A New Look at Weather Forecasting through Analogues

H. M. VAN DEN DOOL

Cooperative Institute for Climate Studies, Department of Meteorology, University of Maryland, College Park, Maryland

(Manuscript received 28 November 1988, in final form 11 May 1989)

ABSTRACT

In the literature, the use of analogues for short-range weather forecasting has practically been discarded. This is because no good matches for today's extratropical large-scale flow patterns can be found in a 30-year data library. We propose here a limited-area model approach for Analogue-Forecasting (AF). In order to make a 12-hour AF valid at a target point, we require analogy in initial states only over a circle with radius of about 900 km. On a limited area there are usually several good analogues, sometimes to within observational error. Different historical analogues may be used at different target points.

The usefulness of the limited area approach is first demonstrated with some examples. We then present verification statistics of 3000 12-hour 500-mb height *point* forecasts in the Northern Hemisphere winter at 38°N, 80°W (over West Virginia, U.S.A.). In order to beat persistence at 12 hours at this point we need an analogue which differs by about 40 geopotential meters or less from the base case. This requirement is met almost all of the time using a 15-year dataset for analogue searching. We find a few percent of the analogue pairs to be within observational error. In the mean, over 3000 cases the initial discrepancy is 33 gpm. When averaging over the first five analogues 12-hour AF over the eastern United States can be characterized by a 52 gpm rms error and 0.95 (0.77) anomaly (tendency) correlation. The forecasts have the correct amplitude, i.e., no damping, in spite of the averaging over five individual forecasts. We then show an example of a 500-mb height forecast *map* on a (roughly) 2000 × 2000 km area over the eastern part of North America. Although different analogues were used to arrive at the 12 hour forecast at each of the 25 gridpoints, the resulting map looks meteorological and the forecast is moderately successful. A verification of a large set of 12-hour forecast maps shows that the height gradients are indeed forecast with some skill. We then proceed to make 24-hour *point* forecasts by finding historical limited-area matches to the 12-hour forecast maps. This second time step indicates that the AF-process holds up, with forecast accuracy increasing its margin over persistence.

Two applications are discussed. Comparing initial and 12-hour forecast error tells us something about "error growth" and predictability at that spot according to a perfect model. Given that there are usually several good analogues, Monte Carlo experiments and probabilistic forecasts naturally suggest themselves. In particular we find the spread of analogue forecasts to be an eminent predictor of forecast skill.

Refinements and applications and extension to longer range forecasts are discussed in the final section.

1. Introduction

The idea that analogues can be used for forecasting future weather is indeed very old. The basis for analogue forecasting [as well as for numerical weather prediction (NWP)] is a very powerful one, namely that if two atmospheric states are very close initially, they will remain somewhat close for some time in the future.

The problem with analogue forecasting (AF) has been that we do not seem to be able to identify any states in the past that can be considered good matches to the present large-scale flow pattern. The very best pair of analogues found by Ruosteenoja (1988) in a very large dataset of extratropical Northern Hemispheric 500 mb heights are as far apart as a 4-day forecast made by a state-of-the-art NWP model is from its verifying analysis [i.e., approximately 65 gpm root-mean-square (rms) difference]. For most flow patterns observed in the last 40 years, the best "analogue" is much worse than that, i.e., close to 100 gpm difference. Lorenz (1969) came to the same conclusion using only five years of data: that the best analogue is at best only mediocre. Both Ruosteenoja and Lorenz quote an astronomical number of years that we have to wait until we can expect to find two atmospheric states that are within present-day observational error (i.e., 10–20 meters rms difference for 500 mb height). Gutzler and Shukla (1984) arrived at essentially the same results. Therefore, the reservations expressed by Namias (1951) seem to remain in effect as long as our historical records are "short", i.e., short compared to "billions of years" (Ruosteenoja 1988).

