DERF Forecasts of Thickness

John O. Roads
Scripps Institution of Oceanography
UCSD, A-024
La Jolla, CA 92093

1. Introduction

In 1985, the National Meteorological Center (NMC) began a concerted effort to evaluate whether numerical predictions might be useful for making extended range predictions on the order of a month. After some initial efforts at evaluating the usefulness of numerical predictions for making extended-range forecasts the NMC global spectral model used for medium-range predictions (1-10 days) was run out to 30 days starting from 0000 UTC of each day from 14 December 1986 to 31 March 1987 (see Tracton et al. 1987, 1988a, 1988b, 1988c). In this paper, an initial evaluation of these forecasts is undertaken.

The original goal of this research was to evaluate the NMC DERF forecasts of surface temperature since that is ultimately what we wish to be able to forecast. Since presumably the temperature has a fairly large vertical structure, it should be possible to determine some of the behavior of the surface temperature from the 850 temperature field. Moreover, since the 850 temperature field is highly correlated with the lower tropospheric thickness (700 mb-1000mb height), since thickness forecasts are available from much longer forecast data sets and since thickness has less of a systematic bias, it is advantageous to examine the thickness field as a proxy for forecasting surface temperature.

2. Climatology

An arbitrary decision was made to approximate the climatology by a quadratic least squares fit to the 138 observations and to then measure the skill with this mean removed. Figure 1 shows the time series of the global mean spectral component for the thickness, 700 mb height, 1000 mb height and 850 mb temperature for the observations along with this estimated quadratic climatology. Note the model bias; the bias was removed by removing the time averages of all fields.

An arbitrary truncation was used for the analysis. Although R30 was available, only R12 was evaluated. The principal reason for this choice is historical, a much larger data set of medium range forecasts is available at this resolution. However, the anomalous thickness variance explained by this truncation is greater than 92 percent.

3. Analysis Method

In the past it has been common to simply evaluate the ensemble correlation, but we now realize that we need to know not only the average correlation but also the temporal variation in this correlation. We would also like to know the time averaged value of the individual correlations. It is therefore advantageous to transform the correlations by the Fisher z transform, which is unbounded and has a normal distribution, and then take the time average of this variable and then transform back to ρ. By this method, we can establish bounds on the pattern correlation distribution. In particular, we can evaluate the standard deviation of the ρ distribution, given by the transformed values of $z+\delta$. We would also like to know how significant, our estimate of the ensemble mean correlation is. If all of the z's were independent, then the estimated standard deviation of the ensemble mean would be simply the standard deviation of the z's divided by the square root of the number of correlations taken. However, the z's are correlated from day to day since the forecasts and observations are correlated from day to day and therefore each correlation is not independent. Leith (1973) suggested a simple model for this autocorrelation, namely a first order process, and with this assumption, it is possible to evaluate an independent time scale from which we can then obtain an estimate for the ensemble mean standard deviation. To summarize, pattern correlations are averaged for each forecast and transformed to z. The z, z^2, and $z(1)z(0)$ are then evaluated to give the bounds of the distribution of ρ as well as the estimated bounds of the distribution of the average ρ.

4. Forecasts of Daily Values

Figure 2 shows the correlation for forecasts of daily events for the numerical model in comparison to a simple persistence model for the US sector, the northern hemisphere and the global domain. The solid lines show the average correlation and the dashed lines the standard deviation of individual variations. The dotted lines give the estimated standard deviation for the average. Three sectors are evaluated: a US sector (which is an area over the continental US), the northern hemisphere, and the global domain. There is not too much difference in skill between the various sectors. Skill decreases toward zero at about the same rate and near zero tends to saturate rather than asymtoting toward 0. Depending upon the level of significance needed, the skill can be indistinguishable from zero at about 11 or 12 days with the most rapid decrease occurring for the US sector. Persistence approaches much more rapidly the zero level; at only a few days skill is insignificant. Persistence seems to show a significant negative correlation for a couple of weeks, which may eventually

