Tellus (1987), *39A*, 492–499

Estimate of errors in lagged time average numerical weather prediction

By JOHN O. ROADS, *Climate Research Group, Scripps Institution of Oceanography, UCSD, A-024, La Jolla, CA 92093, USA*

(Manuscript received 19 June; in final form 11 December 1986)

ABSTRACT

An idealized model for error growth is used to discuss numerical weather prediction model forecasts of time averages ranging from 0 to 30 days in duration at lags of 0 to 30 days. The idealized model allows a simple assessment of how the initial error, error growth rate and serial correlation, influence a forecast model's prediction of a time average. The simplified nature of the idealized model also allows an easy demonstration of how various filters applied to the raw numerical predictions can help to improve forecast skill.

1. Introduction

In a previous paper, Roads (1986, hereafter referred to as R) discussed the error growth inherent in a forecast of a time average by a numerical weather prediction (NWP) model. The time averages referred to in R were those that averaged the forecast from the beginning of the forecast period (day 0, i.e. the initial conditions) to the end of the forecast period (day T). This type of forecast is usually made when the desired forecast time is much larger than characteristic predictability or forecast times of individual events (say 1–3 weeks). For example, the National Meteorological Center (NMC) provides monthly and seasonal forecasts at the beginning of the forecast month or season.

R showed that the best forecasts of these zero-lag time-averages (forecasts starting at day 0 and averaged to day T) were provided by properly filtering (i.e., properly weighting) the raw numerical forecasts. One filter used multiple regression to weight the individual days that go into a time average. Another filter simply ignored the daily predictions once the forecasts had more noise than predictive skill. This cutoff was near to the days where the forecasts of individual days had correlations of approximately 50% or less with the observed data.

Another type of forecast of a time-average currently provided by NMC is a forecast of a 5 day mean 5 days into the future (average of days 6–10), which is a lagged time average. Questions naturally arise then as to what the skill is and what the filters are for this more general case involving arbitrary time averages at arbitrary lags. The purpose of this note therefore, is to extend the results obtained by R using a very idealized model; in addition to the special case of arbitrary time averages at zero lag, the more general case involving arbitrary time averages at arbitrary lags is also treated here.

2. Raw numerical prediction

It will be assumed here that we are discussing predictability of time averages within the context of a perfect model. That is, given an observed variable, ψ_o, and a predicted variable, ψ_p, the climatology or ensemble averages are identical, i.e.

$$\langle \psi_o \rangle = \langle \psi_p \rangle,$$

as are the ensemble averages of the variance, i.e.,

$$\langle \psi_o^2 \rangle = \langle \psi_p^2 \rangle.$$

$\langle \ldots \rangle$ denotes the ensemble averaging. Ensemble averaging is an average over similar events. It is not to be mistaken for time averages. For example, we may desire to discuss the ensemble average of a time average, e.g., $\langle\{\psi\}\rangle$, where

$$\{\psi\} = {}^{\tau}\overline{\psi}^{T} = \frac{1}{T+1}\sum_{s=\tau}^{\tau+T}\psi(s).$$

Here τ is the lag time and $T+1$ is the averaging time. The lag time, τ, is always a measure describing how far from the beginning (i.e., the initial conditions) of a particular prediction we are. It does not necessarily denote any absolute calendar date. For a particular example, $T=4$ and $\tau=6$ denote a 5-day average of a predicted variable, beginning (lagged) 6 days into the future, or in other words, a prediction of a 5-day average, averaged from day 6 to day 10.

The correlation between the predicted and observed state will be used in this paper to be the measure of the error. This measure has the advantage over the root mean square (RMS) error in that dynamical models, as well as filtered models of one kind or another can have the same correlation even if the RMS error is different (see R). For example, if a perfect dynamical model is filtered via a perfect statistical model, the statistical model and the dynamical model will initially have the same RMS error but at long lags the statistical model has $(1/2)^{1/2}$ the RMS error of the dynamical model; however, the correlation between forecast and observed remain the same for both perfect models for all lags. Thus the correlation measure may provide a uniformly good description of the skill over a broad variety of dynamical as well as filtered models.

