention, even at the expense of pursuing questions of utility of the predictions at the extended range. That is, over the next year or so emphasis should be upon upgrading physics, etc. and reducing the model's systematic errors, not trying to emulate possible strategies (such as lagged average forecasting) of the Phase II DERF performance test. Other groups are presently in a much better position to address questions of that sort.

While there was some sentiment at the meeting for no further 30-day integrations by NMC until the model was improved, the mainstream opinion was that extended range predictions would be very useful in providing valuable feedback for model development (e.g., with respect to climate drift) and for establishing a baseline of performance for later testing of improved versions of the model. With this in mind the following specific recommendation emerged: extend the MRF to 30 days once per week (or every other week) starting with the mid-December case and continue to the limit (or near limit) of our approximately 10 case capability. Some resources should be held in reserve for rerunning later in the year with a presumably improved version of the model. If more extensive rerunning was necessary or desirable, it could wait until early next year when the second Cyber 205 (hopefully) becomes available. Such testing of the model's worthiness for extensive DERF experimentation might then be looked upon as the first step of what we're now referring to as Phase II. The specifics for the balance of Phase II strategy need not be addressed just yet.

With respect to the efforts by other groups, ECMWF (Tibaldi) expressed plans at the NCAR meeting to run extended range forecasts for the 15th

and 16th of each month commencing with August, 1985. These experiments will be with their T106 operational model. These runs (only 2 times at 24 hour intervals) are viewed as experiments in deterministic prediction, not lagged averaging. Their principal goals are to gain experience in DERF, accumulate information on systematic errors, and assess factors which might give clues to identifying situations and circumstances related to predictability.

The U.K. Meteorological Office (Palmer) expressed plans to run 7-case ensembles every few months from data centered around the weekends at mid month. Their principal aim is to assess assigning reliability estimates to the forecasts using the lagged averaging approach. Tibaldi and Palmer agreed to coordinate on case selection when the 15th and/or 16th of the month did not fall on weekends.

An agreement on case coordination, in fact, was a major result of the NCAR meeting. GFDL (Miyakoda), NCAR (Baumhefner and Williamson), CCM (Boer), and ANMRC (Bourke), although their specific plans had not yet been defined, agreed upon focusing their experiments on common cases. Also, while it was the general view that using canned cases tends to render results unrepresentative, it was agreed that one FGGE case, January 16, 1979, was of sufficient interest and import that all present tentatively agreed to run from this time. It should be noted that intrinsic in the comparisons which will be possible is recognition that an ensemble used for assessing the range of uncertainty in predictions can be runs from different models from the same dates, not just sequential runs from a given model.

In addition to agreeing upon cases, the group at NCAR decided upon the minimal set of products that should be generated for the purpose of intermodel comparisons. Output should be saved at 24 hour intervals and

ten day running means produced every 5 days. Parameters are 500 mb height and 850 temperature in the extratropics (+/- 20 to +/- 80 deg) and 850 and 250 mb stream function and velocity potential for the tropics. Verification of these same fields will be the anomaly correlation and RMS error. Additionally, the 30-day mean fields and their verifications should be generated. (Tibaldi raised the point that anomalies and verification of same are sensitive to the climatology upon which they are based and agreed to provide the one used at ECMWF (Oort and Crutcher) to ensure consistency amongst groups. S. Tracton will serve as the intermediary in distributing this climatology).

3.2 PHASE I Specifics

The recommendations for DERF activities at NMC over the next year or so, which emerged from the NCAR meeting, have since been refined into a specific strategy that hopefully will reap the maximum gains given our objectives and available resources. The plan, as it now stands, is as follows: an experimental, 40 wave, 16 (or 18) layer version of the MRF with enhanced physics will be run to 30 days from the initial conditions of 1200 GMT January 19, February 16, and March 16, 1986. The selection of cases reflects coordination with Tibaldi at ECMWF and Palmer at the British Meteorological Office. The runs are expected to be made with about a one month lag, i.e., the January case will be run in mid-February, etc. We will use the first 30 days of a 90 day Center for Oceans Land Atmosphere Institute (COLA) integration with the MRF from 1200 GMT 15 December, 1985, as our run for that case.

It is expected that in early summer the above cases will be rerun with a further enhanced version of the model in order to assess the impact of the model changes. Pending the results of this assessment, one version

or another of the model will be used to generate extended forecasts from selected initial conditions during July, August, and September of this summer. Again, the choice of specific cases will be coordinated with other groups.

In addition to the individual forecasts outlined above, one three case ensemble will be run for the period around January 16, 1979. While, in general, our aims this year are best satisfied with a set of largely independent runs, the ensemble approach here will provide us with at least some experience in this area. We will probably run the ensemble experiment sometime during the next few months with the same model used for the Jan, Feb, March, 1986 cases. Resources permitting, we'll then have the option of rerunning later in the year, should that be necessary or desirable.

Evaluating the 30-day runs outlined above is, of course, a critical aspect of our activities. A prime objective is to provide feedback for continuing improvments in the MRF. Additionally, the evaluation process, in combination with the "nitty gritty" of producing the extended range predictions, will provide invaluable experience as we move towards refining the experimental design of the comprehensive DERF test anticipated next year.

4. WORKSHOP SUMMARY

The DERF Workshop at NMC was organized into four sessions with open discussion an integral component of the proceedings. The following is a summary of the presentations and discussions of each session based on the notes of designated rapporteurs. A list of participants appears in Appendix B.

4.1 Keynote: Current Status and General Strategy for DERF, Don Gilman, NMC's Climate Analysis Center (CAC)

This address consisted of three parts: (1) a description of the current status of long-range forecasting at CAC, (2) an outline of the motivation, goals, and basic elements of DERF, and (3) problems likely to be encountered in achieving the goals of DERF.

Currently, CAC issues monthly and seasonal outlooks for temperature and precipitation in three categories (above,below and near normal) with the probabilities for each. The probabilities reflect the level of skill that the outlooks have achieved in the past. Skills are approximately the same for the monthly and seasonal outlooks,and have shown little improvement over the past several years. The tools used in producing these outlooks consist of statistical lag correlations of 700 mb height anomalies, analogue circulations for specifying surface temperature and (especially) precipitation, statistical regression for temperature specification, and, for the monthly outlook, the first and second five-day period mean charts of the 10 day numerical guidance of NMC and/or the European Centre for Medium Range Weather Forecasts (ECMWF). There is evidence that use of the model predictions has resulted in some improvement of the monthly outlooks.