Starting from two atmospheric states that are initially not all that close, the AF method right from the start is at a great disadvantage relative to NWP and to persistence which have small (currently about 15 gpm in 500 mb height) and zero initial error, respectively. This

Corresponding author address: Dr. H. M. van den Dool, Department of Meteorology, University of Maryland, College Park, MD 20742.

disadvantage is so large that any of the intrinsic advantages that AF (a perfect model) may have, seem completely irrelevant in practical forecasting. Ruosteenoja summarizes the issue by concluding emphatically that "hemispheric 500 mb height analogues are practically useless in weather forecasting." It is true that, except for long-range forecasting where NWP is less of a fierce competitor (Livezey and Barnston 1988, plus references therein), AF is no longer used at present to forecast future atmospheric states.

Here we will make yet another attempt at AF. This is done not necessarily to beat NWP as it stands today, but primarily to show that we can do a far better job with the AF method than suggested in the literature. Improved AF might help provide insight in the workings of the atmosphere, and it may help readdress predictability questions discussed originally by Lorenz (1969). Practical applications such as the forecast of forecast skill (Kalnay and Dalcher 1987) also seem possible.

Our optimism to make this new investigation is based on the following considerations:

(i) In order to make a 12-hour forecast (or more generally a forward time step Δt) we do not need analogy over the entire Northern Hemisphere; a limited-area analogue will do. Although it is true that the future of the atmosphere at any given point ultimately depends on the current state of the atmosphere (and ocean, etc.) at all other points, we make explicit use here of the fact that tendencies are determined primarily by quasi-local processes (advection, radiation, etc.). After all, in NWP, limited-area models (LAM) have proven to be successful for short-range forecasts within the limited area.

(ii) Given the success of LAM based on just the barotropic vorticity equation in the 1950s, we do not need analogy in all variables at all levels either. Here we will use 500 mb height, which is the best level, on average, to apply the barotropic model (Berggren 1958).

(iii) Given that the governing differential equations are first order in time, we need analogy at one time level only. In some previous studies (Shabbar and Knox 1986; Dunn 1951) the historical development of the analogue was considered important, thus reducing the chances of finding good analogues.

In order to make a 12-hour forecast of 500 mb height at a gridpoint n, we will first draw a circle of radius r around that point. The radius has to be small enough to allow us to find good analogues and yet large enough to allow us to make a 12-hour forecast for the target point n. The result of this procedure is no more than a 12-h 500 mb height *point* forecast. In much of this paper we will verify point forecasts of height, so as to establish whether the first and fundamental step of AF has any skill over simple control forecasts such as persistence. It is possible to apply AF to a large number of adjacent gridpoints. For each point n a different historical analogue may, and often will, be found/used so that at the end we would have to patch point forecasts together in order to arrive at a forecast map. The spatial consistency of such maps is of some concern and will be investigated here.

Over a small area (at one level, for one variable) it is easy to find good analogues even if the dataset available for analogue search is short. This is a simple matter of the spatial degrees of freedom involved (Gutzler and Shukla 1984; Ruosteenoja 1988). In the extreme case of a single variable at a single point, one can easily find several perfect historical analogues. However, these are not necessarily useful for forecasting because the atmosphere advects information rapidly. So our task is to find good analogues over a small area (so that the initial rms differences are small). This small area, however, must be large enough to make it impossible for boundary and more remote effects to travel to the target gridpoint in a period Δt (which will be a 12-hour period here out of necessity). There is no a priori guarantee that this is always possible with only a few decades of data available, but as it turns out, there is rarely a lack of good analogues.