Now the correlation between two time averages for a perfect model, $(\langle\psi_o\rangle = \langle\psi_p\rangle)$ in which climatology is removed from the time series, is

$$\rho = \frac{\langle\{\psi_p\}[\psi_o]\rangle}{\langle\{\psi_p\}^2\rangle^{1/2}\langle[\psi_o]^2\rangle^{1/2}}.$$

Here

$$\{\psi_p\} = {}^{\tau_1}\overline{\psi_p}^{T_1}$$

and

$$[\psi_o] = {}^{\tau_2}\overline{\psi_o}^{T_2}.$$

In order to compute the correlation for the time averages, we first need to have the form of the variance and covariance for the individual days. For this, R is to be referred to. A first order Markov model was chosen to model the statistical properties of variance functions of observations and predictions. The model is chosen to have a decay constant of $\gamma = 0.3$ day^{-1}. The Lorenz RMS error model (Lorenz, 1969, 1982) along with an assumed exponential decay for noncoincident forecast and observation days was used to describe the autocovariance function. The assumption made in R was

$$\langle\psi_o(s_2)\,\psi_p(s_1)\rangle = \frac{\langle\psi_o^2\rangle(2+e^{-2x})}{(2+e^{2x}+e^{-2x})}e^{-\gamma|s_2-s_1|},$$

where

$$x = \tfrac{1}{2}a(s_1 - T_c).$$

This model is governed by three parameters: γ, a, T_c. For convenience, γ is assumed to be the same as that for the variance functions. The model is most sensitive to T_c and less sensitive to a and γ.

Some justification for this model of the covariance was given by R in a limited analysis of European Centre for Medium Range Weather Forecast data. Much more data is needed in order to verify the model proposed in R or to obtain even better models. Presumably as more extended range forecasts are made, this data will be forthcoming. For now let us assume that the chosen forms for the variances and covariances are reasonable and hence reasonable models describing the forecast skill, by NWP models, of time averages can be obtained.

To summarize, quite arbitrarily, $\gamma = a = 0.3$ days^{-1} along with $T_c = 9$ days was chosen for the analysis to be presented in this note. (Remember, only these parameters are used in the simple models of the variance and covariance functions. They describe the amount of serial correlation in the predicted and observed time series, the error doubling rate as well as the eventual quadratic damping, and the error in the initial conditions.) The basic picture to be presented in this paper is not changed for other parameters that may be slightly more relevant to present day models. In that regard, we must be careful to distinguish what variable as well as what model we are discussing. The parameters chosen here probably correspond best to what some of the future NWP models will be able to predict in the Northern Hemisphere 500 mb geopotential field at extended range. For present day models or for other variables besides the 500 mb geopotential, it would probably be better to lower T_c a bit. Also, even though γ, a and T_c can be adjusted somewhat to fit the desired data, it may eventually be better to use the actual data. Finally, for different lags and averaging times, slightly different parameters may be more applicable. The idealized model and particular parameters chosen here are used simply to illustrate various points of possible interest in extended range forecasts. The parameters should not be taken as the best fit to an arbitrary data set; that needs to be determined by the user.

The dashed curves in Fig. 1 show the correlations of various forecasts of time averages at various lags. The solid curves will be discussed later in the section on multiple regression. Here $\tau_1 = \tau_2$ and $T_1 = T_2$. Again, the τ's indicate the lags or the time from the beginning of the forecast (the initial conditions) to the start of the time average; the T's indicate the averaging time; subscript $_1$ indicates that this value is referring to the NWP forecast; subscript $_2$ indicates that this value is referring to observations. These curves are to be contrasted with those in R, where only $\tau = 0$ was considered and the correlation was plotted as a function of T. Here, each curve represents a different T.