The reasons for seriously considering DERF at this point are fivefold: i) models have reached a level of sophistication that makes them potentially useful for operational monthly predictions, ii) computer power has reached a level that at least borders on permitting the extensive testing and experimentation that will be required, iii) improved global data coverage, especially satellite observations, tempers degradation of forecast skill related to inaccurate initial conditions, iv) predictability estimates do not preclude the possibility of useful skill beyond 10

days in most geographical locations and seasons, and v) some pioneering prediction experiments at GFDL, UKMO, and ECMWF have given encouraging results.

The DERF project really has only one goal: apply a GCM to help produce operationally more useful (not necessarily more accurate) monthly predictions. A vital ingredient of a more useful forecast is a measure of its uncertainty.

The basic elements of DERF are model development, a large set of experimental extended range predictions, and evaluation of the results. The paths from these elements to achieving the aforementioned goal involve three sorts of questions: i) model performance, ii) estimation of uncertainties, and iii) fitting applications.

These categories of questions are discussed in reverse order.

Fitting applications involves questions such as the length of model runs (e.g. 30 vs 20 days), averaging period and relative weighting of the individual days comprising the mean, highlighting extreme anomalies, specificaton of trends and overall measure of variability within forecast period, identification of storm tracks, and specification of precipitation anomalies directly.

Questions on estimating uncertainties include the nature of ensemble runs (e.g., Monte Carlo vs lagged averaging), identification and removal of systematic biases, and regime dependency of skill.

Questions of model performance relate to how much testing is required to certify the model as viable for DERF, the response of the model to boundary conditions, the use of simpler models as controls, and the likely sensitivity of results to changes in the model during the course of the experiment.

In conclusion, although one should be optimistic about the eventual success of DERF, there will be hard decisions, forced in part by limited resources, on precisely how to proceed. No doubt many years of continuing effort will be required.

In response to a question, Gilman suggested that monthly time scales are probably too short for requiring a coupled ocean/atmosphere model, with the possible exception of the tropics. For seasonal outlooks, coupling is obviously more important, but the DERF program at NMC cannot afford to be slowed down initially awaiting development of a coupled model.

In response to another question, Gilman indicated that users want as much information as possible, especially that related to extremes, the amount of variability within the forecast period, the conditions for critical application periods, and a measure of the uncertainty of the prediction.

4.2 SESSION I: Rationale and General Strategy for DERF

The first of the four papers in this session, Physical Basis and Feasibility of Model Long-Range Forecasting, was presented by J.Shukla of the Center for Land Ocean Atmosphere Interaction (COLAI) of the Univ. of Maryland. Shukla began by noting that there are basically two ways of showing that long-range forecasting using GCM's is feasible. The first is through theory and model-based experiments addressing the question of atmospheric predictability. The second is to demonstrate that actual experimental forecasts are useful. This talk focused on the former approach. Four topics were discussed:

1) Dynamic Predictability- In dynamic models and the real atmosphere, most energy is in the lowest frequency waves, especially when one considers the monthly average circulation. Such time averages are predictable

for longer scales than the typical 7-10 days of useful skill for individual days. Results with the GLAS model indicate predictability (relative to persistance) of 30 days for waves 0-4 and about 15 days for waves 5-12. Initial conditions appear important for time-averaged predictions for the period D+8 through D+37, but relatively unimportant for the period thereafter.

2) Boundary Forced Predictability- Small changes in boundary heating can be converted into large and deep atmospheric heating through the CISK mechanism and positive feedbacks, especially in the tropics. In the GLAS model a much improved precipitaton anomaly is predicted when SST anomalies in the tropical Pacific are included. Improvement in mid latitudes from tropical SST anomalies is less clear, and more model studies are needed. Aside from SST anomalies, attention must be given to other boundary forcing mechanisms, including snow cover, sea ice, and soil moisture.

3) Abilities and Limitations of GCM's for DERF- The major problem with prediction models is that of climate drift. Three ways of dealing with this are; i) improve model resolution and physics, ii) remove weak interaction drift and transients at the end of the model run, and iii) multiple overlapping runs such as lagged average forecasting. More crudely, strictly empirical corrections (e.g. in zonal subtropical heating) can be used to improve the prediction by removing systematic errors.

4) Computer Needs and Satellite Data- Global boundary conditions for model experiments must be archived,just as we now archive atmospheric initial conditions. Satellite observations are very important, e.g., through proxy specification of precipitation from radiance data. The inferred precipitation is important both for initializing and verifying a model. In regard to computer requirements, allowance must be made for multiple run techniques for gauging model accuracy and reliability.

In response to a question on the predictability of the planetary waves, given that there is an energy cascade from shorter to longer waves, Shukla said that long waves develop much of their energy at their own length scales, and this part of their development should be predictable for long time periods. Boundary conditions must to a large extent force the quasi-stable long waves, e.g., SST forcing in the tropics. At higher latitudes, baroclinic development is more likely to interfere with the long waves and thus limit their predictability.

In response to a question about correcting for climate drift, Shukla said that continuous adjustment to correct for the drift is preferable to one correction at the end of the forecast period.

The second paper of this session, Major Considerations for Operational Statistical-Dynamical Long Range Prediction, was a joint presentation by E. Kalnay and R. Livezey of the Goddard Laboratory for Atmospheres (GLA). Kalnay began by noting that the objective of the Long Range Prediction Program at GLA is to explore systematically the existence, if any, of monthly and seasonal predictability with comprehensive atmospheric models with emphasis on a priori estimates of forecast skill. The approach is four pronged:

1) Predictability studies; develop parallel series of up to 45 day GCM runs in all seasons in order to study the relative contributions of internal dynamics and external forcing to actual and theoretical predictability.

2) Statistical-Dynamical Prediction; develop methods to extract the maximum information from extended NWP, with emphasis on adaptation of lagged average forecasting (LAF) and new methods of a priori prediction of error.

3) Case studies; conduct case studies of the predictability of persistent anomalies.

4) Low-frequency circulations; conduct diagnostic and global modeling studies to describe the climate system on time scales of interest.

