Estimate of Dynamical Predictability from NMC DERF Experiments

W. Y. CHEN

Climate Analysis Center, NOAA/NWS/NMC, Washington D.C.

(Manuscript received 14 October 1988, in final form 9 January 1989)

ABSTRACT

Experiments on dynamical extended range forecasting (DERF) conducted at the National Meteorological Center (NMC) are utilized to obtain estimates of the upper and lower bounds of dynamical predictability for the NMC operational medium-range forecasting system (MRFS). Owing to the extended range of the integrations (up to 30 days) the upper bound can be estimated without resorting to an empirical model of error growth such as that used by Lorenz (1982). At the root-mean-square error level that equals the climatological standard deviation (about 100 m), the average limit of dynamical predictability is about 14 days, in good agreement with the recent results of Lorenz using the operational forecasting model of the European Centre for Medium Range Weather Forecasts (ECMWF). The current level of forecast skill is also evaluated, indicating loss of useful skill at about 6 days, in agreement with criterion of Saha and Van den Dool. This skill level is much below its potential, thereby leaving considerable room for improvement. The possible contributions to extending the forecast skill from model improvements, as well as initial data error reductions, are also assessed.

1. Introduction

Ever since Richardson (1922) made the first attempt in numerical weather prediction, we have witnessed remarkable advances in this field. The knowledge gained in atmospheric sciences has made numerical models more realistic. The advances in computational and instrumental technologies have made global observation, analysis, and extended-range integration possible. In the early 1970s, experimental forecasts conducted by Miyakoda et al. (1972) showed skillful predictions for up to 3.5 days. A decade later, operational forecasts of the ECMWF have almost doubled this range (Bengtsson 1985). Nowadays, numerical predictions from one to ten days in advance are routinely produced for operational use by NMC, ECMWF, and the United Kingdom Meteorological Office.

While the skill of forecasts has steadily increased and the range of skillful forecasts has been steadily extended, it has also been established that it is impossible to predict the instantaneous state of weather for an extended range. A limit of predictability was first pointed out by Thompson (1957). Lorenz (1963, 1965) subsequently demonstrated that, due to the inherent nature of instability and nonlinearity, atmospheric flows with only slightly different initial states will depart from each other and evolve eventually to flows that are just randomly related. Charney et al. (1966) and Smagorinsky (1969) then used the general circulation models to study the departure of two slightly different initial states. More recently, by using the ECMWF operational forecasts of a 100-day sequence, Lorenz (1982) was able to estimate the upper and lower bounds on the atmospheric predictability of the instantaneous weather patterns. Dalcher and Kalnay (1987) have also used the same ECMWF dataset to conduct error growth and predictability studies. They point out that the commonly used parameter “doubling time of small errors” is not a good measure of error growth because the result is very sensitive to the error growth model used.

In order to have a benchmark dataset for the research community to answer the fundamental questions related to DERF, the National Meteorological Center conducted a very extensive experiment in December 1986. Using this dataset, we follow closely the Lorenz 1982 study and estimate the upper and lower bounds of dynamical predictability for the NMC medium-range forecasting system. Since the DERF integrations are 30 days long, we can follow the growth of error much longer than 10 days and are able to obtain the upper bound of predictability without resorting to an empirical error growth model such as used by Lorenz (1982, 1984) or Dalcher and Kalnay (1987). An initial evaluation of predictability using this DERF dataset can also be found in Roads (1989).

As stressed by Miyakoda et al. (1986), an outstanding problem with extended range numerical forecasts is model climate drift. They report that a substantial portion of the forecast error is due to the model's systematic bias. Adjustment to alleviate the drift problem is also considered in this article.