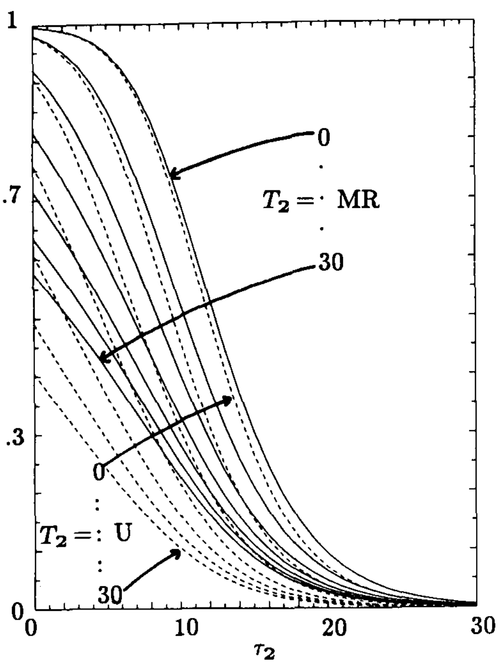

Fig. 1. The correlation, ρ, between forecast and observed for various desired time averages, $T_2 = 0$, 5, 10, 15, 20, 25, 30 days, at various lags, $\tau_2 = 0 \ldots 30$ days. The solid lines refer to the forecasts filtered by multiple regression, MR, and the dashed lines refer to the raw unfiltered forecasts, U.

The uppermost dashed curve shows the correlation decrease for instantaneous predictions at lags (τ) of 0 to 30 days. The lack of perfect correlation at day 0 for the instantaneous predictions simply indicates the magnitude of the initial error. As the lag increases the correlation decreases, initially slowly and then much faster and then once again slowly, toward zero. This correlation is simply the shape of the tanh function.

As longer and longer time averages are considered, the correlation decreases. The reason for this is that the longer time averages include individual days that have a longer lag from the initial conditions and hence a lower skill. Likewise, a time average must have greater skill than the forecast of the individual day at the end of the forecast time average since the time average includes days nearer to the beginning days when the skill is higher. For example, the correlation for an instantaneous forecast at day 30 is ~ 0; on the other hand, the forecast of the 30 day average over days 0 to 30 is ~ 0.4, with the increase resulting from the high skill in forecasting the first few days.

There are a number of ways to increase the skill of the raw numerical predictions given above. One way is to use multiple regression on the individual forecast days and this is discussed in the next section.

3. Multiple regression filter

The multiple regression filter for a particular time average can be written

$$^{\tau_2}\overline{\psi_o}^{T_2} = {}^{\tau_3}\overline{\alpha\psi_p}^{T_3},$$

where $\alpha(s_3)$ are the statistical weights to be determined by the method of least squares. The $\alpha(s_3)$ are determined in the usual least squares way by finding those values that minimize the error variance for the ensemble.