CLIMATOLOGY

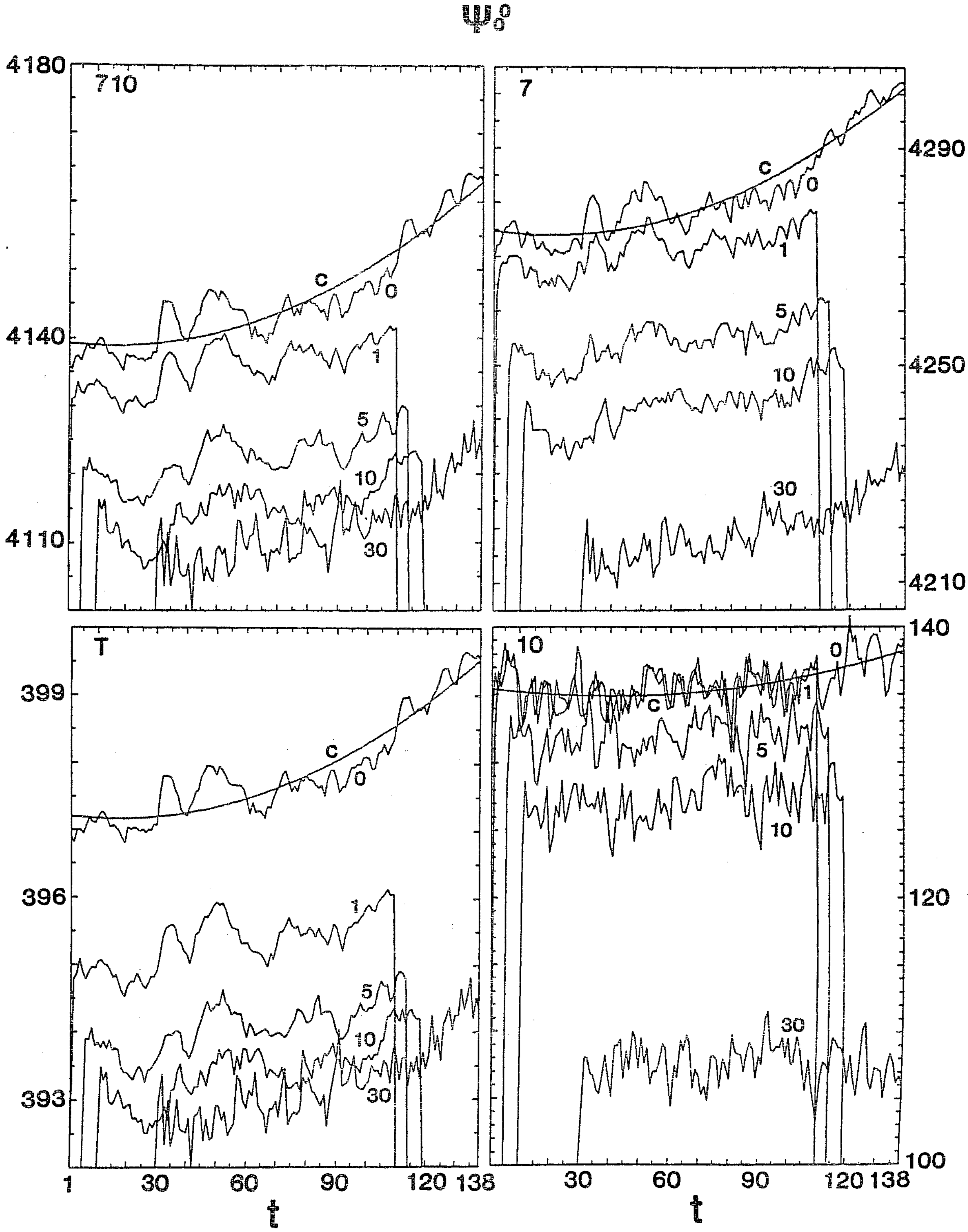

Figure 1: Quadratic climatology, C, for the ψ_0^0 component, the global average times $\sqrt{2}$, for the 850 mb temperature, (T), 850 thickness, (710), the 700 mb height (7), and the 1000 mb height (10). Also shown are the daily observations, 0, and 1, 5, 10, and 30 day forecasts for this component. ψ_0^0 is shown as a function of verifying time with 1 denoting 0 Z December 14, 1986 and 138 denoted 0 Z April 30, 1987.

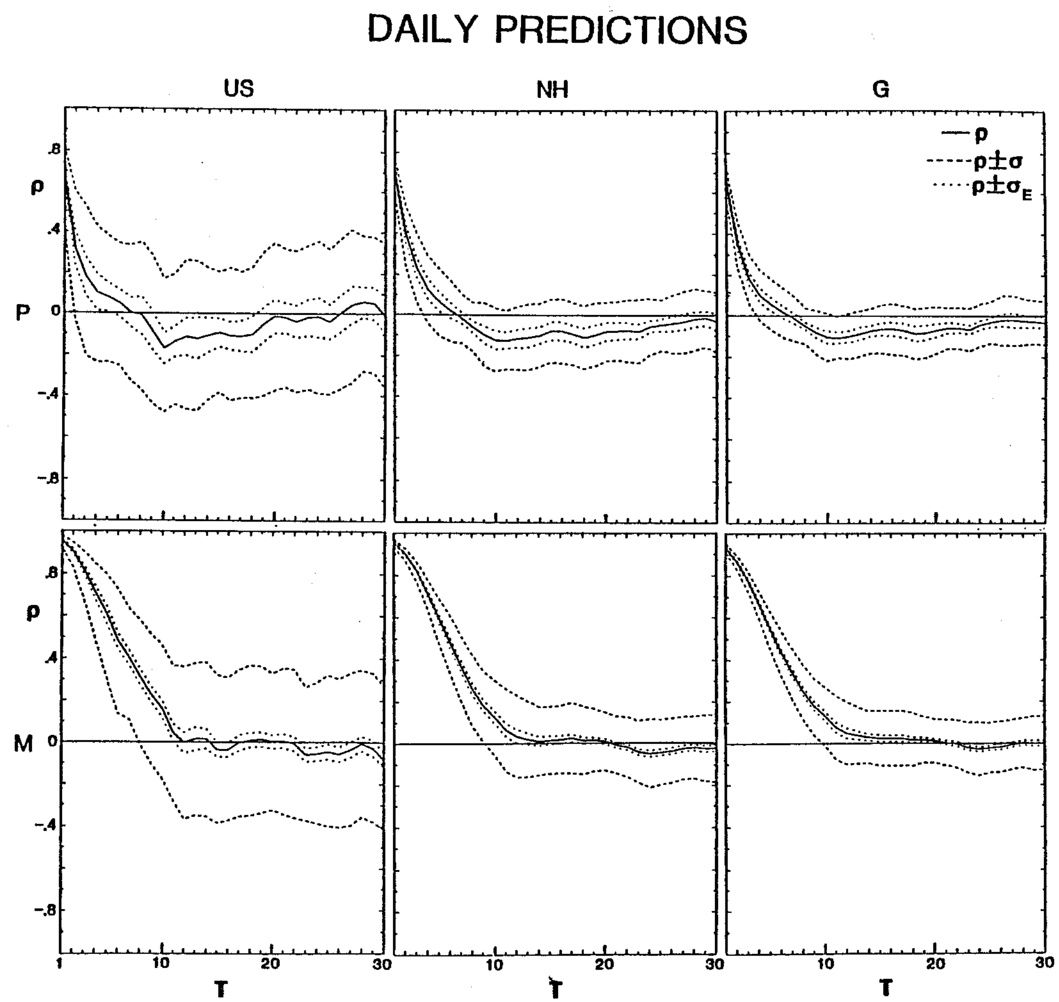

Figure 2: Daily ensemble pattern correlations for persistence, P, and the numerical model, M, for the U.S. sector, Northern Hemisphere and global regions as a function of lag time, τ. The solid line is the ensemble average of the individual z's transformed to ρ. The dashed lines give the standard deviation of daily variations and the dotted lines give the estimate of the standard deviation of the ensemble means.

be of some use in the extended range. It must be concluded from this figure that it is probably not possible to currently predict instantaneous events beyond two weeks on the average with current dynamical models. Of course, as the distributions show it is perfectly possible to sometimes have some spectacular forecasts, especially over the US sector and presumably other limited regions. In fact, it is also perfectly possible to have some spectacular persistence forecasts, the width of the distribution is about the same as for the numerical model forecast.