Corresponding author address: Dr. Wilbur Chen, Climate Analysis Center W/MNC51, National Meteorological Center, Washington, D.C. 20233.
Following a brief description of the data and method of analysis in section 2, the general feature of the error growth up to 30 days is described in section 3. Predictability as a function of the magnitude of the initial data error and an estimate of the upper bound of predictability are obtained in section 4. Adjustment of model climate drift and practical forecast skill are presented in section 5. Possible contributions to dynamical predictability from improvements on the model deficiencies and reductions of initial data errors are discussed in section 6. A summary and a few remarks conclude the article in section 7.

2. Data and methodology

The primary data used in this article are the NMC DERF Phase II 500 mb height forecasts and their corresponding verifying analyses. The DERF dataset is a unique, invaluable series of successive 30-day integrations with initial conditions separated by 24 hours. These runs began on 14 December 1986 and were concluded on 31 March 1987. In addition to these 108 cases of basic integrations, a few small subsets were also generated, with either shorter spacing in the initial conditions or lower model resolutions. In this paper, only the basic dataset of DERF is used. A detailed documentation is described by Tracton et al. (1989). It is sufficient here to say that these integrations were run with NMC’s medium-range global spectral model which has 18 unequally spaced vertical layers and rhomboidal 40 resolution. The NMC MRF model has undergone several improvements over the past few years. The most important changes can be found in White (1988).

We will use the sketch in Fig. 1 to help clarify the notations and definitions. The 30-day integration of interest is represented by the solid line. For illustration, another integration carried out four days before that is also shown, as the dashed line in the same sketch. Following Lorenz’s 1982 study, dynamical predictability will be estimated by comparing two neighboring forecasts for the growth of their difference, assuming the model is perfect. The practical forecast skill will be estimated by comparing the forecasts to the corresponding verifying analysis represented by the solid circles on the extreme left of the sketch. The deficiencies of the model as well as the uncertainties of the initial data will therefore be accounted for in the latter calculation.

The skill of the forecast is conventionally assessed by either the pattern correlation of the geopotential height anomalies (deviations from the climatological normals) or the root-mean-square (rms) error of the geopotential heights between the forecasts and the observations. The former measure is sensitive to the magnitude of the anomaly and the latter tends to give unfair advantage when the fields are smooth. Because of some redundancy between the two (e.g., Chen 1989), and most importantly in order to follow Lorenz’s work closely, only the rms measure will be considered in this paper. The forecast error can be described precisely as follows. Let \( Z \) stand for the 500 mb geopotential height and \( E \) the rms error, then

\[
E_j^2 = \frac{1}{108} \sum_{i=1}^{108} (Z_{ij} - Z_{m,n})^2, \quad \begin{cases} j = 1, 2, 3, \ldots, 30 \\ m = i + j \\ n = 0. \end{cases}
\]

(1)

The first subscript of \( Z \) stands for the case number of the 108 integrations, and the second subscript denotes the forecast range, the 5-day, 10-day forecast, etc. For the growth of difference between two solutions, we let \( D \) stand for the difference and \( l \) the lag between the initial integration time, and define \( D \) as

\[
D_j^2 = \frac{1}{108 - l} \sum_{i=1+l}^{108-l} (Z_{ij} - Z_{m,n})^2, \quad \begin{cases} j = 0, 1, 2, \ldots, (30 - l) \\ m = i - l \\ n = j + l. \end{cases}
\]

(2)
Note that there are only \((108 - l)\) pairs of the equivalent twin experiments. The \(Es\) and \(Ds\) are evaluated at each grid point. They are then spatially averaged. In this article, only the Northern Hemisphere data are considered except in Fig. 11 where the Southern Hemisphere systematic errors are also investigated. The difference in the areal coverage for different latitudes is properly accounted for by applying areal weighting.

3. Growth of difference

In this section we will focus on the growth of rms differences between two neighboring series of forecasts. This should provide a perfect-model error growth due only to the initial uncertainty. Even with a perfect model, a slight initial difference will grow as time goes on due to the dependence of error on instability and nonlinearity. Following Lorenz's approach (1982), we first obtain the error growth between pairs of prognoses with their initial states one day apart. There are 107 pairs from the DERF Phase II experiments. Figure 2 shows the mean and standard deviation of those twin experiments. Since the DERF integrations are 30 days long, we are able to observe the error growth beyond 10 days, unlike those of Lorenz (1982) and Dalcher and Kalnay (1987), and actually determine the time at which the error reaches the saturation level.

