Averaging Techniques in Long-Range Weather Forecasting

A. N. SEIDMAN

The Aerospace Corporation, P.O. Box 92957, Los Angeles, CA 90009

(Manuscript received 3 June 1980, in final form 30 December 1980)

ABSTRACT

A method is investigated for increasing the length of prediction time for intermediate-range forecasting (up to 30 days). The method is to use the ensemble average of a set of forecasts generated by random perturbations from an observed initial state. The effect of ensemble averaging on the predictability time is investigated by means of a simulation study utilizing a three-layer general circulation model. It is found that the technique of ensemble averaging can lead to an increased predictability time when compared either to a single forecast made from the observed state, or to climatological means.

It is found that 1) the distribution of forecasts made from states which are randomly perturbed is Gaussian, within the limits of the numerical experiment; 2) both amplitude (root-mean-square) and phase (correlation coefficient) predictability times are increased for the ensemble-average forecast when compared to the forecast made from the observed initial state; and 3) the number of forecasts necessary to constitute a usable ensemble lies between four and eight. In addition, the procedure of ensemble forecasting was applied to averages over space regions ranging from 4×10^6 to 10^7 km^2, as well as time averages over periods of 5 days. Space-time averaging appears to emulate ensemble averaging in its effect on predictability time for ground temperatures, but not for surface pressure.

It should be emphasized that the results obtained here are based on a specific single initial weather situation and on a particular model's response to that situation. A "perfect model" assumption is made, i.e., that the degradation of the forecasts is due to incorrect initial conditions and not to the model. However, the model used is imperfect. The results obtained here are indicative, but care must be exercised not to extrapolate the results beyond the circumstances and assumptions without further investigation.

1. Introduction

Current numerical weather model forecasting employs space and time averaging. A typical time-average in a global atmospheric forecast model is about 5 min. Thirty-day time-averaged 700 mb weather maps are routinely used in long-range forecasting. Space-averaging is also used routinely. For example, a numerical forecast model might require temperatures, pressures, specific humidity, and winds for a number of volume elements 400 km by 400 km by 1 km.

Another type of averaging proposed is that of ensemble averaging. Leith (1974) has suggested the use of ensemble averages to obtain a Monte Carlo approximation to a stochastic-dynamic forecast (Epstein, 1969a, 1969b; Fleming, 1971a,b; Pitcher, 1977) which would involve a substantial reduction in the numerical computation effort. Jastrow (Druyan et al., 1972) proposed the use of ensemble and space-time averaging as a method to average out random noise-like fluctuations over the intermediate range of month-long forecasts. Here the idea was that before the predictability limit of some 2–4 weeks, but after some initial model readjustment period, averaging might prove useful.

Leith (1974) has examined the ensemble forecast in detail as a best forecast in the least-squares sense based on a two-dimensional model for homogeneous isotropic turbulence (Leith and Kraichnan, 1972). The results presented here are oriented toward examining ensemble and space-time averaging with a somewhat more detailed general circulation model. The point of view is that of an operational forecaster, where applicable.

The present United States weather observing system gives useful forecasts that play an important role in aviation and other sectors of the economy. However, beyond the range of 24–36 h, the accuracy of the forecasts degrades rapidly. Present-day weather forecasts serve up to 72 h, but beyond four or five days it is difficult to extract useable information.

The present investigation is concerned with the degradation of forecasting accuracy by the uncertainty of the initial state. The currently available observing system has a spacing of ~300 km over land areas in the Northern Hemisphere; observations are typically taken twice daily to a precision of 1–2°C in temperature, 3–4 m s^{-1} in winds and 3–4 mb in pressure. Over the oceans, error limits are much larger and coverage is very sparse.

0027-0644/81/071367-13$07.25

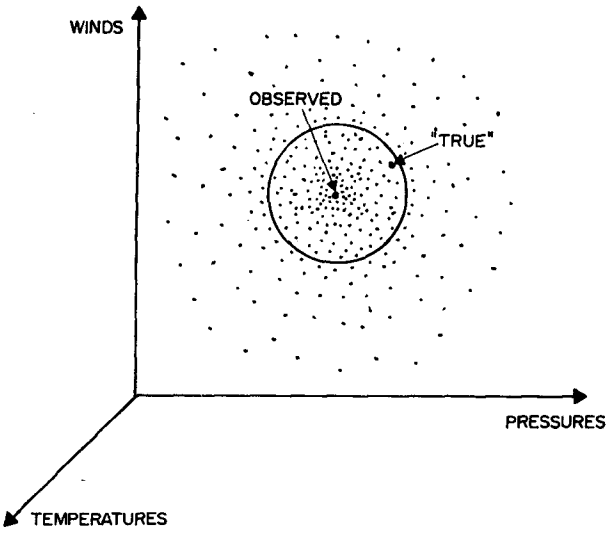

FIG. 1. Distribution of possible true states around the observed state (schematic representation).

The combination of substantial errors over Northern Hemisphere land, plus extremely sparse coverage over other large, fixed regions of the globe, gives rise to initial-state errors that are both systematic and random in nature. Additional errors of both systematic and random character are introduced by the methods used for analyzing the observations and preparing them for insertion into the numerical integration routines.

The systematic errors arise, for example, from biases in the instruments that measure atmospheric temperatures from satellites; and from large-scale, fixed gaps in coverage, e.g., over the oceans and polar regions.

The random errors can arise from individual measurements in the observing system (e.g., the satellite temperature soundings); the analysis schemes, which also generate random errors (e.g., when the observations are interpolated from the weather stations at which they are made to the locations of the computing grid); and finally, a random component, mainly due to clouds as far as remote global observations are concerned. When clouds are present, temperature soundings of the atmosphere from satellites are degraded. Since the position and extent of clouds vary irregularly, the errors introduced into the satellite temperature soundings by cloud effects have a strong random element.

The present research explores the reduction of the impact of initial-state errors on forecasts in the one-month interval. *Only random errors are considered; systematic errors are assumed to be zero in this study.* The technique consists in generating an ensemble of forecasts (possible true state evolutions) from a cluster of initial states (possible true state initial distributions) grouped in the neighborhood of the observed state. Each initial state in this cluster differs from the observed state by small random variations in temperature, wind and pressure fields. The magnitudes of these variations represent the estimated errors in the global field of observations.

The temperature, wind and pressure fields for each individual forecast are then averaged over the ensemble to produce ensemble-averaged fields. The specific objective of this investigation is to determine how much better these ensemble-average fields do compared with representative single forecasts. The assumption is that the model itself is perfect, i.e., degradation of the forecast is due to incorrect initial conditions and not to any imperfections in the forecast model.