5. Forecasts of Time Averages

Figure 3 shows the skill for a time averaged forecast for the same sectors. An a priori hypothesis was made here to limit the time averages to just the first 10 days if the desired time average beyond 10 days was desired. Again, the ulterior motive was to be able to compare the much longer time series of archived medium range forecasts applied to the monthly forecast problem, but in fact forecasts of time averages greater than 10 days are best made by considering the average of the first 10 days or so of the forecast, at least for the data and analysis of skill considered here. (See also Roads, 1986, 1987, 1988a). For example, figure 4 shows the appropriate averaging time for individual forecasts of 30 day averages along with the corresponding correlations for the optimal forecast time and correlations for the 10 day forecast forecast proxy and the 30 day forecast average. There is an extremely wide variation in this forecast time, ranging from the initial day to the entire 30 day forecast average. However, an optimum time of the entire 30 days is an extremely rare event just as the initial day is an extremely rare event. An envelope of 10 days seems to capture most of the variance in the time series. In fact, for the (global, NH, and U.S.) sectors, the 10 day forecast average proxy has a correlation of (.91, .86., .83) with the optimal forecast whereas the 30 day forecast average is correlated at (.65, .60, .59) with the optimal forecast; persistence is correlated at (*, 62., .70) with the optimal forecast. Suffice it to say that it seems to be more advantageous to deal with the shorter 10 day forecast average proxy than with the entire 30 day forecast average which has more extreme variability, less average skill, and smaller correlations with the optimum forecast variability.

These proxy numerical time average forecasts are compared to persistence forecasts in which persistence of the initial day is used as a proxy for the initial day. This type of persistence is certainly convenient to use for the comparison and some studies have indicated that the inclusion of previous days will be worse than simply using the initial day. In any event, as shown in Figure 3, for forecasts of time averages the skill

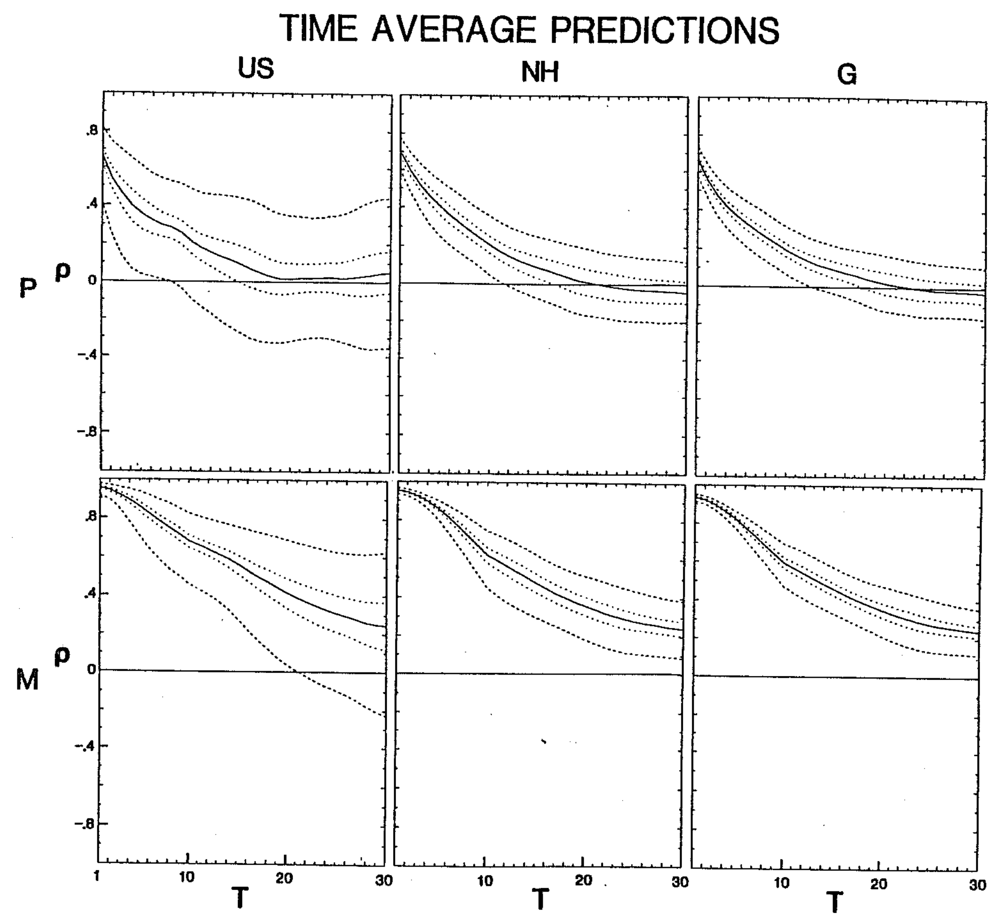

Figure 3: Same as Figure 2 except that this figure shows the correlations for forecasts of time averages (averaged from one day to τ days). When forecasts beyond 10 days are desired then only the first 10 days of the numerical forecast are used as a proxy.