From Eq. (2), the saturation level can be precisely defined as the \(D\) with \(m \neq i\) an arbitrary case number. In other words, the saturation level is the rms difference between pairs of forecasts with arbitrary but different initial states. Saturation levels for \(j\) from 0 to 29 days were calculated for this DERF dataset. For small \(j\), the saturation level is about 140 m. When \(j\) increases, its saturation level decreases gradually and settles at about 130 m. The gradual decrease is due to the fact that the MRF model also suffers from the problem of leveling off in its variance as integration progresses—as experienced in most other operational or experimental models. With the saturation level thus defined, it becomes obvious that when a rms difference grows to a size comparable to its saturation level, the forecast of interest is no longer better than an arbitrary prediction. A limit to predictability can therefore be defined. In section 4 we will discuss further the saturation level as well as the limit of predictability.

Note that the error growth rate slows on about the eighth day and gradually assumes another slower linear growth rate to reach saturation. In order to see how stable this kind of statistic is, we divided the DERF Phase II period into three equal time segments, each consisting of 35 pairs, and repeated the error growth calculations. Figure 3 shows the mean for each segment. As expected, the saturation level decreased as the months advanced from December to April and the disturbances decreased in magnitude. However, all three segments show the same feature and reach their own saturation level at about the same time. When normalized with their own saturation value, all three fall more or less into one curve (not shown here). Therefore the statistics appear to be rather stable intraseasonally for this DERF period.

Following Baumhefner 1984, we compare in Fig. 4 the error growth of the smallest eddies, those with the two-dimensional wavenumbers greater than 18, to the total error. The results shown here were normalized with their respective saturation rms error so that both curves saturated at 100%. We see that the small-scale errors (curve b) grow rapidly to reach saturation on about the eighth day, while curve (a), which contains eddies of all sizes, takes another 12 days to reach saturation. The slower growth rate at the later stage for the latter case appears to result from the large-scale
eddy. The recent Monte Carlo forecasting experiments of Schubert and Suarez (1988) show a similar feature.

Stratification of results by latitude bands was also investigated. Figure 5 shows the midlatitude (35 to 55°N) and high-latitude (60 to 80°N) results. When normalized with their own saturation values, these two also fall more or less into one curve (not shown). In the tropics, the geopotential height field does not vary much and is perhaps not suitable for predictability study. Furthermore, as shown in Fig. 6 (for 0 to 20°N) in normalized form, the 1-day rms difference is already about 80% of its saturation level, and, instead of growing, the subsequent rms differences actually shrink in the first few days, reflecting strenuous internal adjustments and suggesting a compelling need for further improvements on MRFS as well as GDAS (the Global Data Assimilation System of NMC). We should therefore avoid using the tropical data in this kind of study until the model is improved to simulate this area more realistically. In the following calculations, therefore, only the latitude band from 20° to 80°N will be considered.

4. Upper bound of predictability

Similar to the way the curve in Fig. 2 was obtained, the error growth for various magnitudes of initial data difference were also obtained. Figure 7 shows the results for lag between pairs from 1 to 4 days. As mentioned in the previous section, a limit to predictability can be defined as the time at which the rms difference reaches its saturation value when two forecasts verifying on the same day cease to resemble each other any more than those being chosen randomly. The limits of predictability (when the rms difference reaches 130 m) for various initial error magnitudes are marked with a cross in the figure and listed in Table 1. Plotting the initial data error against the limit of predictability, as shown in Fig. 8, we see a rather linear relationship. Note that at 130 m the error is already saturated so that the limit is zero. That is how we also mark a solid circle at that coordinate. Since the relationship appears to be linear, it is reasonable to extrapolate the relationship linearly to near zero initial data error. In doing so, we arrive at a limit of predictability of about 24 days for an infinitesimal initial data difference. Note that the conceptual infinitesimal initial data difference here refers to the error in those grid-point network resolvable scales. Errors due to the unresolved scale will always be present even if the error of resolved scale was nil. The above estimate of predictability is in excellent
agreement with other recent studies, such as Dalcher and Kalnay (1987) and Schubert and Suarez (1988).