Predictability studies at GLAS show that the long wave (0-4) errors are still growing after 10 days, whereas waves 6 and larger have mostly saturated error variances after 10 days. When predictability is defined as existing until the error variance reaches 95 per cent of the saturation value, the ECMWF model indicates two week predictability is feasible through waves 4 or 5. As illustrated through an investigation of the summer 1980 heat wave in the U.S., the GLAS model can be skilful even well beyond the normally expected predictability.

Livezey discussed some considerations for operational statistical-dynamical long range prediction at NMC. They include:

1) Motivation- why DERF; NWP has already made an impact on the monthly outlook at NMC. Use of the D+3 and D+8 numerical guidance appears to have resulted in some improvement of the forecasts. Also, as noted by Kalnay, skill remains in NWP in an ensemble sense after 10 days.

2) Requirements for delivery of the optimum product to users; Model results must be appropriately postprocessed for use in objective and subjective specification or extrapolation processes. In particular, the model climate drift must be corrected for, and unpredictable and/or unimportant higher frequencies should be filtered out. Additionally, measures of the uncertainty of the predictions need be provided.

3) Range of approaches required and some major unsolved problems; One of the major problems requiring innovative measures is the fact that model error varies with several parameters- time of year, location, scale of disturbance, length of forecast period, averaging period, errors in

initial conditions, and circulation regime. The importance of the last two is generally not well recognized. LAF may help the former of these since runs starting from differing initial conditions may average out the detrimental effects, e.g., of omitting an important circulation feature in a given analysis. The dependence of error on circulation regime is related to the problem of climate drift and is probably as important as dependence on model configuration.

4) Research priorities and computational issues; Priorities are regime dependence of forecast reliability, minimizing vulnerability to analysis inconsistency, minimizing effects of non-stationarity, and best treatment of scale dependence and temporal averaging. Computing issues are how long (e.g. 20 vs 30 day runs), frequency of runs, and how large a supplementary history is required to address questions such as the dependence on regime.

R. Daley, in the context of a discussion on minimizing the strain on computer resources, suggested using a simplified version of the ultimate model used for DERF experiments. To the extent that simple and full blown models had the same error and bias characteristics, the simple, more computationally efficient model could be used to enhance the sample e.g., in studies of regime stratification.

In the third paper of this session K. Miyakoda discussed Experimental Extended Range Forecasting at GFDL. The talk consisted of two parts. The first part dealt with predictions to 30 days in the context of 8 winter-time cases from the years 1977-1983. In comparison to persistence and climatology the results beyond 10 days were not impressive. The ensemble bias in the 10 - 30 day range contributed about 60% of the total rms error. Correcting for this bias (climate drift) improved the forecasts in all but one of the cases . The January 1983 case forecast was degraded,

possibly because the climatic regime that month was significantly different from the other cases. For the ensemble, application of the drift correction increased the average skill over climatology and persistence to about 20 days.

Part two of this talk focused on the importance of very accurate specification of boundary conditions (especially SST) in extending predictions beyond one month. In one experiment use of actual vs climatological SST did not result in improvement of the forecast of the 30-60 day mean. This was likely related to inaccuracies in the SST field because of sparse data in the tropics. Modifying the SST's on the basis of the relationships between satellite measured outgoing longwave radiation (OLR), inferred convection, and the anomaly of SST resulted in slight improvement of the predictions. The improvement was especially noticeable in the tropics, but improvements poleward of 25N were also noted. Miyakoda concluded that the best forecasts would be achieved through both correcting for the climate drift and SST modification.

In the ensuing discussion, Hollingsworth pointed out that the upper-level, cloud-tracked winds in the Eastern Pacific in January 1983 were deficient, and that should be taken into account in evaluating this case. Van den Dool suggested that improvement of the experimental results with the modified SST's was really the result of implicitly specifying upper-atmospheric heating in the tropics due to convection more accurately. Miyakoda responded that higher latitude effects are "secondary" in nature and would probably be about the same whether SST or latent heating anomalies were specified more precisely.

The next presentation of this session was by Andrew Gilchrist, who discussed DERF related activities at the United Kingdom Meteorological Office (UKMO). The UKMO strategy for DERF research developed in 1976 with

the following broad aims; i) determine a practical methodology for using GCM's within the existing LRF group, ii) provide documentation of the existing 5-level hemispheric model properties relevant to LRF, e.g., its ability to reproduce the observed atmospheric spectrum on relevant time scales, and iii) provide a reference against which the performance of future models could be assessed.

To achieve these aims, two 50-day runs per month were made using real initial data and climatological (or partly climatological) SST's for several years. Additional runs were performed, mainly for successive days or with observed SST's, as computer time allowed. A total of around 70 integrations were made. The main results were; i) predictive skill was greatest in winter and non-existent in summer (assessment has concentrated on winter), ii) despite comparatively low resolution (approx 330km) and use of a hemispheric model important aspects of the geographical and time variations of blocking were reproduced, iii) an estimate of potential predictability (defined as the time for RMS differences between forecasts one day and one year apart to become insignificant) was 26 days for daily forecasts and 33 days for 15-day averages, iv) actual predictive skill was much less, about 7 days on a similar basis, v) some forecasts were much more skilful than others, and tended to be so throughout the forecast period - the best eight winter forecasts had significant skill (based on anomaly correlation scores) to at least 30 days, and vi) use of observed SST's improved skill in 4 out of 5 years - for the two cases available in 1977 the 15-day anomaly correlation was about 0.6 to beyond 30 days.

The experiment is continuing using a global 11-layer model (2 1/4 X 3 3/4 deg resolution). At present, ensembles based on analyses 12-hr apart are being run once per season. Results of the autumn forecasts completed

on 15 September 1985 showed considerable variation, but forecasters preparing the autumn forecast considered them helpful.

The final paper of this session was by A. Hollingsworth on DERF Related Activities at ECMWF. The following issues were discussed; i) developments in DERF at ECMWF during 1980-1985, ii) variations in forecast skill with the ECMWF model, and iii) results of DERF experiments.

Hollingsworth discussed the need to improve the model rather than empirically correcting for systematic errors. Subtracting out the climate drift improved forecasts, but not significantly so. Amongst the improvements suggested (and , in fact, already implemented operationally) are use of the envelope orography and increased resolution (T106). DERF experiments with variable resolution and orography showed that the model exhibits larger sensitivity to orography with higher resolution. The same applies relative to model sensitivity to parameterizations of physical processes. Overall, the synergy between enhancement of resolution and improvements in physical parameterizations (radiation and convection) reduces (but does not eliminate) systematic errors.