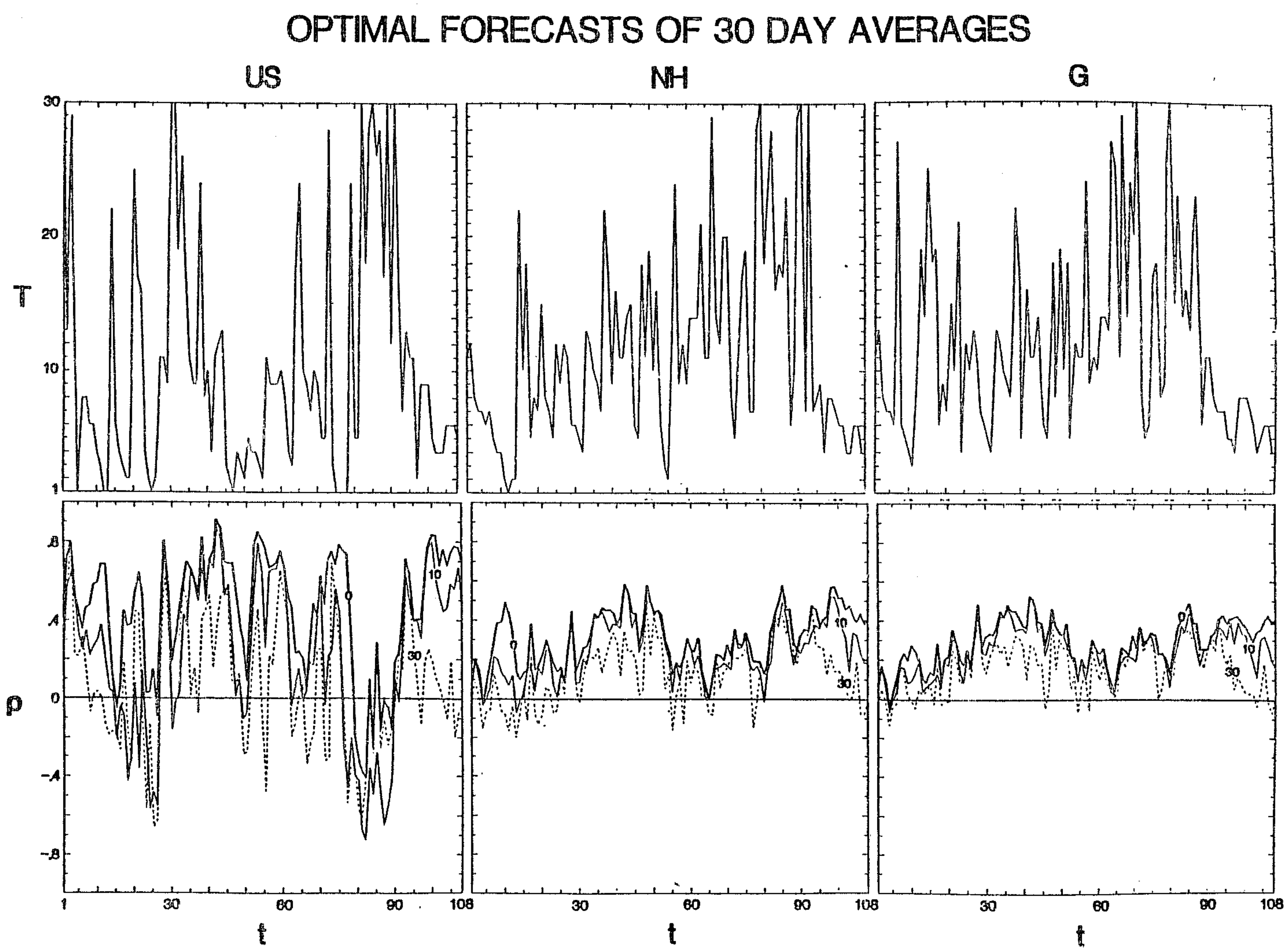

Figure 4: Daily correlations of 30 day averages with the 10 day forecast average proxy, thin line, and 30 day forecast average, dashed line, and the optimum forecast average, solid line, for the U.S. Northern Hemisphere and global sectors. The T of the optimum forecasts are shown in the upper panels.

remains above zero for the entire 30 days; in contrast persistence is insignificantly beyond about 2 weeks. The level of skill for the numerical proxy is small but appears to be significant. Moreover the distribution indicates that some really spectacular forecasts were made. About 1/6 of the time the correlation was probably greater than .6 for the US sector. Persistence also shows the potential for some spectacular forecasts although less likely on the average than the numerical model due to the overall lower level of skill.

6. Temporal Variations in Skill

On the average, the skill of a 30 day forecast is quite low. On the other hand the distribution of skill is fairly large. For example, Figure 4 shows the skill of each numerical forecast as well as the skill of a persistence forecast. For the U.S. sector, there are roughly 3 episodes of high skill, at the beginning, toward the middle and toward the end of the ensemble with lots of temporal structure in between. Persistence also has roughly the structure of the numerical forecasts with a lower overall mean. However, the correlation of the skills for the numerical model and persistence is fairly low, approximately .3. For the northern hemisphere and global domain the correlation is larger, approximately .6. However, even if there were a better connection between persistence and numerical model skill, it still would not help us to predict when the skill is likely to be large and when it is likely to be small, since we only know after the fact whether persistence was a good choice. We need other indicators.

To put more of a perspective on why it would be advantageous to have this knowledge, Figure 5a shows the average of the 1/3 highest skill numerical forecasts, and the 1/3 lowest skill forecasts in comparison to the normal forecast skill. The composite was done in terms of the 30 day forecast but in fact this extended range skill also shows up in forecasts of shorter term averages. It also shows up in having higher skill persistence forecasts indicating the slight correspondence between the two forecast methods.