When all of the curves of Fig. 7 are composited into one panel as shown in Fig. 9, we acquire better insight as to the shape as well as the position of each curve. Knowing that the limit of predictability for the infinitesimal initial data error is about 24 days (position of each end) and knowing also the general shape, we can reasonably extrapolate those results to obtain the error growth curve for the infinitesimal initial data error, shown in the figure as the bold faced curve. The upper bound of predictability for the NMC MRF model is therefore obtained. The area to the right of the curve represents the unattainable domain, in an ensemble sense. At the rms difference level that equals the climatological standard deviation (about 100 m), the average limit of the predictability is about 14 days.

5. Lower bound of predictability

So far we have assumed a perfect model and compared the neighboring forecasts for the growth of their difference. In this section we will deal with the forecast skill by verifying the forecasts against the corresponding verifying analyses. In a recent investigation regarding
the feasibility of one-month forecasts based on a dynamical approach, Miyakoda et al. (1986) point out that it is essential to remove the systematic drift of the prediction model in order to raise the skill of forecast for the extended range. Like most of the general circulation models (e.g., Hollingsworth et al. 1980; Derome 1981; Bengtsson and Lange 1981) the NMC medium-range forecasting system also suffers from a pronounced systematic cooling of the model atmosphere. For example, the 108-case mean geopotential height of the 15-day forecasts, their corresponding verifying analyses, and the difference are displayed in Fig. 10. Although the mean of the forecasts (10a) shows certain similarities to that of the observed (10b), we see an overwhelming drop of height over most of the hemi-

![Graph](image-url)

**Fig. 9.** Extrapolation of results of Fig. 7 to yield an estimate of upper bound of predictability for an infinitesimal initial data error, shown here as the bold faced curve.

![Graph](image-url)

**Fig. 8.** Initial data error versus limit of predictability. The dashed line is the extrapolation of the solid line which is a linear fitting of those solid circles. The intercept at infinitesimal initial data error represents an estimate of the ultimate limit of predictability.

![Graph](image-url)

**Fig. 10.** (a) DERF Phase II 108 cases mean height for the 15-day forecasts, (b) mean of their corresponding verifying analyses, (c) the difference between the two. The dashed contours denote underforecast in height.

sphere (10c). The severity of the drift can be seen in Fig. 11, which shows the zonal averages of the mean height error from the South to North Pole and a forecast range up to 30 days. For the area north of 50°N, the
drop in the mean height does not appear to be abating even after 30 days.

In forecast skill assessment, the systematic error must be accounted for. According to Epstein (1985), empirical correction of error has been practiced even for the short-range forecasts (Glahn and Lowry 1972; Glahn 1980; Hughes 1982). The causes of the systematic drift are subjects of recent intensive research. Investigations from many angles have been conducted (e.g., Wallace et al. 1983; Burridge and Sadourny 1982; Arpe and Klinker 1986; Palmer et al. 1986). Miyakoda and Sirutis (1984) further looked into the effects of parameterizations for cumulus convection, orographic representation, boundary layer fluxes, and turbulent processes. Options other than the GCM approach were also considered (Miyakoda and Chao 1982). The basic approach to remove the systematic error may take many more years of intensive effort. For a temporary remedy, Shukla (1983) suggests that the drift be simply adjusted by the difference between the model climatology and the real climatology. Following this suggestion, the mean error of the DERF period is obtained for each forecast length, and each individual forecast is then adjusted by that particular mean error, i.e.,