In order to examine the improvement in an ensemble forecast, we must compare the error in the ensemble forecast with the error in the forecast from the single true state. We do not, however, know what the true state is. We do know the single observed state and the assumed distribution of possible true states around the observed state; this distribution is taken to be the same as the distribution of observed states around the true state (Leith, 1974) for the case where the distribution of states depends only on the relative separation of the states and not on the specific numerical value associated with the true state. Thus, we can generate a distribution of true states around the observed state and use this true state distribution to compare with the forecast based on the ensemble mean and with the forecast based on the single observed state. By comparing the mean errors in the two forecasts, we can evaluate the usefulness of the ensemble method.

Although we do not know which state is the true one, we know that the true initial state is a member of the distribution of the possible true initial states shown schematically in the cluster in Fig. 1. Since the distribution of observed states is Gaussian because of the random nature of the error in the observed state, the distribution of true states around the observed state is taken as Gaussian. The width of the Gaussian distribution is known from the observers' estimates of the probable errors in their data.

We can make forecasts from the initial conditions represented by points in the schematic "phase space" picture of Fig. 1. Let us also represent the forecasts generated from these various initial states schematically, as shown in Fig. 2. In this figure, the state of the atmosphere is represented one-dimensionally on the vertical axis. Time is plotted horizontally starting from zero.

The method of producing a possible true state distribution about each observed weather map is described in Section 2. Briefly, a Gaussian random distribution is produced on each map with a mean

of zero and a standard deviation of 1°C in surface temperature, 3 mb in surface pressure, and 3 m s^{-1} in each of the wind components. These error limits are optimistic with respect to the current global observation network.

One further point requires attention: how large an ensemble is necessary to achieve the improvement in forecast accuracy indicated above? Since a heavy computing burden is involved in generating the forecasts in the ensemble, this question is of practical importance in judging the usefulness of the ensemble-averaging technique. In anticipation of the discussion of the results in Section 3, it appears that an ensemble of four to eight forecasts is adequate, at least for the particular numerical experiment involved here, while little change is noted in the results between an ensemble of eight cases and the full ensemble of 34 cases.

A second objective of the present research is to determine the effect produced on intermediate-range forecasts by averaging the forecasts over substantial geographical areas and intervals of time. Many practical applications of intermediate-range forecasting do not require detailed day-to-day information at a given place; in agriculture, for example, soil moisture is important and a forecast of average temperature and rainfall for a 1-week period averaged over a state or region would be of great value. The combined effect of space-time averaging and ensemble averaging is also discussed in the section on results.

2. The numerical experiments

Ideally, one would use a comparison with actual observations, i.e., with the real atmosphere, as the basis for this investigation of forecast accuracy. However, present-day atmospheric general circulation models do not forecast the real weather beyond a few days or so with any degree of accuracy. Since we wish to test the impact of ensemble-averaging on forecasts out to 30 days we instead utilize simulated data generated by a three-layer general circulation model. Simulated real weather was produced over a 30-day period. The initial conditions were randomly perturbed 33 times and 34 separate 30-day forecasts were made from these different initial data sets (one original plus 33 perturbations). The 34 forecasts constituted an ensemble of forecasts which could be analyzed.

a. The three-layer model

The model used is a three-layer version of the Goddard Institute for Space Studies (GISS) general circulation model (Somerville *et al.*, 1974). The three-layer model differs from the GISS nine-layer model in its treatment of shortwave radiative transfer; it uses different ground and ice-snow albedos,

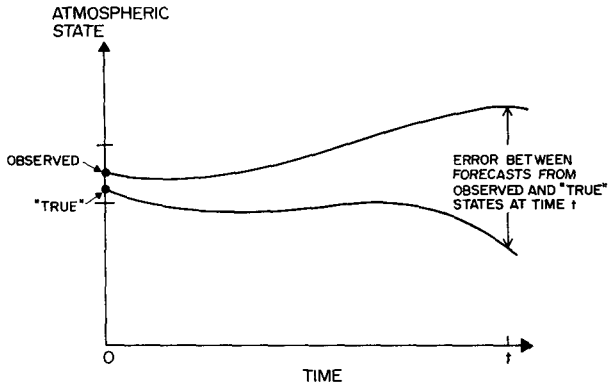

FIG. 2. Error between forecasts from observed and possible true states as a function of time (schematic representation).

a different ground wetness array, and a different formulation of the heat and moisture diffusivity in the vertical. It also differs in that, although the same longwave radiative transfer physics is used, a cruder interpolation scheme for temperature and specific humidity is necessitated by the fewer number of levels. In addition, the cloud-layer specification for interaction with the longwave radiation is different. A description of the model with the associated general circulation model statistics can be found in Seidman (1976).

The model temperature analyzed is the ground temperature T_g. It corresponds to the soil temperature for observed data. Another quantity available from the model is the surface temperature T_s which is a boundary layer temperature for the constant flux boundary layer of the model. It corresponds to the temperature at a height that can range from the ground level up to the mid-level of layer three (nominal pressure: 850 mb or roughly 1.5 km). The shelter level temperature, which is the temperature that is used for surface verification at the weather stations, falls somewhere in between. The decision was made to use the T_g as the temperature to analyze, even though it does not correspond exactly to observed surface temperature observations. This should not result in any systematic error in the daily mean temperature, as the diurnal cycle of $T_g - T_s$ is averaged out.

b. Procedure for simulating random errors in initial conditions

The basic initialization data set used in these forecasts was obtained from the National Meteorological Center (NMC). This was a specially produced global data set using an experimental analysis procedure developed by Flattery. The analysis was based on a spectral representation, which today has become the operational analysis. However, at that time (0000 GMT 9 December 1972) the available

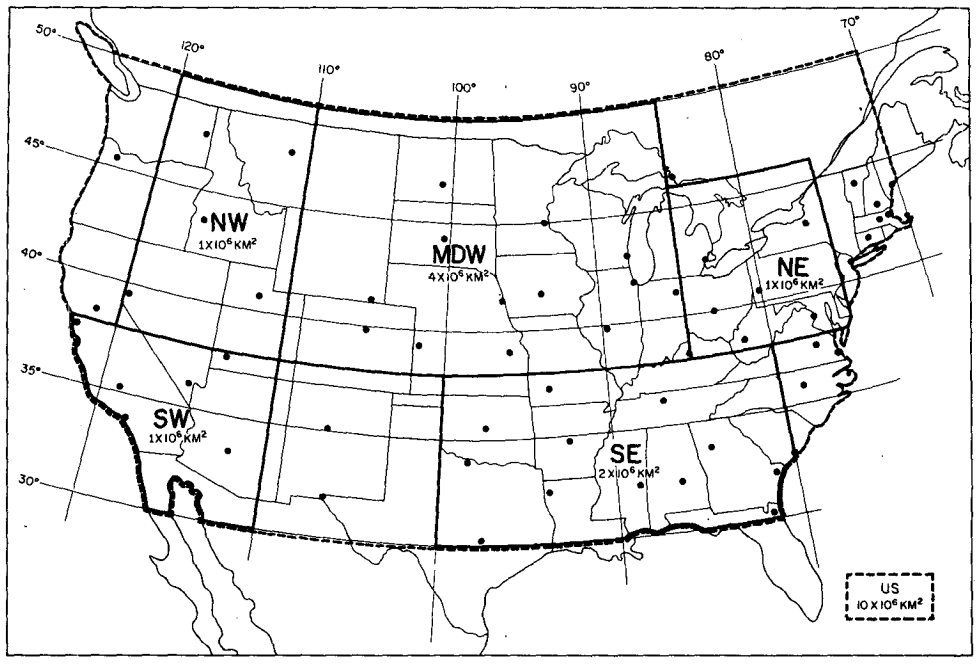

FIG. 3. Map of the six standard regions.