One way to predict this forecast variability is to determine the characteristics of the initial state, observed 30 day average and model state when the forecast is likely to be large and when the skill is likely to be small. Figure 6 shows the composites of high and low skill for the US sector. The composite was done by averaging the 1/3 high, normal and low cases for the lag forecasts of 30 day averages. There are clear differences between the high and low skill cases. For the high skill case the country is as a whole initially cool and above normal thickness exists off the west coast. During the course of the 30 day forecast, the initial warmth along the coast spreads

A POSTERIORI DISCRIMINATING FORECASTS

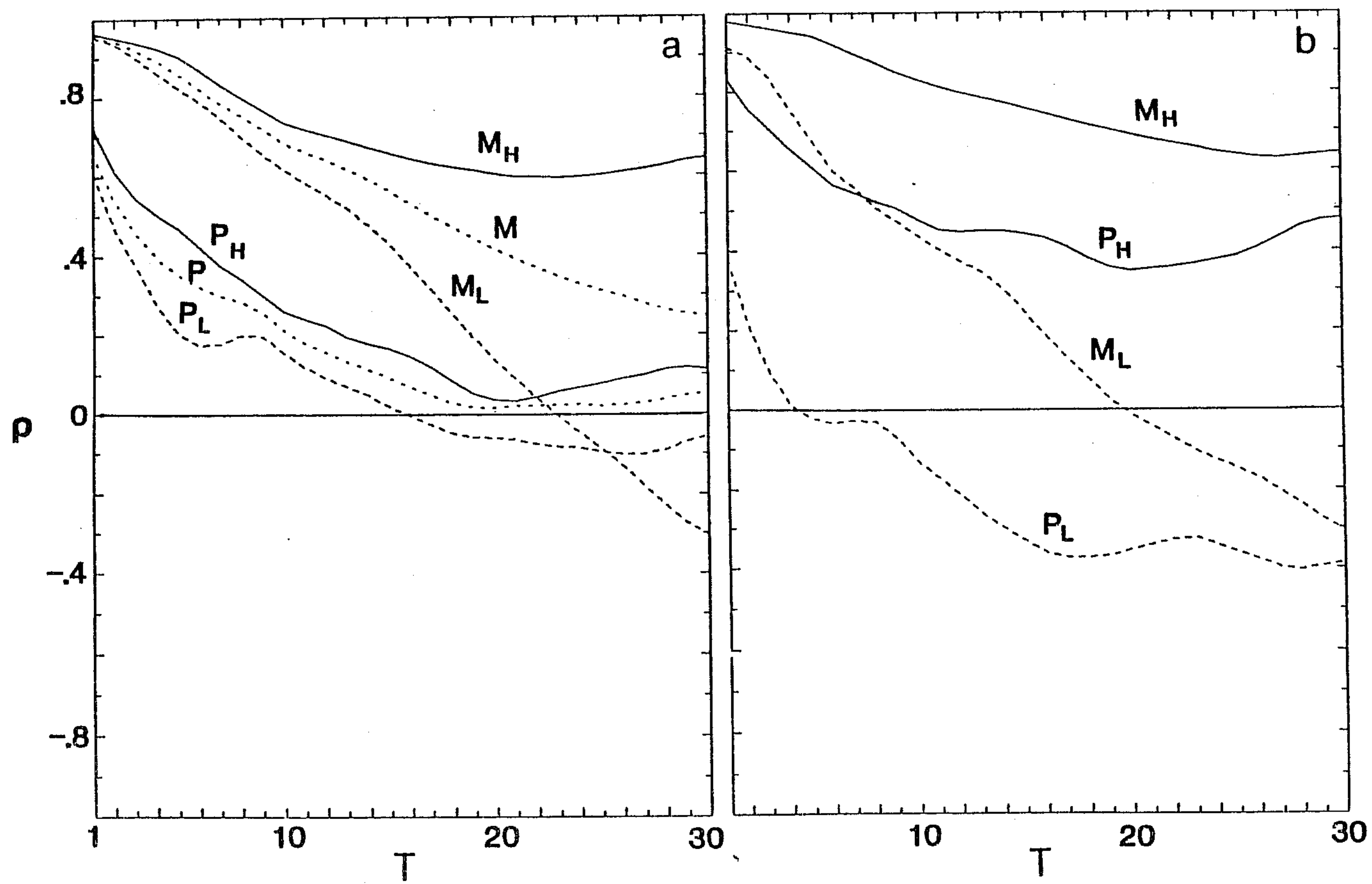

Figure 5: Skill of discriminating forecasts if perfect knowledge of above normal and below normal forecasts for 30 day averages were available. (5a) shows the skill of the model and persistence for the 1/3 of the time that the 30 day forecasts by the numerical model 30 day forecast proxy were high or low. (5b) shows the skill for the model and persistence forecasts for the 1/3 of the time that the numerical model or persistence forecasts were high or low for each forecast day.