$$m_j = \frac{1}{108} \sum_{i=1}^{108} (F_{ij} - O_{ij})$$

$$aF_{ij} = F_{ij} - m_j$$

where $F$ stands for forecast, $O$ observation (analysis), $m$ mean error, and the subscript $i$ denotes the case number of the 108 integrations, $j$ the forecast length, and $a$ the adjusted forecast. Since the results in this section are based on the adjusted forecast, the subscript $a$ is dropped for simplicity of presentation. We note that the above climate drift adjustment is not operationally possible since $m_j$ can only be obtained after the entire experiment is completed. The systematic drift of the model is found to be mainly in the wavenumber zero component. We have compared the adjustment by Eq. (4) to that by simply suppressing the wavenumber zero component, which is operationally possible and simple to carry out, and found the results to be surprisingly similar. For practical purposes, the climate drift adjustment by the latter procedure appears to be acceptable.

Figure 12 displays the forecast skill of the NMC MRF (86) model up to 30 days in terms of rms errors. The dashed curves represent one standard deviation from the 108-case mean. We see that the rms error grows rapidly, reaching 85% of its saturation value by the 8th day and then, with rapidly decreasing growth rate, gradually approaches its saturation value on about the 14th day. Similar to the earlier definition, the saturation level is defined precisely from Eq. (1) by taking $m$ arbitrary but not $i$ itself. The saturation levels for various $j$ are calculated and shown in Fig. 13. When the error of a forecast reaches the level of that in Fig. 13 (about 140 m), it is not better than an arbitrary forecast and there is no skill whatsoever in that prediction. Note that the spread in Fig. 13 decreases with the forecast lead time. It implies a certain loss of variability in the model as the forecast time increases.

When the rms error grows to be as large as the climatological standard deviation (about 100 m), the forecast is then considered to contain little useful skill. We see from Fig. 12 that the average limit of forecasts that contain useful skill is only about 6 days, in good agreement with Saha and Van den Dool’s assessment (1988).

The rms errors discussed above are those adjusted for the model climate drift. The unadjusted ones are shown in Fig. 14 and compared to those adjusted ones. The unadjusted total error keeps increasing in mag-

![Fig. 11. The zonal averages of the mean height error for forecasts of various ranges, from 5 to 30 days.](image-url)

![Fig. 12. The forecast skill in terms of rms error of the NMC MRF (86) model. The mean errors obtained by Eq. (3) were subtracted from each forecast before the calculation of the rms error. The dashed curves represent one standard deviation from the 108-case mean.](image-url)
Fig. 13. Similar to those of Fig. 12, except the verifying analysis was randomly picked. The rms errors here thus represent those from arbitrary prognoses.

Fig. 14. Comparison of rms errors calculated with adjustment of the model climate drift (a) and without adjustment (b).

6. Potential skill improvements

The skill level obtained (Fig. 12) is much below its potential (Fig. 9 bold faced curve), so there exists considerable room for further improvement. In this section we attempt to estimate the contributions that reduction of initial data uncertainty and improvement in model deficiency can possibly make to extend the predictability. For this purpose, we plot the upper bound of predictability and the practical forecast skill curve, as shown in Fig. 15 (curves c and a, respectively). Starting at D, the current average magnitude of the 1-day apart

\[ u E^2 j = a E^2 j + m^2 j. \]

Fig. 15. (a) The lower bound of predictability that represents the current level of forecast skill; (b) and (b') the predictability curve for an average 1-day apart data difference of magnitude D and D', respectively; (c) the upper bound of predictability, for an infinitesimal initial resolvable data error. The A represents the current average level of skill of a 6-day forecast; B and B' the forecast length with the same level of skill if the model were perfect but the one-day apart data difference is of magnitude D and D', respectively; C the forecast length if the model were perfect and the initial resolvable data error were infinitesimally small. The area to the right of curve c represents the unobtainable domain.