operational analysis used a Cressman interpolation procedure and was used where available north of 18°N. The point was to use a realistic initial condition to avoid a prolonged initial readjustment time.

Thirty-three separate initial states were generated, each being a perturbation of the basic data set of 0000 GMT 9 December 1972. The perturbations correspond to Global Atmospheric Research Program (GARP) error limits, i.e., 3 mb in surface pressure, 1°C in temperature, and 3 m s^{-1} in winds. These error limits are not achieved yet in global observations; they are a well-defined set, however, and represent a lower bound on present observational errors.

The IBM scientific subroutine random number generator was used to add random errors with a Gaussian distribution having these error parameters. The surface pressure for each initial state was perturbed at each of the surface grid points of the model (46 meridional, 72 zonal) with a standard deviation of 3 mb and a mean error of zero. The global standard deviation of the error applies to each initial state, so that each pressure map has an independent rms error of 3 mb. The temperatures of the three layers and the surface temperature (T_s) which corresponds to a planetary boundary layer temperature, were perturbed at all grid points with a standard deviation of 1°C and a mean error of zero. In each initial state, the meridional and zonal wind velocity components were perturbed at all three layers at all grid points with a standard deviation of 3 m s^{-1} and a mean error of zero for each initial state.

c. *Classification of forecast regions*

Although the model produces global forecasts, the analysis in this experiment was limited to the continental United States. For the purpose of forming spatial averages, six regions were defined. These are designated as US, NE, SE, SW, NW and MDW. The boundaries of the regions are given in Fig. 3. The US region was chosen to encompass the continental United States, and it is of the order of 10^7 km². The NE or Northeast region, SW or Southwest region, and the NW or Northwest region were chosen to be $\sim 10^6$ km² and to approximate the geographical regions suggested by their names. The SE or Southeast region is of the order 2×10^6 km² and the MDW or Midwest region is of the order of 4×10^6 km². The regions were picked to represent regions of population and agricultural interest, while giving a mix of area sizes.

3. Results

The difference between two meteorological states may be expressed in terms of the rms difference. The measure may be applied to the difference of two weather maps or states, such as the ground temperature over a given area, and may be used to

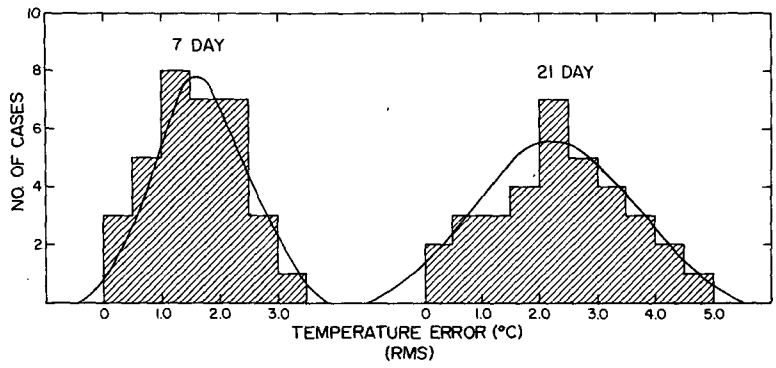

FIG. 4. Point ensemble distribution for ground temperature (US region) at 7 and 21 days.

determine the growth rate of separation between states (Charney, 1966). Other measures are possible, such as a correlation coefficient, absolute mean difference, or skill score. Each of these measures may give rise to different predictability times.

a. Definition of rms differences

The rms difference between two atmospheric states X and Y is given explicitly as

$$\text{rms}^n = \left[\frac{\sum_j \{\cos\phi(j) \cdot \sum_i [X^n(j,i) - Y^n(j,i)]^2\}}{\sum_j [N(j)\cos\phi(j)]} \right]^{1/2}, \quad (1)$$

where the symbols have the following meaning: $N(j)$ is the number of zonal grid points on row j, $X^n(j,i)$, $Y^n(j,i)$ represent a state variable as a function of grid-point coordinates (j,i) at a specific time (n), $\phi(j)$ is the latitude corresponding to the meridional grid point j, and i represents a grid-point coordinate in the zonal direction.

The above definition applies to a *point* difference, i.e., at a specific place and time. It carries a superscript n to indicate that it is evaluated at a specific time. Two other measures of difference may be defined. For the computation of a *space*-averaged rms difference, the $X^n(j,i)$ and $Y^n(j,i)$ are averaged over area before computing the rms difference. The space-averaged quantities $\overline{X}^n$ and $\overline{Y}^n$ then replace X^n and Y^n in (1) at each grid point for the calculation of the *space*-averaged rms difference. For the *space-time* averaged rms difference, an additional time-averaging is performed over $\overline{X}^n$ and $\overline{Y}^n$. The space-time averaged quantities, $\overline{X}$ and $\overline{Y}$, are substituted for X^n and Y^n in (1) for the calculation of the space-time-averaged rms difference. The averaging time period used is five days.

b. Apparent Gaussian character of the ensemble distribution

As a first step we examined the ensemble ground temperature distribution. This is the point ensemble distribution which means that the rms differences are computed over the individual points of a spatial region and averaged over the ensemble. We find that the distributions are compatible with a normal distribution. However, if the underlying distribution is normal or Gaussian, the mean-square distribution would be χ^2 and the rms distribution would then be χ (e.g., Cramer, 1946). As the number of degrees of freedom increases, however, both the χ^2 and the χ distributions asymptotically approach a normal distribution. With the number of degrees of freedom $\sim$ 30, the χ^2 and χ distributions are not empirically distinguishable from the normal with the information at our disposal.

Keeping the above in mind, we find that the distributions after 7 and 21 days are Gaussian, or normal, for the ground temperature averaged over the US and MDW regions. Fig. 4 shows the distribution of rms point error for the ground temperature for the US region. The distributions are shown for 7 days and 21 days after the initial time. The moment fit Gaussian is shown for each case. Fig. 5 shows similar distributions for the MDW ground temperature. Table 1 summarizes the characteristic features of these distributions, namely, the mean, standard deviation, skewness and flatness.