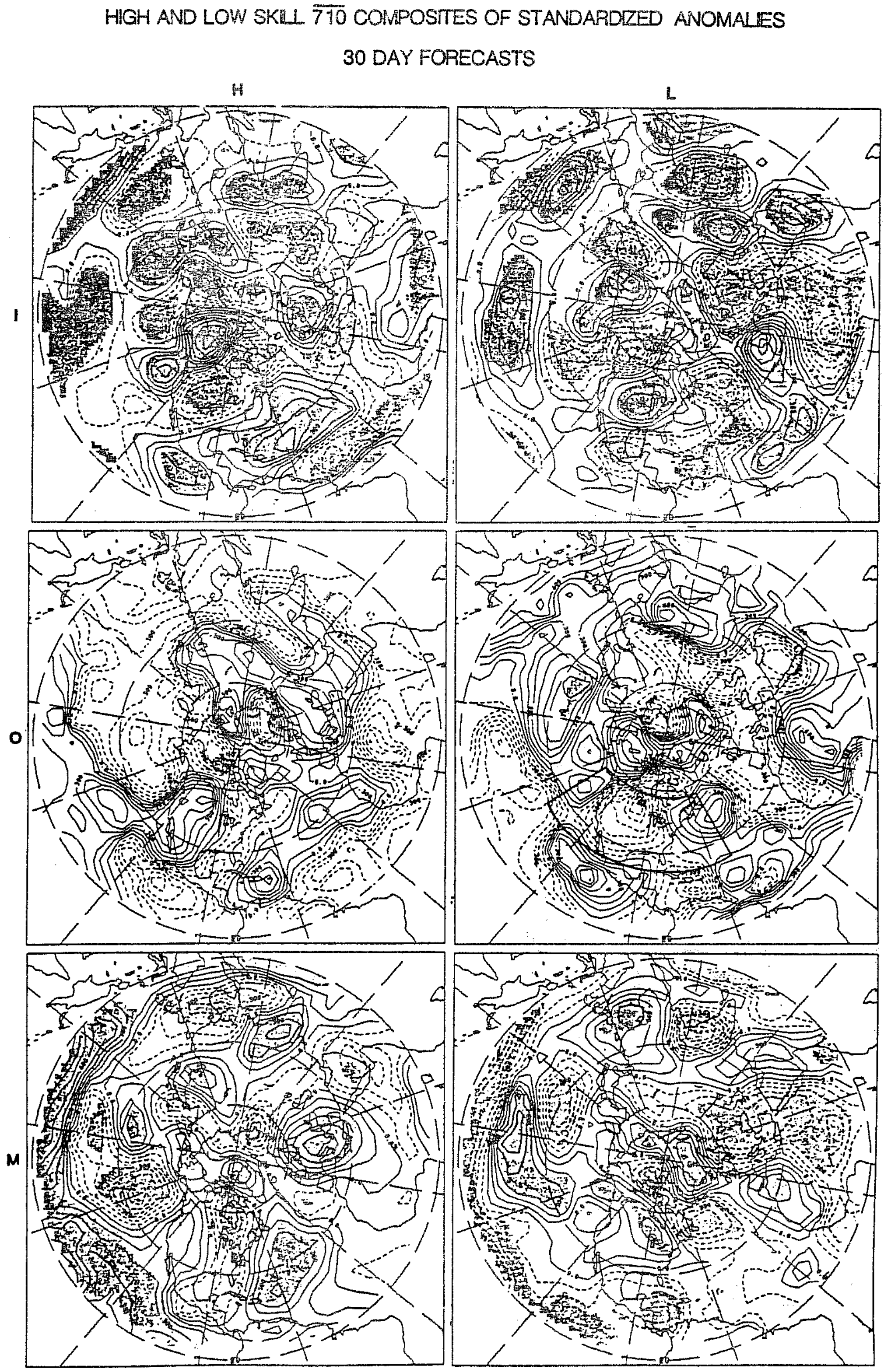

Figure 6: Physical space composites of the thickness field for the cases in which the 10 day forecast average, M, was highly correlated, H, or highly uncorrelated, L, with the observed 30 day average, O. I denotes the initial state. Shaded regions indicate those regions in which the composite was greater than the estimated ensemble standard deviation.