Hollingsworth cited recent studies at ECMWF examining the question of the variability in skill as a function of regime. In one investigation the skill scores were higher if a block exists in the initial data or develops within 3 days thereafter. In another study, marked correlation was demonstrated between hemispheric forecast skill and the level of high frequency activity in the western Pacific. Hollingsworth stressed the need for similar studies using other models.

In one case (17 Jan 1984) the ECMWF model demonstrated remarkable correspondence to reality through the 20 - 30 day range in describing the daily sequence of an evolving blocking situation.

Studies of LAF at ECMWF suggest that improvements decrease as resolution is increased, but more work is necessary in this area.

In the open discussion Blackmon and Gilchrist reemphasized the need to assess how various models treat spells. In response to a question from Kalnay, Gilchrist indicated there has been no comparison of UKMO skill scores and those of US models. Shukla commented that the real question is how much better we can do compared to the relatively modest skill of Gilman's outlooks, not how models do relative to one another. He further noted that there is validity in claims that prediction at > 10 days is feasible, but Blackmon, referring to Chervin/Tribbia studies,differed on this.

The question arose whether there is need for a low resolution model to generate a large number of cases if a good high resolution model is available. Hollingsworth argued for staying with the higher resolution version in DERF experiments, at least to the extent that computer resources permit. Shukla asked whether high resolution with good physics always yields better forecasts as Hollingsworth suggested. Hollingsworth responded that there was no question about it. Gilchrist argued vigorously for performing ensemble average studies with various models irrespective of model resolution.

Hollingsworth pointed out that as the forecast range is extended variations in forecast skill increase. The uncertainty of predictions deserves close attention, for it is a crucial ingredient for the consumer. Roads noted that, in addition to improving forecasts through enhanced physics and increased resolution, one must also be concerned with improving initial conditions.

4.3 SESSION II NMC/COLAI EXPERIMENTAL DESIGN

S. Tracton began this session with an outline of NMC's proposed DERF plan. The next year (through FY 86) is viewed as a buildup period leading to a comprehensive test of DERF beginning early in 1987. The key activities during this phase include: i) improvements in the NMC medium range prediction model (MRF), especially in regard to physical parameterizations, inclusion of internal diagnostics, refinements to the lower boundary conditions, and addition of a diurnal cycle, ii) production and analysis of at least 10 30-day experimental integrations, ii)collaboration and coordination with other research groups (e.g., GFDL,ECMWF,GLA and NCAR) on relevant questions of strategy, experimental design, and evaluation, and iv) refinement of the design of the DERF performance test.

The principal objective of the DERF performance test is to produce and make available for research within and outside NMC a large, homogeneous sample of extended range forecasts in order to establish the feasibility of operationally useful monthly predictions. As presently envisioned (consider this a "strawman" proposal) the main components of the performance test include:

1) extension of 10-day MRF to 30 days on 3 consecutive days biweekly for a period of at least 3 years.

Presumptions here include; i) a 3-year period is the minimal time necessary for obtaining an adequate data set for statistical analysis of the significance of the experiment, and ii) forecasts on 3 consecutive days is the minimum necessary for evaluating LAF and variations of error dispersion.

2) extended range forecasts on alternate weeks with a less sophisticated model (w.r.t. horizontal resolution and physics)

The presumption here is that a baseline of model capabilities can be established with a simpler model, and sample size can be enhanced.

3) evaluation

The principal elements here include; i) extensive diagnostics (internal and after the fact) of model behavior (systematic errors, variance, etc.), ii) assessing approaches for extracting the maximum possible useful information (e.g. LAF, Epstein statistical corrections, most appropriate averaging period,etc), and ii) reliability and variability of skill.

Finally, it was noted that aside from the question of the utility of extended runs the DERF experiment should provide significant feedback relative to increased understanding and improved performance of the NMC model in the medium range.

In the discussion the question was raised as to whether the model would remain fixed during the experiment. Tracton acknowledged probably not, and the ramifications of this must be carefully examined. NMC most certainly encourages suggestions on alternative strategies, but they must keep in mind the resource limitations and operational responsibilities of NMC. The idea of pooling the resources of several groups (e.g., in parallel runs on selected cases with differing models) was aired and received an enthusiastic response.

In the second presentation of this session J. Gerrity and G. White discussed the structure and performance of the NMC spectral model. The model has 18 layers with rhomboidal 40 wave resolution in the horizontal. Physics presently includes dry adiabatic adjustment, a Kuo type cumulus convection parameterization (to 300 mb), and a radiation package based

upon that of GFDL. Soil moisture and snow cover are allowed to change during the integration and the current SST analysis is used. The terrain is the "silhouette" topography of F. Messinger. At present there is no diurnal cycle.

Evaluation of the one extended integration performed to date suggested the error patterns at the 30 day range were similar to the systematic errors observed in the MRF day 10 fields. These in turn, especially at low levels, appear to emerge in the first 12 hours and probably reflect some sort of initial imbalance. Another likely source of error in the model is that due to improper treatment of tropical convection which results in, among other possible effects, misplacement of the Indian monsoon. Intense efforts are underway to upgrade the model, especially in the areas of representing physical processes and diagnostics.

J. Shukla introduced a series of presentations by COLAI staff on ancillary experiments with a research version of the NMC spectral model. The research plans include i) analysis of forecasts at the medium and extended ranges, ii) climate simulation and analysis, iii) predictability and sensitivity studies (e.g. w.r.t SST), iv) dynamical diagnostics, v) interactive biosphere, and vi) studies on the sensitivity to physics parameterizatons.

J. Kinter discussed various aspects of creating a research version of the NMC model. Among the projects completed are the updated documentation and inclusion of GCM diagnostics. Studies are planned on the model climatology through summer and winter case runs to 90 days and analysis of 10 and 30-day systematic errors.

E. Schneider outlined a project for developing the tools to analyze qualitatively and quantitatively the questions of understanding the "why's" of good and bad predictions. A direct approach is to deduce the roles of

various processes which force the time mean flow on a case study basis. Practical approaches are linear and non-linear model simulation and analysis.