Fig. 2 shows the α's as a function of s_3 ($s_3 = 0$

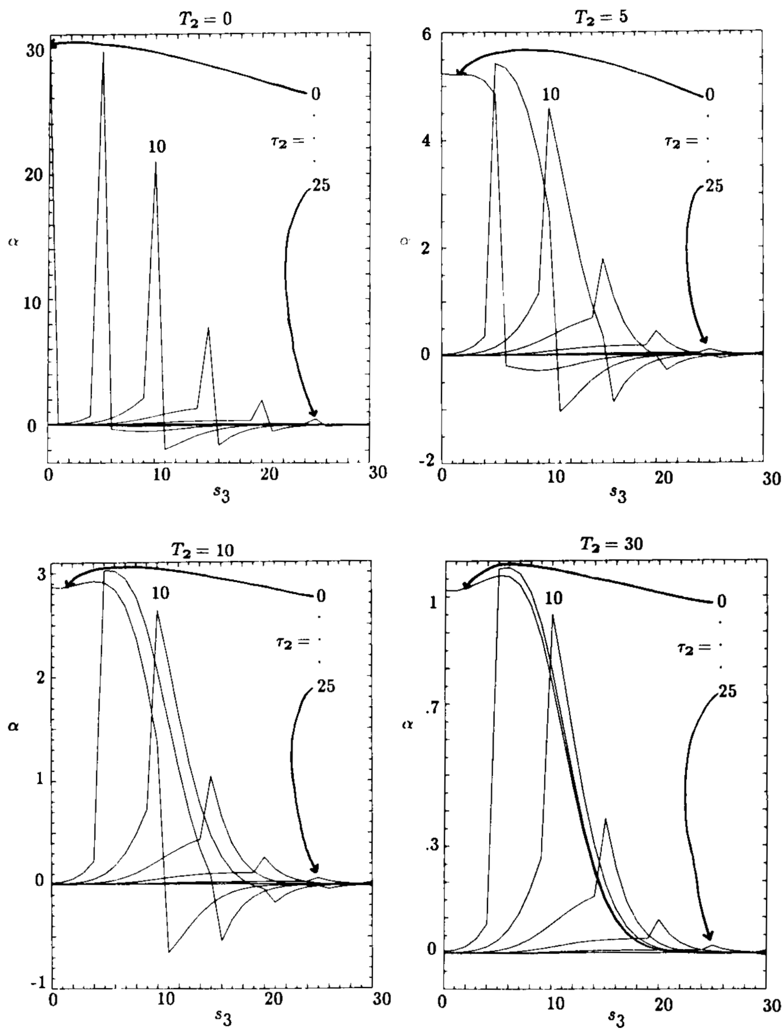

Fig. 2. Regression coefficients or statistical weights, α, as a function of each predictor day of the NWP model, $s_3 = 0 \ldots 30$ days, for various desired time averages, $T_2 = 0, 5, 10, 30$ days, at various lags, $\tau_2 = 0, 5, 10, 15, 20, 25$ days.

to 30 days) for various predictions. Although the weights were determined both for $s_3 = 0$ to 60 and $s_3 = 0$ to 30 days, no discernible difference in the first 31 α's was noted for the parameters chosen here; this indicates that the $\alpha(s_3)$ for $s_3 > 30$ days are unimportant for the predictions considered here.

For the daily predictions ($T_2 = 0$), at short lag times ($\tau_2 \sim 0$), essentially only the day corresponding to the prediction is participating ($\alpha(s_3) \sim 0$ for $s_3 \neq \tau_2$). For longer lead times, some marginal contribution from the initial days as well as the days after the lag (τ_2) day participate. Curiously, the weights are negative after the initial day of the desired lag. That is, $\alpha(s_3) < 0$ for $s_3 > \tau_2$.

For longer time averages ($T_2 > 0$), the peaks in the weights are broadened for $s_3 \sim \tau_2$ and $\tau_2 \sim 0$, but for increased lags ($\tau_2 \gg 0$) the α's tend to sharpen again toward the initial day of the time average. That is, a clear maximum exists for $s_3 = \tau_2$. Again, as in the forecasts of the daily values, the weights are initially positive, rise sharply to a peak for the initial day of the desired average, then decay toward the end of the desired average ($\tau_2 + T_2$). For days after the desired average ($s_3 > \tau_2 + T_2$), the coefficients are again negative but asymptotically approach zero.

Comparing the skill of these predictions (solid lines) with the skill of a straightforward unfiltered prediction (dashed lines) in Fig. 1, we can see a substantial improvement in all forecasts using the multiple regression filter. This is especially noticeable in the 30 day averaged forecasts. Here the variance explained (ρ^2) is almost double that of a forecast that uses only a top hat (standard time average) weighting profile for the individual forecast days.