CORRELATIONS BETWEEN SKILL AND SKILL INDICES

FOR TIME AVERAGE FORECASTS

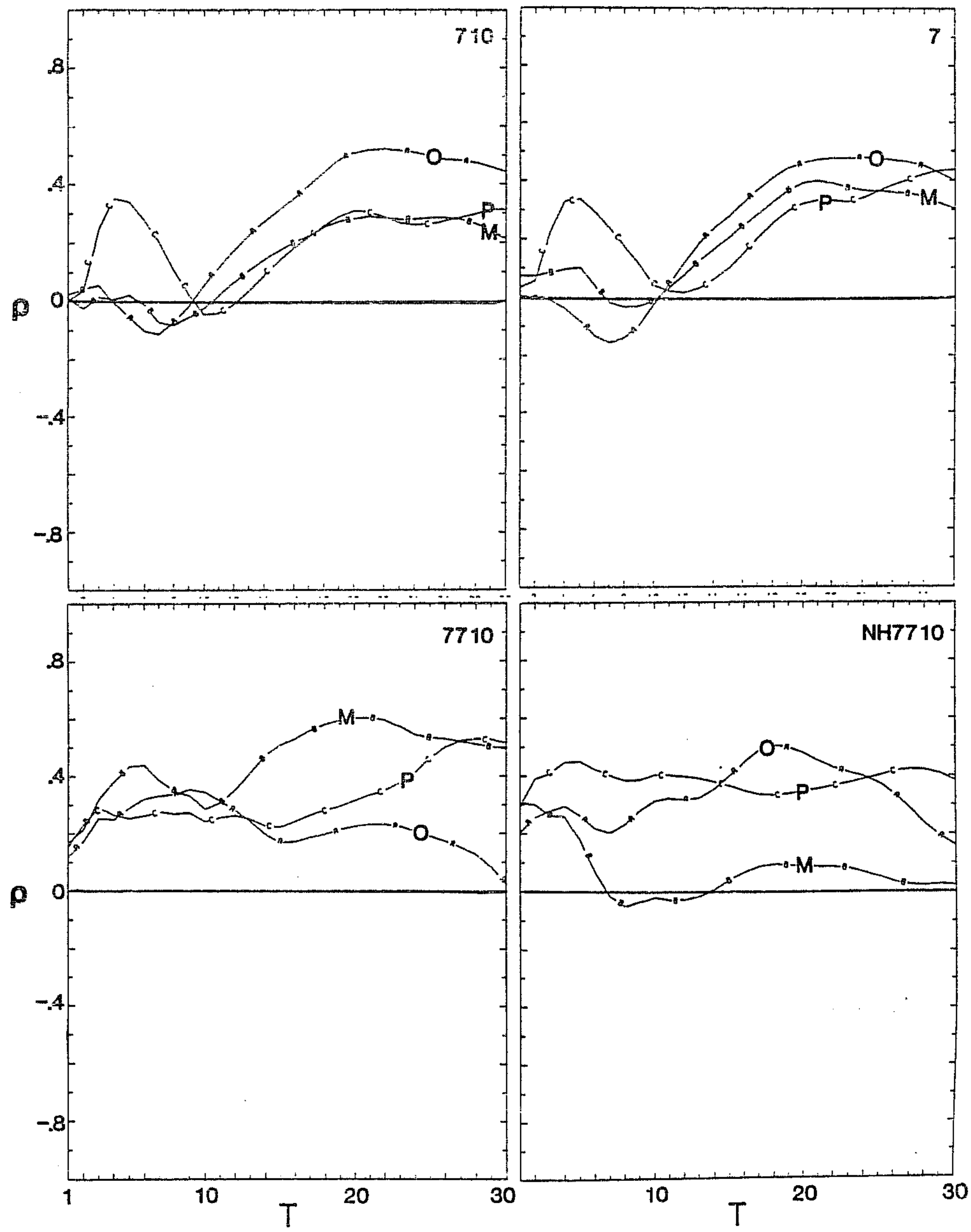

Figure 7: Correlation of the various skill indicators with the actual z of the 10 day forecast average proxy and observed 30 day average for the U.S. sector. P denotes the index constructed from the high skill composite initial state. M, the index from the high skill composite of the 10 day average forecast proxy, and O, the index constructed from the observed high skill 30 day average. 710 denotes the thickness high skill indices, the 700 mb height indices. 7710 denotes the correlations between the model skill and the barotropic index as measured by the correlation of the 700 mb height and thickness. The graph labeled NH7710 shows the correlation between the 7710 index and the model skill for the entire Northern Hemisphere sector.

inland to result in a warm west coast and cool east coast. The model has the ability to predict the slow variation by the 10 day forecast average proxy. In contrast, during the low skill episodes, the initial state has a warm interior and cool states along both coasts. During the course of these forecasts, the observed average state completely reverses the initial state. The model is unable to reverse the initial situation and therefore predicts a state similar to the initial state and almost the reverse of what occurs. Similar features occur at the 700 mb level.

Given these features, it is possible to construct a skill index. The skill index was taken here to be simply the correlation of the high or low composites with the actual state. This generates another time series of correlations which can then be correlated with the actual variability to determine how useful this skill index is for predicting the actual variability. The correlation of the skill index with the 1-30 forecast averages is shown in Figure 7. As may be seen, the composite is heavily tied to the forecast of 30 day averages. The greatest skill in making the prediction is found for the forecasts of 30 day averages.

Another hypothesis discussed in Roads (1985) is that the baroclinicity of the initial state may have something to do with the forecast skill. A barotropic index was constructed by correlating the 700 mb height and the thickness over the U.S. sector and the correlation of this skill indicator with the actual skill (Fig. 4) shows that this too provides a useful indicator of model variability. That is if the initial state is baroclinic then the forecast is likely to much worse than if the initial state is barotropic. Unfortunately this skill index only worked for the US sector, if the entire northern hemisphere domain is considered then it provides much less indication of the skill and if other sectors are considered then the skill indication is still lower. To summarize, the initial state baroclinicity as measured by the correlation between the thickness and 700 mb height is a useful indicator of forecast variability over the U.S. and the northern hemisphere for this data set; it is not as useful for other small northern hemisphere sectors.

7. Predictability

Lorenz (1982) suggested a predictability experiment that could be performed with the model to determine what might happen if the model were perfect and the atmosphere behaved like the assumed perfect model. That is instead of trying to verify the forecasts with observations, we can verify the forecasts with each other. In particular, we can verify the one day forecast with the two day forecast, the two day forecast with the 3 day forecast etc for each verifying day. The difference