P. Sellers described a biosphere model which, among other factors, accounted for the effects of vegetation drag, reflectance, and Bowen ratio on the surface boundary layer.

In the next presentation of this session G. Ohring discussed satellite data for initial and boundary conditions and for validation of forecasts. Satellite observations can provide global coverage of vertical temperature structure, cloud tracked winds, relative humidity measurements, SST, outgoing long wave radiation (OLR), albedo, cloudiness, precipitation (indirectly from OLR and directly from microwave), snow cover, sea ice, skin temperature, and sea ice. Recommendations were solicited on where NESDIS ought to place priorities in providing data for DERF experiments. Soil moisture is one area where more information is required but is especially difficult to obtain.

The final presentation of this session was by R. Kistler on the "nitty gritty" of computer requirements, archival and data storage considerations, internal vs externally derived fields and parameters, post processing, etc. Execution of a 30 day run requires about 7 hours of CPU on the Cyber 205 for the current model. Upgrading the physics and including diagnostic quantities will increase this requirement by an as yet unspecified amount. The primary challenges relative to postprocessing are what quantities, their frequency of output, resolution and format.The pros and cons of disk vs tape storage were discussed. Tapes are portable but have limited capacity, are fragile, cumbersome, and data is sequential. Disk storage has fast random access, is reliable, and has enlarged capacity, but is not easily portable. In summary there are many logistical

questions that must be resolved in setting up, running, archiving, disseminating the output, etc. before the comprehensive testing can begin. Finally, a key element in regard to computer resources is the expected acquisition of a second Cyber 205 in early 1987.

4.4 SESSION III: Topics in Diagnosis, Evaluation, and Interpretation

The first paper of this sesion was presented by D. Williamson (R. Daley-coauthor) on the climate drift in the NCAR Community Climate Model (CCM). Specifically, focus was upon determining how the terms in the temperature tendency equation related to the drift towards or away from the models climatology. Among the interesting results was that some processes behaved very differently over land as opposed to over oceanic regions. For example, in the lowest model layer the adiabatic tendency term drove the ensemble of forecasts away from the model climate over oceanic points and towards it over land areas. The sensible heat flux was the only term driving the ensemble toward climate at the surface over the oceans. Williamson expressed the hope that the techniques used here to establish the mechanisms related to the drift toward the model climate can be applied to the same problem relative to the drift away from the climatology of the real atmosphere.

In the next paper D. Baumhefner addressed the question of estimating realistic values of predictability error growth from twin-pair numerical integratons. The NCAR model was used to obtain these estimates from an ensemble of forecasts based on the FGGE SOP1 data set. The resulting error growths proved to be insensitive to the initial size of the error, which ranged from 1 to 20 RMS in the 500 mb geopotential. The typical rate of growth was doubling at the 2-3 day time frame for errors in the 20-40m range. This agreed quite well with other results obtained at ECMWF and

GLAS using real atmospheric initial states as well. If the consensus growth rate of 2.5-day doubling is applied to a realistic initial error of 12m RMS, the errors grow as large as climatological variability in 8 days and become totally uncorrelated by 16 days. These limits of predictability pertain to Nothern Hemisphere wintertime situations of unfiltered patterns in the middle troposphere.

In the next presentation R. Chervin reported upon the influence of boundary conditions on predictability of the time averaged state of the atmosphere. It was noted that potential predictability exists at extended ranges only to the extent that the variance associated with boundary forcing contributes to the total, since the variance related to internal dynamics is unpredictable. Comparisons of real atmospheric variance with that of CCM runs from internal dynamics alone indicate that significant potential predictability exists only in tropical regions. Additional results from parallel multi-year integrations with and without anomalous global ocean surface temperatures were shown to be consistent with this conclusion.

J. Tribbia presented results relating to the predictability of time averages. For time averages up to 20-30 days the NCAR CCM does a respectable job of reproducing real atmospheric transient variability, a necessary attribute for perturbation experiments yielding reliable estimates of predictability. The results of such experiments indicate that on average 10-day running means lose skill at day 8 (i.e. 8-18 day mean is the last with any skill), 20 day means are not skilful beyond day 5, and 30 day means have no skill at all.

M. Blackmon shared several insights from studies of SST anomaly experiments with GCM's. He cited the importance of eddy-mean flow interaction and the position of the the mid-latitude jet to anomaly propagation

from the tropics. He also noted the importance of barotropic energy conversions and their strong model dependency. It was noted that the NCAR CCM was relatively insensitive to SST anomalies in mid latitudes. Complementing the results presented by Chervin, Blackmon noted that 80-85% of the variance in the CCM is related to internal dynamical processes. This led to an expression of pessimism on the prospects of skilful extended range prediction.

J. Walsh followed with a review of the role of SST, snow cover, and soil moisture variability in extended dynamical prediction. Of the three, SST is the most persistent of surface boundary influences, and soil moisture the least. Each affects local climate to some extent, but the effects outside the immediate area of influence (especially with regard to snow cover and soil moisture) is not certain and requires further study. Model simulations, which require a high level of complexity to depict accurately the interaction between boundary conditions and atmosphere, often give conflicting indications. A major problem is lack of soil moisture data.

A lively discussion followed the six presentations above. The importance of understanding model behavior relative to observations was repeatedly stressed. An oft cited example was the variance and climatology of real vs model atmospheres. Questions of this sort are critical in addressing the problem of the model dependency of results, e.g. the relatively pessimistic vs optimistic results of predictability experiments with the NCAR CCM and GLAS models, respectively.

The next formal presentation was by J. Roads on forecasts of time averages and temporal variations in predictability. In this study Lorenz's rms error model was applied to estimate the error growth of forecasts of time averages. Conclusions were; i) present day NWP model forecasts have "useful" skill to five days, whereas time averages extend utility to 10

days, ii) averages are improved by filtering, and iii) forecast skill varies with time of year, location, and the synoptic situation.

In the next talk E. Kalnay discussed experiments in LAF based upon operational forecasts from ECMWF. Conclusions were i) the winter hemisphere dominates the statistics, ii) LAF outperforms the straight (unweighted) average of all available dynamical forecasts, as well as the latest available dynamical prediction, iii) LAF shows no improvement at three days, but there is marked improvement at days 5 and 7, and iv) the spread of the LAF ensemble was a good a priori estimator of skill.