Although the filter based upon the multiple regression coefficient weights of various days is likely to produce an increase in skill for a large ensemble, it is not altogether certain that this skill can be achieved in practice due to potentially high artificial skill. That is, although actual forecasts can be used to generate the statistical weighting coefficients, unless a very large number of forecasts can be used, the multiple regression filter is likely to be deceptively good on the data set used to generate the coefficients and deceptively bad on an independent sample (see Davis, 1976). Therefore, we should always prefer methods that provide almost equivalent skill but have fewer parameters.

4. Forecast window filter

An obvious simple filter, consisting of only two parameters, is finding the discrete interval of the prediction which is most closely correlated with the desired lagged time average. That is, we wish to maximize the correlation by finding the τ_1 and T_1 of the forecast that provide the optimum forecast for a desired lagged time average given by τ_2 and T_2. This optimum window is determined numerically in this note by taking a particular desired prediction (say $\tau_2 = 5$ and $T_2 = 30$ days) and then calculating the best correlation between this time average and a forecast time average. The forecast time average ranges over a two dimensional space, ($0 \leqslant \tau_1 \leqslant 30$ and $0 \leqslant T_1 \leqslant 30$ days), for a particular desired time average (i.e., a specific τ_2 and T_2).

An example of such a calculation is given in Fig. 3 for $T_2 = 30$ and $\tau_2 = 0, 5, 10, 20$ days. The thing to note here is that there is a maximum in the correlation field at somewhat surprising values for τ_1 and T_1. For example, for a forecast of a 30 day average lagged at 10 days, Fig. 3 shows that the best correlation is found by using $T_1 \sim 6$ and $\tau_1 \sim 9$. Thus the forecast should only be made out to day 15 and the average of the forecast over days 9 to 15 should be used as the optimum forecast of the 10 day lagged 30 day average.

Many correlations of this sort can be evaluated and from a limited number of experiments it appears that $\tau_1 \sim \tau_2$ for $\tau_2 \sim 0$. As τ_2 increases, τ_1 approaches an asymptotic value approximately equal to the time at which the correlation for the daily prediction drops to about 50% ($\sim \sqrt{2}T_c$). As for the optimum averaging time, $T_1 \sim T_2$ for τ_2 and $T_2 \sim 0$. However, T_1 asymptotically approaches a limiting value for τ_2 small but T_2 large. As both τ_2 and T_2 increase then T_1 increases such that $T_c + T_1 \sim \tau_2$. Thus, an increased averaging time does help to marginally improve the skill at this very extended range. It must be stressed again, however, that the optimum forecast average and lag (the forecast window) is still much different than the desired lagged time average.

Fig. 4 shows the correlation using the window filter (dashed lines) versus the multiple regression (solid line) filter. Multiple regression is superior to the window filter, although, in comparison to the increase from unfiltered to filtered prediction data, only slightly. Therefore, since fewer filter coefficients must be defined it will probably be better, at least with limited data sets to use the window filter.

5. Conclusions

An idealized model for error growth of forecasts of time averages by an NWP model has been examined in this note. These time averages were of arbitrary duration and arbitrary lag. Hence this note extends R's special case of zero lag. As in R, it was helpful to filter the forecasts since the variance explained could be substantially increased with the proper filter.