rms difference, we extrapolate to obtain the equivalent predictability curve for this initial uncertainty, shown as curve b. It can be seen that if we keep improving the model until it is perfect, we might be able to obtain 10-day forecasts as skillful as the current 6-day forecasts (from A to B at the climatological error level). In addition, if we are able to reduce the initial data error to D', another predictability curve b' could be obtained and we would add another 2 days to about 12 days (A to B to B'). By reducing the data error from D to D' alone, an imperfect prediction model will yield a curve nearly parallel to curve a, starting from D'. The resultant extension of skillful forecast may be less than those of curve b except for the first few days, but can still be very significant. The above partitioning of errors into those due to model imperfections and those from observation/analysis is only intended as a conceptual exercise. In the real MRFS the initial data error is critically dependent on the forecast model in the data assimilation cycle. The two cannot actually be separated. Nevertheless, the exercise does illustrate that both reduction of data error as well as model improvement are urgently needed, especially for the extended-range forecasts.

7. Summary and concluding remarks

In this paper we have obtained the upper and lower bounds of the dynamical predictability for the NMC operational medium-range forecasting system (MRFS). The data used are the NMC DERF Phase II 500 mb
height forecasts and their corresponding verifying analyses. Owing to the extended range of the integration, the upper bound of predictability can be estimated without resorting to an empirical model of error growth such as used by Lorenz (1982, 1984) or Dalcher and Kalnay (1987).

Following Lorenz' approach (1982), we first evaluate the error growth between pairs of prognoses with their initial states one day apart. The subsequent rms differences grow steadily for about a week before slowing down on about the 8th day and gradually approach saturation on about the 20th day. The rms differences for the small-scale eddies show growth with essentially the same rate to reach saturation. Therefore, the later stage slower growth rate appears to result mainly from the large-scale eddies. We divide the DERF Phase II period into three equal time segments and repeat the error growth calculations. All three show essentially the same feature, suggesting that the statistics obtained are rather stable intraseasonally.

The limits of dynamical predictability (defined here as the time the rms difference reaches its saturation level) for various magnitudes of initial data error are then determined. Their relationship appears to be linear. Based on this feature, we obtain the upper bound of predictability for an infinitesimal initial resolvable data error. The lower bound of predictability is then obtained with proper adjustment for the systematic drift of the model atmosphere. The skill level obtained is much below its potential, leaving considerable room for further improvements. An attempt is made to estimate the possible contribution to extending the predictability by reducing the initial data uncertainty and improving the model. At the rms error level that equals the climatological standard deviation (about 100 m), the results indicate that if we keep improving the model until it is perfect, we might be able to obtain 10-day forecasts as skillful as the current 6-day forecasts. In addition, if we were able to reduce the initial resolvable data errors to very small values, we could stretch another four days to about 14 days. These estimates appear to be in excellent agreement with the recent assessment of Lorenz (1984) using the ECMWF model data. The useful forecast intervals for time averages (not investigated in this article) are reported to be longer than those of instantaneous events (Roads 1986). The predictability is also known to depend on flow regimes (e.g., Palmer 1988) and long-lived anomalous forcings (e.g., Shukla 1983). Due to the limited scope of this paper, these latter issues have not received their deserved attention.

Acknowledgments. Without the dedication of many NMC staff members to produce this massive DERF experiment, this article would not have been possible. Particular thanks goes to Robert Kister, who provided me with the condensed dataset. I am also grateful to Drs. Huug van den Dool and Steven Tracton for their reviews of this manuscript and valuable comments. Ms. Bessie Bando and Mr. Michael Gaidurgis provided their expert assistance in manuscript preparation.

REFERENCES


Epstein, E. S., 1985: Procedure for the statistical correction of medium-range forecast spectra. NMC Office Note No. 301. [Available at NMC, Washington, D.C. 20233.]


Hughes, F. D., 1982: An objective method to modify NMC/NWP model mean sea level progs. NMC Office Note No. 259. [Available at NMC, World Weather Building, 5200 Auth Road, Washington, DC 20233.]