E. Epstein presented an outline of his studies on the statistical correction of spectral forecasts (NMC model). Conclusions indicate that simple regression applied to spectral coefficients is an economical and reliable means of improving the forecasts of standard level height fields. An a priori assumption that the longest waves contained the greatest predictability was substantiated.

The next presentation by H. van den Dool first examined aspects of the low frequency variability of the NCAR CCM and observations. In the context of the assumption that this variability (at least in the real world) is due to internal dynamics and interaction with the boundary, conclusions were that i) either the CCM is too insensitive, or ii) the extra variance associated with ENSO events, for example, is not additive, but competetive, and iii) perhaps the CCM is right and the observations are biased by data problems.

Another aspect of van den Dool's study was the assessment of whether forecast skill increased with time averaging relative to individual day forecasts. Conclusions were; i) time averaging may improve skill at long lead times, but only if the daily forecasts have non zero skill, ii) time averaging increases the signal to noise ratio, but a signal must first

exist, iii) time averaging makes it easier to demonstrate that skill differs significantly from zero providing that there is skill in the instantaneous forecasts, and iv) applying time averaging over a-priori determined lengths is useful only if we know the skill loading of the instantaneous flow.

In the final paper of this session F. Baer discussed observed characteristic structure evaluation of forecast quality. Using the FGGE data set, Baer's objectives were to examine the spectral characteristics of fields in the context of developing a tool for diagnosing model behavior.

Among the remarks in the open discussion at the end of this session were i) if there is skill in the instantaneous values, it makes sense to time average. Beyond the point of skill in the instantaneous fields, averaging only adds useless information, ii) perhaps we have not examined enough the question of why a forecast is not particulary good, and iii) even if the often used anomaly correlation drops to below 0.5, there still may be some useful skill remaining. Beyond this (or some other somewhat arbitrarily selected skill level), one should start time averaging.

4.5 SESSION IV: Summary and Discussion

This session, judged the most important of the workshop, considered i) where do we stand, ii) issues in concept, design, execution and evaluation of DERF experiments, iii) consultative/collaborative efforts, and iv) coordination.

Co-chairperson W. Bonner invited debate on the practical questions of experimental design, collaboration, and data handling after discussion led by co-chairperson J. Shukla on the scientific and conceptual issues. As it turned out the distinctions between practical and scientific questions were often difficult to separate.

The first issue was the question of whether, in fact, there exists a scientific basis for DERF. The optimists were asked to find the most realistic statement they would concede and the pessimists to identify the most optimistic statement they would accept. There was no direct response to this query, though it was clear from remarks here and earlier that varied opinions existed on the prospects for DERF. It was suggested that the question really ought to be,can we develop tools which will be useful for Gilman's forecasters? After some discussion on the relative merits of good individual forecasts vs. ensembles, it was generally agreed that dynamical forecasts and statistical procedures could enhance skill when used together.

The next question considered was the tradeoff between model development and accumulation of a homogeneous forecast sample. The comments were highly varied, but there was strong sentiment in favor of homogeneous statistics, i.e., a stable model. This being the case an operational center like NMC can be involved only during periods of relative model stability or when computer resources permit running offline with a fixed, non-operational version of the model. The importance of NMC documenting and circulating information on model status was stressed. The suggestion was made of running the NMC model for DERF experimentation at another computer facility. An argument against this is the presence of experienced forecasters at NMC who would be involved in the crucial evaluation of forecast utility. In response to questions of whether a simpler version of the operational model could be used, the general opinion seemed to be that using the more sophisticated model is desirable. Simpler models might be useful in enhancing the size of the statistical sample, if it can be shown that their characteristics and systematic errors correlate with those of the more advanced model.

On the question of choosing a large number of individual cases or a smaller number of cases of ensemble runs, there was a mixture of opinion. Several people noted that skill at the extended range is generally low but variable and ,therefore, ensembles must be used. Others indicated that it is useless to examine higher order moments of an ensemble if the mean is bad or if none of the ensemble members are skilful. The comfortable middle ground was that both many cases and ensembles are required to evaluate both skill and the characteristics of ensembles. It was noted that ensembles can be sets of different models, not just ensembles of differing initial conditions. The suggestion was made several times this session for GFDL, NMC/COLAI, ECMWF, UKMO, and NCAR to consider parallel experiments.

Several practical issues enumerated by Bonner were:

1) What constitutes an interesting model?

The MRF would qualify when its climatology was the best possible and its climate drift the slowest possible. That point, as illustrated by the performance results presented by G. White earlier, has probably not yet been attained. Improvements in the model are necessary before proceeding with the comprehensive DERF experiment at NMC. Again, use of simpler models was suggested to augment runs with the more sophisticated versions. The possibility was raised that the simpler models might have a better climatology, but this needs to be assessed through intercomparison experiments. On the question of to what range should models be integrated in DERF experiments, there was some sentiment for restricting it to 20 days, but the concensus was for extending the runs to at least 30 days.

2) How do you extract a useful signal from DERF?

Among the possibilities are using time averages, LAF, defining confidence limits, more relevant and/or appropriate measures of skill, and

package (MRF-86) was used in the January run. Additional 30 day forecasts with MRF-86 (with or without some relatively minor variations) will be run from 1200 GMT 16 February and March 1986. These and subsequent experiments are in accord with the outline presented in Section 3.2. The MRF-86 physics, unlike that of MRF-85, includes shallow convection,convective heating extending above 300 mb, and stability dependent diffusion. Additionally, MRF-86 has 18 unequally spaced vertical layers, with the enhanced resolution primarily at lower levels. The new physics and enhanced vertical resolution is expected to become operational in April 1986.

Evaluation of the first two 30 day experimental forecasts has just begun. As an overview, Figures 1 and 2 present the observed and forecast 30 day mean 500mb height/anomaly charts for the December and January cases, respectively. Also shown are the fields of standard deviation about the 30 day mean height fields. In both cases the forecasts simulate most of the principal features seen in the analyses. There are, however, some large errors, for example, in the mean ridge along the west coast of North America in both cases and in the mean trough over Europe in the January run. With the exception of the overprediction of the trough in the Gulf of Alaska in both periods, the model tends to underpredict the amplitude of systems. Also note the preponderance of negative anomalies in the forecasts at higher latitudes, reflecting a fairly strong cold bias in extratropical latitudes. The "new physics" actually seems to increase the bias, though a similar bias in tropical regions (not shown) is reduced with the MRF-86 run. Care must be exercised, however, in interpreting differences in errors between these two cases as being only the result of model changes, because the meteorology is also different.