The moment fit is determined by using the standard deviation of the sample as the standard deviation parameter of the Gaussian, with the total area of the fitted Gaussian normalized to the total area of observations. The more values in a given class, the more confidence one has in the chi-square values. A minimum number of five values in each class is usually required, although ten is more desirable. Since we had only 34 values total, a minimum of five in each class was required. This gives about six

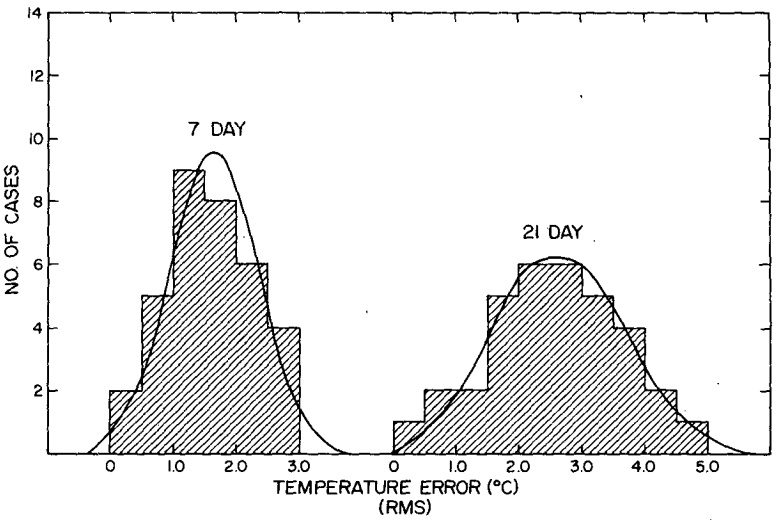

FIG. 5. Point-ensemble distribution for ground temperature (MDW region) at 7 and 21 days.

points to be fitted, and this is approaching a requisite lower limit for the number of points.

Consider the hypothesis that the observed distribution of rms errors relative to the ensemble average is Gaussian. We reject this hypothesis if $P(\chi^2 \geq \chi^2_\alpha) \geq \alpha$, $\alpha = 5\%$. If the inequality does not hold, we accept the hypothesis at the 5% level. It is found for all the cases shown that the hypothesis that the ensemble distribution is Gaussian can be accepted at the 5% level.

Fig. 6 shows the moment fit Gaussian to the rms error distribution for 7, 14, 21 and 28 days after the initial time. These Gaussian curves are for the ground temperature over the US region. Fig. 7 shows the ensemble standard deviation versus time for the above Gaussian curves.

From Table 1 we conclude that the distribution over the 34 cases is close to Gaussian. From Fig. 7 we see that the width of the distribution increases fairly rapidly during the first two weeks of the forecast period. Thereafter, however, it increases very slowly, and probably does not change appreciably beyond 21–28 days (dotted line in Fig. 7). We infer that the averages over the ensemble of forecasts are approaching the model's climatological state beyond 14 days and have reached this state at 28 days.

TABLE 1. Parameters of the theoretical normal and the observed distributions.

	Mean (°C)	Standard deviation (°C)	Asymmetry (skewness)	Flatness (kurtosis)
Normal distribution (theoretical)	—	—	0	3
US—7 days	1.59	0.70	−0.03	2.20
US—21 days	2.73	1.07	0.02	2.23
MDW—7 days	1.51	0.78	0.05	2.12
MDS—21 days	2.37	1.16	0.01	2.22

c. *Root-mean-square error growth*

We expect the forecast error to grow rapidly with time at first, eventually leveling off to a roughly constant plateau or asymptotic level. The plateau is reached when the difference between the control and the forecast reaches the level of the mean dif-

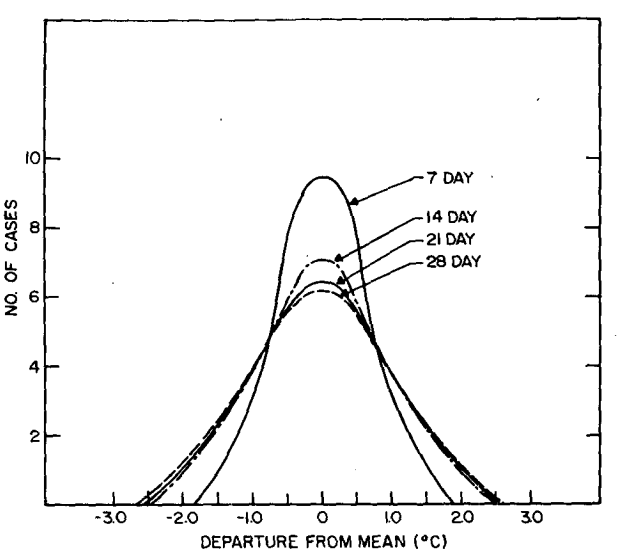

FIG. 6. Point-ensemble distribution for ground temperature (US region) at 7, 14, 21 and 28 days.

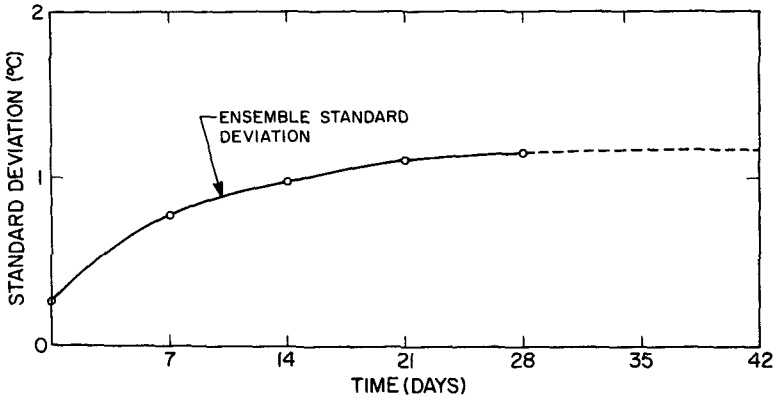

FIG. 7. Standard deviation of ensemble versus time for ground temperature (US region). The growth of the ensemble error toward the complete loss of predictability is indicated.

ference between randomly chosen states of the atmosphere. In simulation studies based on a model, the model is often less variable than the real atmosphere; accordingly, the difference in temperature, etc., between two circulation, but at the same observing point, will be smaller in the model than in the real atmosphere. In the present three-layer model computations, the differences between randomly chosen states are 8 m s^{-1}, 5 mb and 3.8°C for the average global wind, pressure and temperature, respectively. In the real atmosphere, the corresponding differences are about 13 m s^{-1}, 9 mb and 7°C (Halem, 1975, private communication). The fact that the corresponding levels for the model are lower than those for the real atmosphere is a measure of the lesser variability of the model's circulation.