One filter was multiple regression. Here each

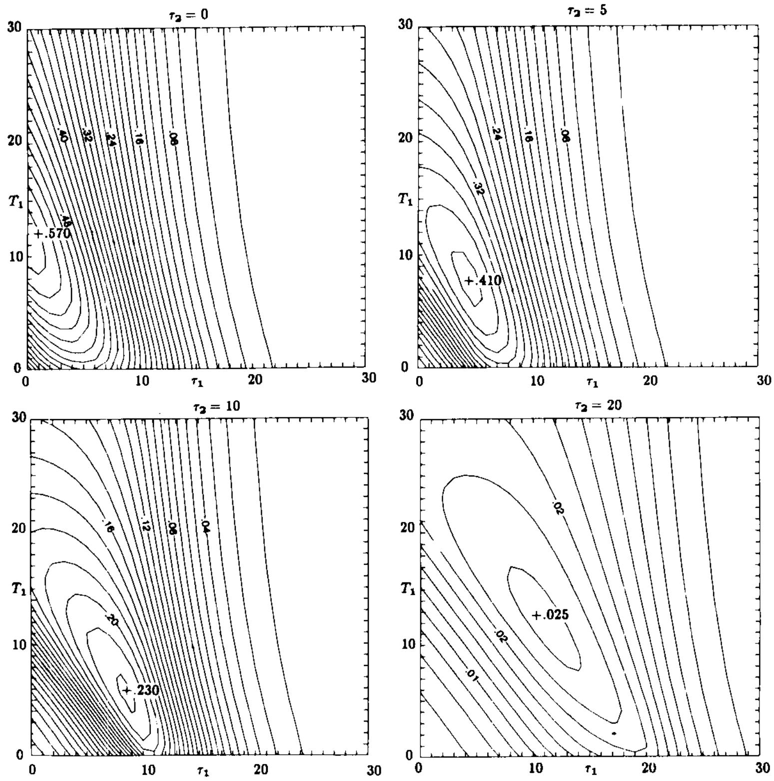

Fig. 3. Correlation contours, between forecast and observed, for a desired time average of $T_2 = 30$ days lagged at $\tau_2 = 0, 5, 10, 20$ days using NWP model prediction time averages of T_1 days lagged at τ_1 days. Different contours are used for the different τ_2. The maximum value for the correlation is indicated on the interior of each figure following the + sign.

day of the forecast was statistically weighted. For the forecasts of discrete time averages lagged from the initial day, the initial days of the forecast marginally contributed with greater contributions occurring as the actual forecast and desired forecast days approached each other. A strong combination then occurred when the regression and desired lag days coincided. The weights subsequently decayed with increasing forecast time and then became negative for forecast times past the desired time average.

A two parameter (window) filter provided almost as much skill. Here the best time average and lag of the forecast (or, in essence, the best forecast window) were determined. The optimum forecast lag reached a limiting value near to where the daily correlation approached 0.5. The optimum prediction time average increased, however, such that the lag plus time average was approximately equal to the lagged first day (or first few days) of the desired time average. That is, after the daily forecasts begin to drop in skill below about a 0.5 correlation, it is better to average the forecasts up to about the first day or so of the desired lagged time average. Forecasts beyond this point are not very useful.

To summarize, let us assume that we wish to predict the time average, T_2 days in duration, averaged from day τ_2 to day $\tau_2 + T_2$ days into the future. Again, τ denotes the lag from the initial conditions and T denotes the averaging length. One way to make this prediction is to use a numerical weather prediction model forecast starting at day τ_1 then averaging the forecast from day τ_1 to day $\tau_1 + T_1$. τ_1 and T_1 can be the same as or different from τ_2 and T_2. As was

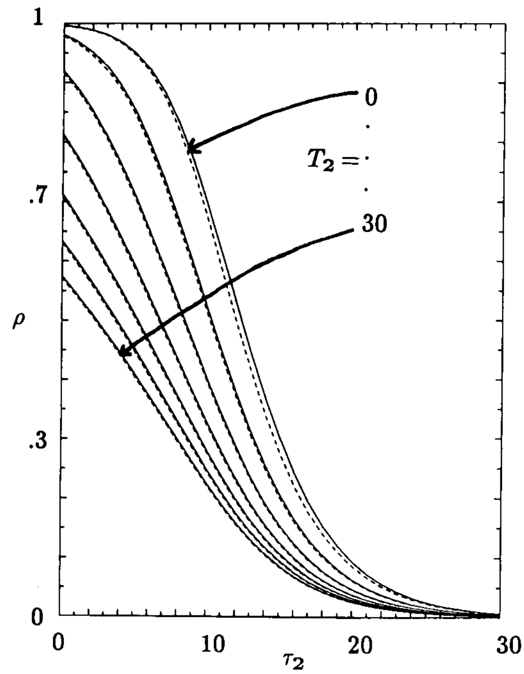

Fig. 4. The correlation, ρ, between forecast and observed for various desired time averages, $T_2 = 0$, 5, 10, 15, 20, 25, 30 days, at various desired lags, $\tau_2 = 0 \ldots 30$ days. The solid lines refer to the multiple regression forecasts and the dashed lines refer to the forecasts using the optimum NWP model τ_1 and T_1.