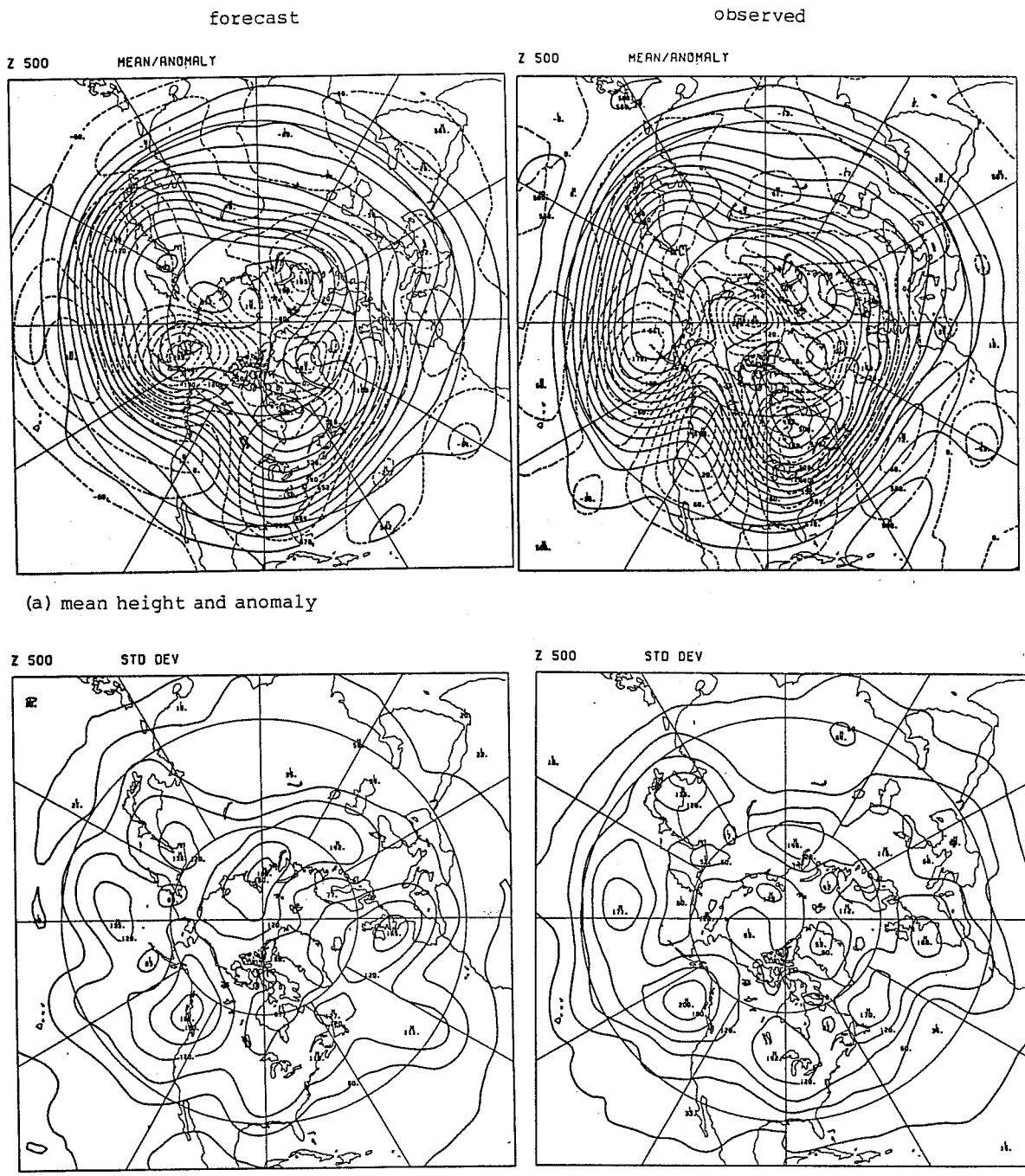
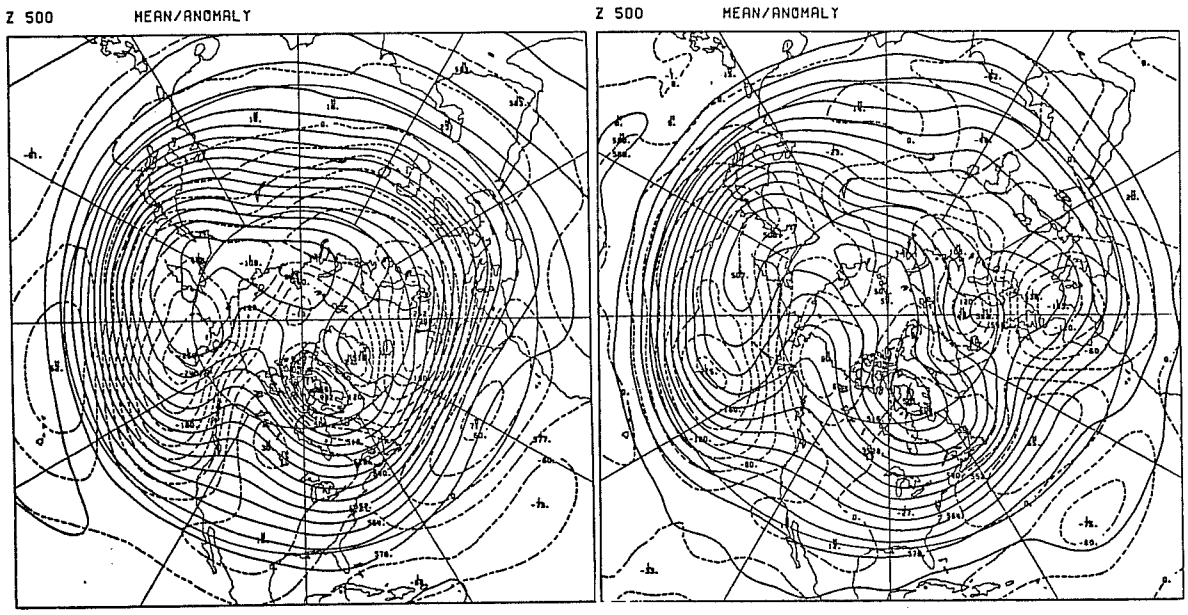
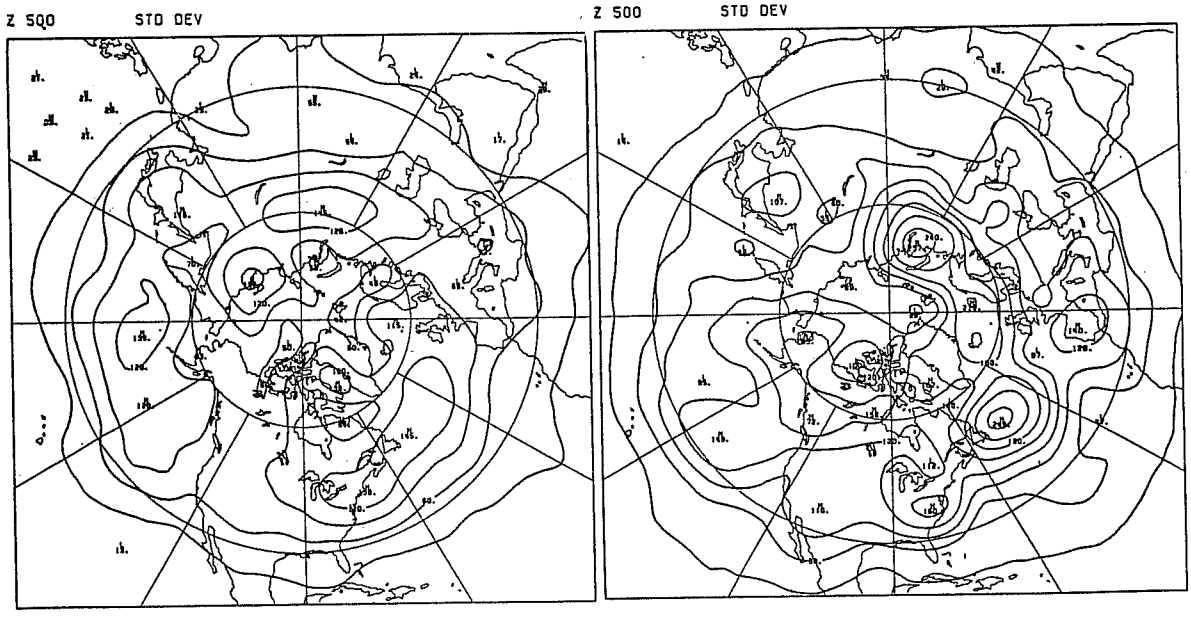