In examining the growth of the rms error curves for the single forecast from the observed state (i.e., individual) and for the ensemble forecast based on the perturbations of the initial observed state (ensemble) it should be noted that the values obtained in the figures and tables are based on the comparison of the forecast under consideration with all of the different 33 members of the possible true state distribution. The average of the growth rates is then calculated.

In comparing the present results with those of Leith (1974), where all pairs of solutions are considered, the rms error growth curves are different in that the forecast from the observed state (f_0) is used as a reference so that the asymptotic rms error of the average growth of the individual forecast is not $2^{1/2}$ times the asymptotic rms error of the average growth of the ensemble average. Symbolically,

$$\langle (f_i - f_0)^2 \rangle = \langle [(f_i - f_e) - (f_0 - f_e)]^2 \rangle = \langle (f_i - f_e)^2 \rangle + \langle (f_0 - f_e)^2 \rangle ,$$

since $\langle f_i \rangle = \langle f_e \rangle = f_e$. Here f_i is a member of the possible true state distribution, f_e the ensemble average forecast; $\langle\ \rangle$ indicates ensemble average, and i runs over the members of the ensemble. The "axis of rotation" of our "moment of inertia" has been shifted from f_e to f_0. It is transparent that $\langle (f_i - f_j)^2 \rangle_{i \neq j} = 2\langle (f_i - f_e)^2 \rangle - 2f_e^2$ remains true as $t \to \infty$, where, for comparison, Leith (1974) has chosen $f_e = 0$.

Because the point of view adopted here has been to refer to the forecast from the observed state and not to an unknown (although approximatable) true state, the rms asymptotic variances do *not* become twice the ensemble rms variances. This is an artifact of how the present calculation was done and does not contradict Leith's (1974) analysis. It can be taken as a confirmation of the statement that the centroid of the initial ensemble does not remain the centroid of the time-integrated distribution (Epstein, 1969a).

Figs. 8 and 9 show two rms error curves, one is the average of the forecast from the observed initial state (individual forecast), relative to all of the other cases (i.e., possible true states), and the other is the average rms error of the ensemble average relative to all of the cases (i.e., possible true states). Ground temperature and surface pressure for the US region are shown in the two figures, respectively.

The noise level is higher in these curves compared with the usual presentation of global rms error curves. For the US region we are dealing with a fraction of the 3312 or so grid points on the global surface grid. In fact we have 60 points, or ~2% of the grid points for the global case. For the MDW region we have 27 points or about 0.8% of the total global grid number. The degree of fluctuation shown in the figures therefore is not unexpected.

Table 2 summarizes the asymptotic error levels for the ground temperature and the surface pres-

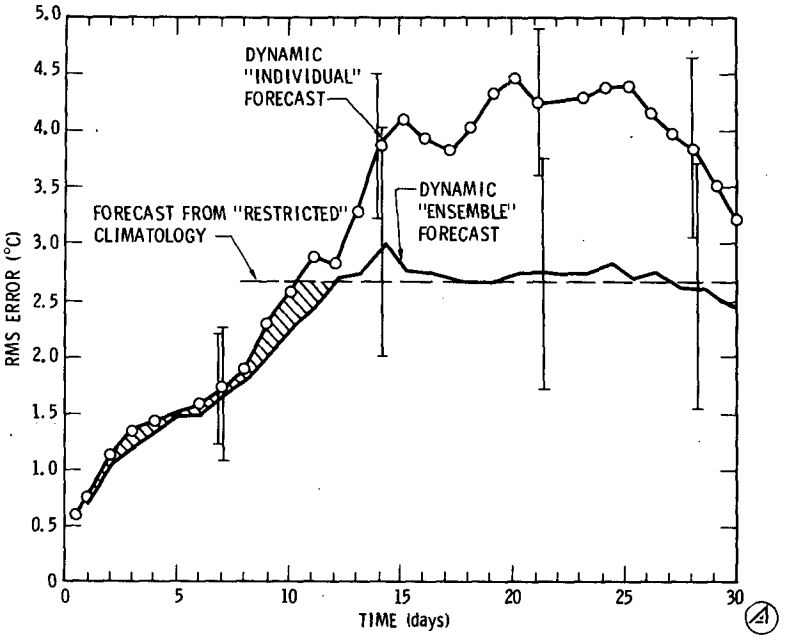

FIG. 8. Effect of ensemble averaging on point forecasts for ground temperature (US region). The average single forecast, the ensemble average and a "restricted" climatological forecast are shown together with one-standard deviation error bars.

sure. Table 2 also shows the differences and the relative percentage decrease due to ensemble averaging over the unaveraged forecast.

It should be emphasized that the time of interest is the time before total decorrelation occurs; detailed information on the ensemble average forecast versus the single forecast is sought. In the language of stochastic-dynamic forecasting, does the

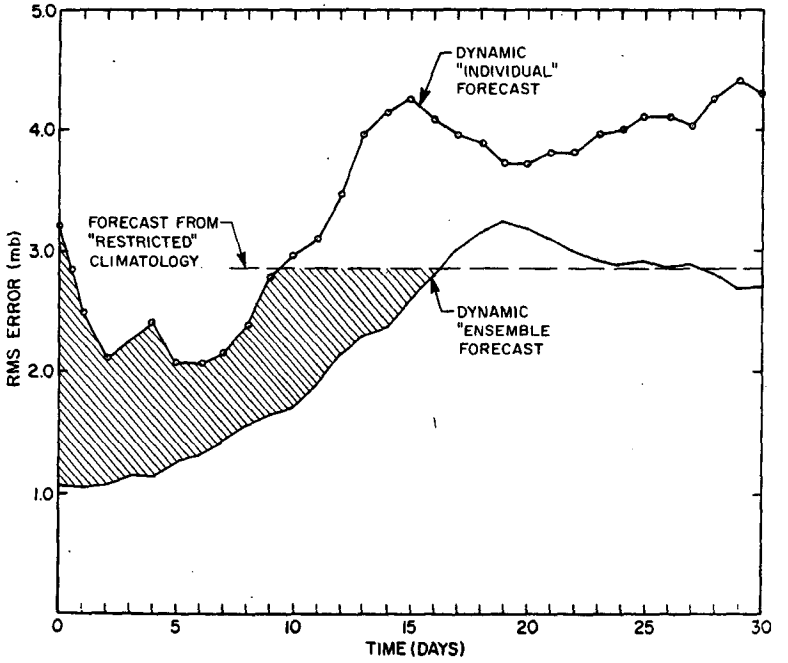

FIG. 9. Effect of ensemble averaging on point forecasts for surface pressure (US region). The average single forecast, the ensemble average, and a "restricted" climatological forecast are shown.