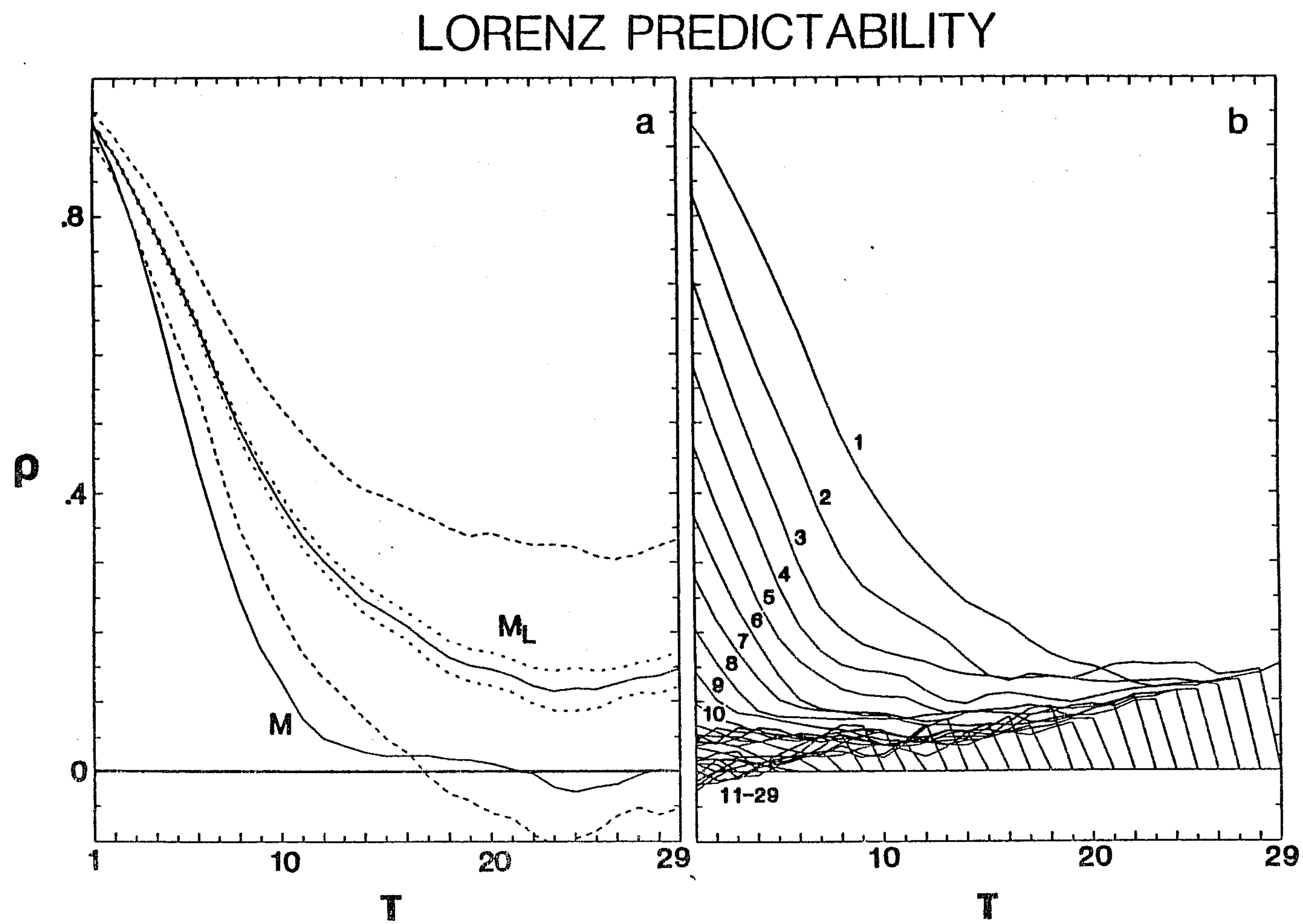

Figure 8: Thickness correlations for the Lorenz predictability experiment in comparison to the forecast experiment. (a) 1 day lag. The meaning of the dashed and dotted lines are described in Fig. 2. (b) 1-29 day lags. Here the ρ are ensemble correlations. The numbers refer to the lag of the initial state in days. The lag 1 ensemble pattern correlation is shown in Fig. 8a.

between the two runs is thus the result of amplification of the initial error between the one day forecast and the two day forecast verifying on the same day and so on. A fundamental problem with this procedure is that we are not sure that the forecast model is an accurate representation of nature; however this problem is common to all predictability experiments and we are certainly safer in performing this experiment with the NMC model than we would with the simpler models that have been used in the past to describe predictability error growth.

The results of this predictability experiment are shown in Figure 8. Note that unlike the forecast experiments, the predictability does not saturate at zero. There is now in essence an indication of predictability beyond the classical three weeks. If the optimum time average and lag of the forecast are used for the desired average and lag then forecasts of a 30 day average would have a skill now achieved for a 5 day instantaneous forecast or a 10 day average. Forecasts of lagged time averages would be made better than forecasts of the initial day of the lag. It would be advantageous to average the forecast over almost the entire period. In short, we might have to revamp our present feeling that most of the skill in making a forecast of a time average results from the skill of the initial part of the forecast.

8. Summary

To summarize, we are still a long way from understanding what the ultimate level of prediction skill is. We only have a better idea of what these models are presently capable of forecasting and what they might be capable of forecasting due to the DERF experiment.

Further details are given in Roads (1988b).

Acknowledgements

J. Alpert, R. Kistler, N. Maisel, and S. Tracton of NMC were extremely helpful to this research which was supported by CAC, NSF Grant ATM85-20540, and NOAA Grant NA86-AA-D-CP104. I also thank P. d'Ursin for typing the manuscript, M. Schultz for help with the figures, and D. Nghiem and E. Bainto for help with the programming.

References

Leith, C.E., 1973: The standard error of time-average estimates of climate means, *J. Appl. Meteor.*, **12**, 1066-1069.

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

Roads, J.O., 1985: Temporal variations in predictability, *J. Atmos. Sci.*, **42**, 884-903.

Roads, J.O., 1986: Forecasts of time averages with a numerical weather prediction model. *J. Atmos. Sci.*, **43**, 871-892.

Roads, J.O., 1987: Estimate of errors in lagged time average numerical weather prediction. *Tellus*, **39A**, 492-499.

Roads, J.O., 1988a: Predictability in the extended range. *J. Atmos. Sci.*, **44**, 3495-3527.

Roads, J.O., 1988b: Dynamical extended range forecast of the lower tropospheric thickness, *Mon. Wea. Rev.*, (submitted)

Tracton, M.S., 1987: Application of dynamic extended range forecasting (DERF) to the monthly forecast problem. Proceedings of the 12th Annual Climate Diagnostics Workshop, Salt Lake City, UT, October 12-15, 1987.

Tracton, M.S. and R. Kistler, 1988a: Activities and dynamic extended range forecasting (DERF) at the National Meteorological Center, Numerical Weather Prediction Conference, AMS, in press.

Tracton, M.S., E. Kalnay, W. Chen, K. Mo, G. White, and R. Kistler, 1988b: Dynamic extended range forecasts at NMC. European Centre for Medium Range Weather Forecasts Conference on Predictability in the Extended Range, May 11-14, 1988.

Tracton, M.S., E. Kalnay, W. Chen, K. Mo, G. White, and R. Kistler, 1988c: Dynamic extended range forecasts at NMC. To be submitted to *Mon. Wea. Rev.*