shown, the best forecasts do not necessarily have $\tau_1 = \tau_2$ and $T_1 = T_2$. In fact, as was shown in R, for $\tau_1 = \tau_2 = 0$ and $T_2 \gg 0$, $T_1 \ll T_2$. For example, to make the best forecast of the average of the next successive 30 days, it is probably best to average an NWP forecast over the first few days (say 10 until further analysis is done, see R) and call that average the forecast of the monthly average.

The extension here to arbitrary lags shows that for $\tau_2 \gg 0$, $\tau_1 \sim \sqrt{2}T_c$, where $\sqrt{2}T_c$ is the approximate time at which the RMS error curve crosses the climatology curve or $\rho \sim 0.5$. Also, the extension in this note shows that for $\tau_2 \gg 0$, $T_1 + T_c \sim \tau_2$. For example, to make a forecast of a 30-day average starting 30 days from now (i.e., forecast the January average starting at the beginning of December) it is best to average the daily forecasts from the time the daily forecasts have a correlation of 0.5 (say 1 week) up to the beginning of the desired forecast (beginning of January). That is, the average of the forecast for the last 3 weeks of December provides a better forecast for the month of January than the actual NWP forecast of the January average.

If a sufficiently large data set were available, an even more skillful filter would be to properly weight each day of a forecast by using multiple regression. Unfortunately due to the present limited data sets, the artificial skill is likely to be way too high for this latter multiple regression method and hence the actual forecast skill is likely to be way too low, in practice. Moreover, the increase in skill with the multiple regression filter, over the window filter, may be small even under the ideal conditions that the statistics are estimated with no sampling error. This was observed in the present study (Fig. 4).

Finally, the conclusions of this paper are based upon a very simple and idealized model for error growth. Much additional work needs to be done in order to determine how real NWP forecast models behave (see R). In that regard, the scientific community is eagerly awaiting the results of the dynamical extended range forecast experiments to be conducted over the next few years by various NWP modeling groups. Also, the influence of anomalous boundary conditions or specific synoptic situations may change some of these conclusions, especially at the very extended ranges or limits of present day NWP forecasts.

6. Acknowledgements

This research has been supported by NSF Grants ATM85-05435 and ATM85-20540, NOAA Grant NA81-AA-D-00054 and NA86AA-D-CP104, and Calspace. The multiple regression filter was originally suggested to me by R. Davis of Scripps. I am also grateful to E. Epstein of NMC for showing me some of his unpublished multiple regression results. I am grateful to J. Tribbia of NCAR for showing me some of his unpublished research on predictability of time averages with the NCAR Community Climate Model. Useful comments by J. Horel of Scripps are appreciated. I also thank V. Roberts and P. d'Ursin for help with the word processing. I also thank the anonymous reviewers for their helpful, constructive criticism.

REFERENCES

Davis, R. E. 1976. Predictability of sea-surface temperature and sea-level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr. 6*, 249–266.

Lorenz, E. N. 1969. Atmospheric predictability as revealed by naturally occurring analogs. *J. Atmos. Sci. 26*, 636–646.

Lorenz, E. N. 1982. Atmospheric predictability experiments with a large numerical model. *Tellus 35*, 505–513.

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

Roads, J. O. and Barnett, T. P. 1985. Forecasts of the 500 mb height using a dynamically oriented statistical model. *Mon. Wea. Rev. 112*, 1354—1369.