"certain" knowledge decrease slowly or rapidly, mostly at first, or mostly toward the point at which predictability is lost?

There is a considerable relative improvement due to ensemble averaging for both the point and for the space-time-averaged cases. For example, we see that the random error from uncertainties in the initial conditions has been reduced from 4.0 to 2.5°C, i.e., a relative reduction of 38% for the ensemble-averaged ground temperature compared to the non-averaged case point prediction (United States region).

The statistical significance of the difference between the curves presented in Figs. 8 and 9 is obtained by examining the error bars (± one standard deviation) for the two cases (cf. Fig. 8). Sixty-eight percent of the cases are likely to lie within one standard deviation, 95% within two standard deviations, and 99% within three. It can be seen from the figure that from 14 to 28 days there is relatively little significant overlap in the error bars. The hypothesis that there is no significant difference between the individual and the ensemble curves can be rejected at the 5% significance level for all cases from day 10 to day 30 with the exception of day 12 where the two curves blurred together. Here, the significance level of no difference was reduced to 15%. The surface pressure curves were different for all days at the 1% level of significance. The student t-test was used to evaluate the hypothesis of no difference between the curves.

d. *How do we compare the ensemble forecast with what an operational forecaster would obtain?*

In answering this question, we first recall the meaning of the averages over the individual forecasts. The true state can be any one of the states in the ensemble. The observed state is initially at the center of the initial states that generated the ensemble. At a later time, the true state is a member of the ensemble of the forecasts being generated from the true initial states. The forecast from the observed state no longer coincides with the mean of the ensemble of forecasts. With this explanation in mind, we return to the opening question of this paragraph.

The operational forecaster will use the forecast generated from the observed state up to the point where it is no better than a climatological forecast. That is, the operational forecaster would use a combination of a dynamic computer-generated forecast coupled with a knowledge of the real climatology.

For the case of this simulation study, as has been observed above, the ensemble average becomes the model climatological state after ~14 days into the forecast. Thus, we may compare the ensemble forecast to the average individual forecast and to the "restricted" climatology of the model, as shown for the United States in Fig. 8. Using this measure of our ability to improve the forecast by ensemble averaging, we see that, for the US region, the improvement in predictability time is roughly from 10 to 12 days or +20% (shaded area in Fig. 8). Results for the predictability times for both the US and the MDW regions are listed in Table 3 (based on ground temperature and on surface pressure). Note that the predictability time is based on the climate mean crossover point, i.e., when the rms difference equals the square root of the climate variance.

The difference is greater in the surface pressure results. For the surface pressure, the predictability time is increased, for example, from 9 to 16 days, or nearly twofold (US region, point, individual and climatology versus ensemble).

The fact that the ensemble average produced a larger improvement for surface pressure than in the

TABLE 2. Regional asymptotic rms error level for the US and MDW regions (°C ground temperature; mb, surface pressure); and the difference between individual and ensemble asymptotic error level (°C and percent; mb and percent).

	rms error level			
	US (°C)	MDW (°C)	US (mb)	MDW (mb)
Point				
Individual	4.0	4.5	4.0	4.0
Ensemble	2.5	2.0	2.75	1.8
Space-time				
Individual	1.0	1.5	1.8	2.0
Ensemble	0.75	1.0	1.0	1.5
	Individual and ensemble error level differences			
	US (°C)	MDW (°C)	US (mb)	MDW (mb)
Point	1.5 (−38%)	2.5 (−56%)	1.25 (−31%)	2.2 (−55%)
Space-time	0.25 (−33%)	0.50 (−33%)	0.8 (−44%)	0.5 (−25%)

TABLE 3. Regional predictability times (days for ground temperature and for surface pressure).

	Ground-temperature times		Surface-pressure times	
	US (days)	MDW (days)	US (days)	MDW (days)
Point				
Individual + climatology	10	9	9	8
Ensemble	12	11	16	14
Space-time				
Individual + climatology	13	12	6	7
Ensemble	15	15	15	18

ground temperature indicates that an indirect approach may be the best, namely, forecasting an ensemble-pressure map, and then deducing ground temperatures from the location of the air masses by a subjective analysis.

Beyond 12 days, the temperature error in the ensemble forecast is the same as the error in the model's climatological state. This means that a simulated operational forecaster, switching to climatology beyond 12 days, would do as well in his forecast as if he were to use the ensemble forecast; i.e., the ensemble forecast would not offer any advantage over climatology. However, in an important respect this estimate of the impact of ensemble averaging may be unduly conservative. It is based on a restricted definition of the model's climatology, which may diminish the errors calculated in the climatological forecast, making the climatological forecast seem more favorable in comparison to the ensemble forecast than would be the case in real life.

The reason for this distinction between the model's restricted climatological state and the real climatology is the following. In the model's climatological state defined in the "restricted sense" we see the effect of variations in the initial state used to generate the forecasts. However, we do not see the effect of variations in the boundary variables, i.e., large-scale patterns of sea surface temperatures, snow and ice cover and soil moisture (not to mention possible variations in solar luminosity). These variables, which form the boundary conditions in the numerical integrations, are the same for every member of the ensemble.

In the real world, the above quantities may vary from year to year, and almost certainly vary by considerable amounts over the course of several years. Thus, the width of the spread about the climatological mean may be greater in actuality than for the model's climatological state used in this simulation study.

Madden and Shea's (1978) estimates of the natural variability of North American surface temperatures range from a standard deviation of about 1.5 to 4.0°C for January and 0.5 to 4.0°C over all seasons. Madden's (1976) estimates for the sea-level pressure standard deviation range over 1–8 mb for January and for all seasons. The estimated ratio of interannual variability to natural variability range roughly over a factor of 1–3 times the natural variability (i.e., 1.5–12°C), while for pressure the range is also a factor of 1–3 (i.e., 1–24 mb).

The increased predictability time in the 30-day period suggests that the ensemble-averaging technique should lead to significant improvements in intermediate range forecasting, if implemented operationally. The space-time averaging led to mixed results: always better for the ground temperature predictability times, worse for the surface pressure predictability times except for the MDW where the space-time ensemble was better than the point ensemble case.

e. Correlation coefficient predictability

In measuring the degree of predictability in simulation experiments, one may consider two factors. The first consideration is how close is the predicted amplitude to the observed amplitude of fluctuation. The second consideration is how close is the predicted phase to the observed phase, i.e., does the predicted ground temperature rise and fall in agreement with observed temperature variations. The first factor is represented to a large extent by the rms difference between predicted and observed states, while the second may be estimated from the correlation in time between the two histories. The ability to forecast the sign of the change in a meteorological variable such as the ground temperature is clearly a useful skill, even if one cannot predict the magnitude of the change.