Figure 1. 30-day mean 500 mb mean height and anomaly (a) and standard deviation (b) for December 1985 case.

forecast observed



(a) mean height and anomaly



(b) standard deviation

Figure 2. Same as Figure 1, except for January 1986 case.

Note from the fields of standard deviations that the forecasts underspecify the variability in the 500 mb height field. This is especially so over the Western Atlantic and Northern Siberia ahead of the mean troughs in the January case. Inspection of the 3 ten day mean charts comprising each of the 30 day periods shows that the forecasts fail to identify some meteorologically significant changes occurring on that time scale. The anomaly correlations for the individual 10 day means and the thirty day mean charts are shown in Table 1. The skill, especially in the January case, drops off markedly between the first and second 10 days. The thirty day scores approach values generally ascribed as indicating some degree of utilitarian value. The relatively good scores in the thirty day means reflect the higher values in the first 10 days plus a tendency for errors to average out during the course of the full forecast period.

Additional results as available will be presented at the Workshop.

TABLE 1. 500 MB ANOMALY CORRELATION SCORES

		DEC	JAN
DAYS	1-10	.76	.70
	11-20	.66	.34
	21-30	.44	.10
	1-30	.45	.44

APPENDIX A. Participants at NCAR DERF Meeting

R. Kistler, NMC; J. Shukla, U/Maryland; S. Tibaldi, ECMWF; U. Cubash, ECMWF; T. Palmer, British Met. Office; D. Baumhefner, NCAR; M. Blackmon, NCAR; R. Chervin, NCAR; D. Williamson, NCAR; G. Boer, Canadian Climate Centre; W. Bourke, Australia; K. Miyakoda, GSFD/NOAA; and H. Von Storch, Germany.

APPENDIX B. Participants at National Meteorological Center DERF Workshop

Dr. David Williamson, NCAR
Dr. Roger Daley, Canadian Climatology Center
Dr. David Baumhefner, NCAR
Dr. Joseph Tribbia, NCAR
Dr. David Rodenhuis, Climate Analysis Center/NWS
Dr. Donald L. Gilman, Climate Analysis Center/NWS
Dr. Eugenia Kalnay, NASA/GSFC
Dr. Robert Livezey, NASA/GSFC
Dr. Andrew Gilchrist, British Met. Office
Dr. Anthony Hollingsworth, European Center for Medium Range Weather Forecasts
Prof. Ferdinand Baer, U/Maryland
Prof. John Walsh, U/Illinois
Prof. Richard Sommerville, Scripps Inst. of Oceanography
Dr. Huug van den Dool, U/Maryland
Dr. Kiko Miyakoda, GFDL/NOAA
Dr. Joseph Gerrity, National Meteorological Center/NWS
Dr. Edward Epstein, Climate Analysis Center/NWS
Dr. George Ohring, NESDIS/NOAA
Dr. Maurice Blackmon, NCAR
Dr. William L. Sprigg, National Climate Program Office/NOAA
Dr. Hassan Virji, National Science Foundation
Dr. Ed Schnieder, U/Maryland
Prof. Jim Kinter, U/Maryland
Dr. P. Sellers, U/Maryland
Dr. Larry Marx, U/Maryland
Dr. M. Fennesy, U/Maryland
Dr. Bert Katz, U/Maryland
Jim Miller/Climate Analysis Center/NWS
Dr. Ken Bergman, Climate Analysis Center/NWS
Dr. Vern Kousky, Climate Analysis Center/NWS
Dr. Robert Chervin/NCAR
Dr. John Brown, National Meteorological Center/NWS
Dr. John Roads, Scripps Inst. of Oceanography
Prof. J. Shukla, U/Maryland
Prof. John M. Wallace, U/Washington
Glenn White, National Meteorological Center/NWS