The aspect of predictability that is examined in this section is the correlation in time of both the ensemble average and the forecast from the observed initial state with the forecasts (i.e., possible true states). That is, in addition to the absolute difference between forecasts, how do the forecasts correlate in time? This will give a measure of the "correlation predictability time." The level at which one can say unambiguously that correlation has been completely lost is fairly clear, namely, when the correlation coefficient is zero. However, even nonzero correlation coefficients may be without significance.

Since the characteristic time between independent events based on a time-lagged autocorrelation-weighted integral (Leith, 1973) is ~6.7 days (Leith, 1974), a 30-day time series could be considered as having 4–5 independent events, or degrees of freedom. Madden (1976) provides the range of values over the globe of from two to seven days for independent surface pressure events based on the autocorrelation, and Madden and Shea (1978) show a range for autocorrelation-based independent sea-level temperatures of from 3–7 days for the continental United States over all seasons. The direct calculation of the correlation coefficient would suffer from the small number of degrees of freedom available, e.g., four.

In order to establish a higher degree of precision in the correlation coefficient results, the correlations are converted to the Fisher (1958) z-statistic, averaged and converted back to correlation coefficients. This allows one to increase the precision estimate of the correlation coefficient by a factor of $(N - 3)^{1/2}$, where N is the total number of corre-

TABLE 4. Correlation coefficient and z statistic parameters, where r is the correlation coefficient, z the value of the corresponding z statistic, NV the value of the normal variate for the given z statistic and P the probability of the value of the NV occurring by chance.

r	0.1	0.2	0.3	0.4	0.5	0.6
z	0.11	0.21	0.31	0.43	0.55	0.70
NV	0.60	1.15	1.70	2.36	3.01	3.83
P	0.55	0.25	0.09	0.05–0.01	0.01–0.001	0.001–0.0001

lation coefficient calculations being averaged. This results from the z-statistic approximating a normal distribution, whereas the correlation coefficient does not.

The running correlation coefficient is defined as

$$r = \left[\sum_{i=1}^{M} \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{M \sigma_x \sigma_y} \right]^{1/2},$$

where $\overline{X}$ represents the average ground temperature state of the control case over the 10.25-day interval moving "window" and X_i denotes a 6 h state within the interval, or, in the case of the ensemble correlation, the average ground temperature state for the 33 perturbed cases for a 6 h period. $\overline{Y}$ and Y_i represent ground temperature states for one of the perturbed cases.

M is the number of pairs being correlated in the moving "window" ($M = 41$). The initial time was allowed to move from 0000 GMT 13 December 1972 to 0000 GMT 3 January 1973. A running sequence of 41 pairs of states at each time interval is considered. The ground temperatures from the forecasts, computed at 6 h intervals, taken five days about a specific time, make up the 41 elements in the computed correlation. Consequently, the total time series covering the running correlation over 30 days consists of correlation coefficients individually spanning 10.25 days.

The quantities σ_x and σ_y are the estimates of unbiased standard deviations, defined as

$$\sigma_x = \left[\sum_{i=1}^{M} \frac{(X_i - \overline{X})^2}{M - 1} \right]^{1/2},$$

and similarly for σ_y.

Space-averaged correlation curves are obtained by considering the quantities X and Y as representing averaged quantities over a region. Space-time averaged correlation curves have the additional feature of having X and Y averaged over time as well. Finally, the ensemble correlation coefficient is defined by averaging the 33 forecasts to form X.

The point correlation coefficient is computed by taking the average of the pairs of values over the grid points of the regions. Ensemble versus individual correlation curves for point and space-time averages are calculated. The time period for the space-time average is five days.

The observed state is correlated in time with each of the remaining 33 cases and the results averaged using the z-statistic to give an average individual correlation coefficient. The ensemble correlation coefficient is calculated by first averaging the values for the ground temperatures of the 33 cases and then this single time series is correlated in time with each of the individual cases. The average correlation coefficient is then computed using the z statistic, which is defined as

$$z = \tfrac{1}{2} \ln[(1 + r)/(1 - r)].$$

The *standard error* of z is approximately $\sigma_z = (N - 3)^{-1/2}$.

The correlation coefficient, the corresponding z-statistic, the normal variate, and the probability of a value greater than a given z-statistic normal variate occurring by chance are listed in Table 4. A correlation coefficient of 0.5 is different from zero at least at the 1% level of significance.

Fig. 10 shows the correlation coefficient versus forecast time for the ground temperature for the US region. Table 5 shows the predictability times for

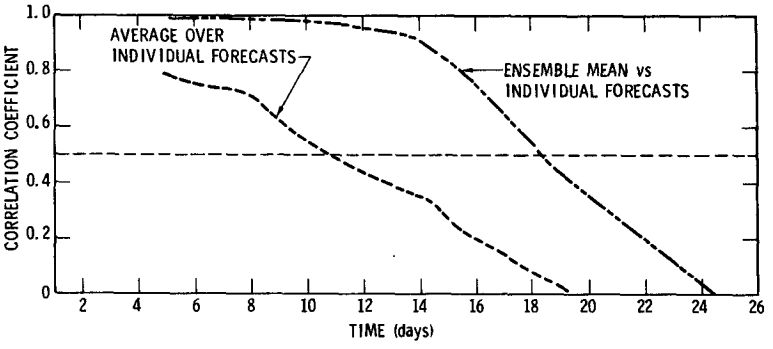

FIG. 10. Effect of ensemble averaging on point forecasts for ground temperature (US region), using z-statistic calculation of correlation coefficient.

TABLE 5. Correlation predictability times (days) for a correlation coefficient of 0.5.

	US (days)	MDW (days)
Point		
Individual	11.0	8.0
Ensemble	18.5	15.5
Space-time		
Individual	19.0	14.0
Ensemble	19.5	16.5

the 0.5 level of correlation. The best improvement, percentage-wise due to ensemble averaging is for the point cases, rather than the space-time averaged cases. Space-time averaging always gave longer predictability times. Its effect was greatest on the individual forecast, with not much effect on the ensemble average.

f. How many forecasts are necessary?

An ensemble of 34 forecasts has been generated and analyzed individually and as an ensemble average. The ensemble average is found to have potentially useful properties. Can a smaller number of forecasts be used to generate an ensemble average forecast which will be as good as the 34-case ensemble average? This is an important question for a practical point of view, because a heavy computational burden is required to produce an ensemble of 34 forecasts. If the more realistic nine-level model had been used, the burden would be unacceptable, and the ensemble averaging technique would have no practical value whatsoever as a technique for the operational forecaster.

We form, successively, ensemble-average forecasts from 2, 4, 8, 16 and 34 cases, the subensembles chosen at random from the total ensemble. Figs. 11 and 12 show the various errors (rms) of the ensemble average versus individual. We see that the 2-case ensemble average fluctuates widely and shows a major departure from the 34-case ensemble average; the 16-case is much closer. The 4-case ensemble average is reasonably good, but departures from the 34-case ensemble average are still evident. A reasonable choice of the number of forecasts would appear to be between four and eight.

4. Summary

The ensemble averaging technique seems to yield significant improvements in the accuracy of intermediate range (under four weeks) forecasts pointwise and over areas of 10^7 km^2 (US) and 4×10^6 km^2 (MDW) and for a time averaging period of five days. The improvement in the point predictability time is of the order of two days for ground temperature and one week for surface pressure. The predictability time averaged over space and time for the United States and Midwest is increased by about three days for ground temperature, but usually decreases for surface pressure.

The phase predictability time defined by the correlation coefficient is roughly 11 days for the US region, for the forecast from the observed initial state (single forecast). The ensemble average increases this phase predictability time to 18.5 days. Similar results obtain percentage-wise, but with somewhat shorter predictability times for the MDW region.

Between four and eight forecasts are necessary

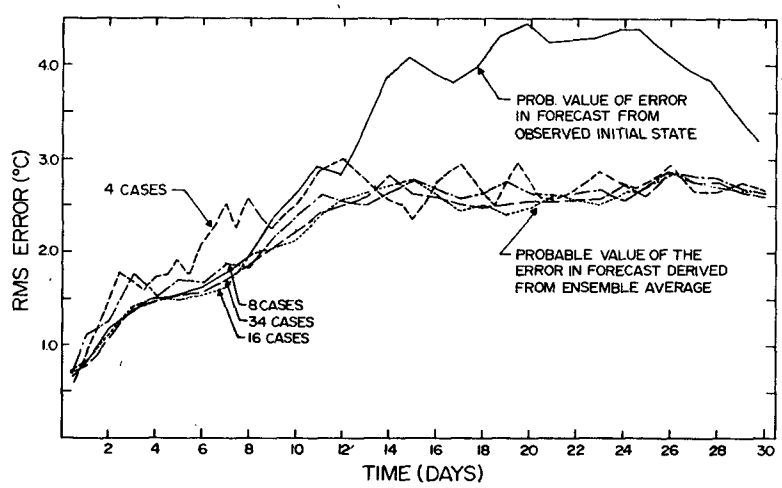

FIG. 11. Error in ground temperature forecast, showing forecast from the observed state and randomly chosen ensemble averages of 4, 8, and 16 cases, and the total 34-case ensemble.

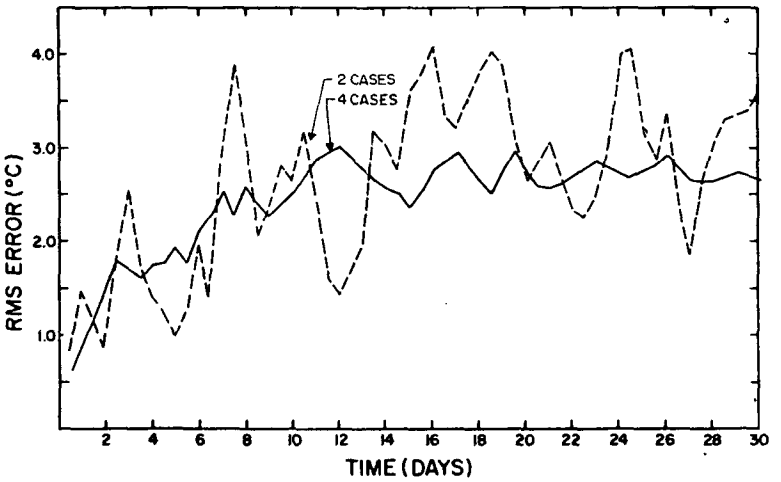

FIG. 12. Error in ground temperature forecast, showing randomly chosen 2-case and 4-case ensemble averages.

to obtain an ensemble sufficiently large to achieve the indicated increases in predictability time.

It should be emphasized that the results obtained here are based on a specific single initial weather situation and on a particular model's response to that situation. A "perfect model" assumption is made, i.e., that the degradation of the forecasts is due to incorrect initial conditions and not to the model. However, the model used is imperfect. The results obtained here are indicative, but care must be exercised not to extrapolate the results beyond the circumstances and assumptions without further investigation.

Acknowledgments. This work was based on the author's doctoral dissertation done at Columbia University. I wish to thank Prof. Robert Jastrow, my dissertation adviser, and Profs. Peter Stone, Jerome Spar and Dr. Milton Halem, for many useful suggestions and discussions. In particular, Prof. Jastrow, as Director of the Goddard Institute for Space Studies, made available computational resources without which this research would not have been carried out.

REFERENCES

Charney, J. G., *et al.*, 1966: The feasibility of a global observation and analysis experiment. *Bull. Amer. Meteor. Soc.*, **47**, 200–220.

Cramer, H., 1946: *Mathematical Methods of Statistics*, Princeton University Press, 575 pp.

Druyan, L., M. Halem and R. Jastrow, 1972: Preliminary studies on long-range forecasts. *Institute for Space Studies: Research Review*, Goddard Space Flight Center, NASA, 223 pp. (see pp. 112–114).

Epstein, E. S., 1969a: The role of initial uncertainties in prediction. *J. Appl. Meteor.*, **8**, 190–198.

——, 1969b: Stochastic dynamic prediction. *Tellus*, **21**, 739–755.

Fisher, R. A., 1958: *Statistical Methods for Research Workers*. Hafner Publishing Co., 356 pp.

Fleming, R. J., 1971a: On stochastic dynamic prediction: I. The energetics of uncertainty and the question of closure. *Mon. Wea. Rev.*, **99**, 851–872.

——, 1971b: On stochastic dynamic prediction: II. Predictability and utility. *Mon. Wea. Rev.*, **99**, 927–938.

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

——, 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.

——, and Kraichnan, R. H., 1972: Predictability of turbulent flows. *J. Atmos. Sci.*, **29**, 1041–1058.

Madden, R. A., 1976: Estimates of the natural variability of time-averaged sea-level pressure. *Mon. Wea. Rev.*, **104**, 942–952.

——, and Shea, D. J., 1978: Estimates of the natural variability of time-averaged temperatures over the United States. *Mon. Wea Rev.*, **106**, 1695–1703.

Pitcher, E. J., 1977: Application of stochastic dynamic prediction to real data. *J. Atmos. Sci.*, **34**, 3–21.

Seidman, A. N., 1976: Numerical experiments on long-range weather prediction. Ph.D. thesis, Columbia University, 216 pp.

Somerville, R. C. J., *et al.*, 1974: The GISS model of the global atmosphere. *J. Atmos. Sci.*, **31**, 84–117.