A 12-hour forecast for each point of interest is only the first step of a larger process. Once the 12-hour forecast map is available we can search for analogues that match the 12-hour forecast map so as to take the next "time step." More generally we would like to present our procedure as a method to integrate from t to $t + \Delta t$. Starting from analyses and presenting results primarily pertaining to the first 12-hour time step may give the impression that we are in the business of short-range forecasting only. But we will also present preliminary results of the second 12-hour timestep so as to emphasize that nothing in the AF procedure, although working on small areas, limits its application to the very short-range forecasts alone. *Changing analogues from point to point in space (as well as in time) is the major feature that distinguishes our effort from all previous studies on AF. In fact, the AF procedure technically becomes very much like NWP, except that tendencies are determined empirically from a data library, rather than from a set of discretized approximate equations believed to govern the atmosphere.*

In section 2 we describe the data, terminology and the analogue-searching method. In section 3 the main body of results will be presented. First, we discuss some examples of the AF method (3a, 3b). Then an extensive verification of 3000 12-h 500 mb height point forecasts in winter for 38°N, 80°W (§ 3c). In section 3d we give a detailed discussion of an example forecast map for eastern North America, followed in section 3e by a verification of 15 such maps. This includes verification of gradients of the height forecast, so as to address the question whether patching the point forecasts leads to acceptable spatial consistency. Finally, in section 3f, fifteen 24-hour forecasts will be discussed. Applications will be discussed in section 4. This includes a discussion

of atmospheric predictability, using the atmosphere as its own model, forecasting forecast skill, etc. Conclusions and discussion are in section 5. Finally, in appendices, we speculate about the governing equations of AF and compare the skill of AF to that of NWP.

2. Data, terminology and analogue-searching

The data used here consist of twice-daily 500 mb height analyses from the National Meteorological Center, Washington, DC, on a 4° × 5° latitude–longitude grid over the area 18°N to the North Pole and around the earth longitudinally. The period is 15 years, from 1963 to 1977. A few bad and missing data (about 1.1%) have been replaced through linear interpolation, so that the resulting dataset has no gaps. The corrections were made by B. Doty and K. Mo, and this dataset has been used in Gutzler and Shukla (1984) (only 0000 UTC data) and in many other studies.

The process of finding analogues and using them in forecasting is almost self-evident. However, to define the terminology we present the process symbolically in Fig. 1. Horizontally we have plotted time t, representing month, day and hour (0000 UTC or 1200 UTC), while the vertical axis refers to the year (1963–77). The *B*ase is the situation for which the analogues are sought; operationally this would be the 500 mb height map valid at the present time, but in research mode the Base could be anywhere in the 15-year dataset. The *A*nalogue (A) is the best match found in some year at some time. The *V*erification and the *F*orecast are valid 12 hours after B and A, respectively. If necessary more than one Analogue (A_1, A_2, . . .) and subsequent Forecasts (F_1, F_2, . . .) can be considered. The initial "error" of AF can be measured by $B - A$ (segment 1 of Fig. 1) while the forecast accuracy is measured by $F - V$ (segment 3) which should be compared to *P*ersistence $V - B$ (segment 2). By drawing segments 1 and 3 slanted we express that the times t (of the Base) and t' (of the analogue) are generally not the same.

Analogues are sought here by calculating the rms difference between the Base height field (Z^B) and the height field of a Candidate Analogue (Z^{CA}) over all co-located gridpoints i at or within a circle of radius r around target point n. The quality (Q) of the Candidate Analogue is defined by

$$Q = \left\{ \frac{1}{N} \sum_i (Z_i^B(t, j_B) - Z_i^{CA}(t', j_{CA}))^2 \right\}^{1/2} \quad (1)$$

where Z denotes 500 mb heights, j is the year index, and t and t' denote the time of year ($|t - t'| <$ about 1 month); N is the number of gridpoints within distance r from target point n. By limiting $|t - t'| <$ about 1 month, we restrict ourselves to picking analogues from about 1800 candidates in a 15-year dataset. The probability of finding worthwhile analogues outside the one month window is small (Ruosteenoja 1988). In Fig. 2, an example target point is indicated by an asterisk (at 38°N, 80°W). The other gridpoints are labeled by their distance to the target point. Distance r is measured in units of grid-spacing, which is 444 km in the meridional direction; for convenience we ignore that the grid-spacing in the zonal direction is slightly different. Note that we use a regular grid only for convenience. The AF procedure could be applied equally well to all observed height values (wherever they are) within distance r from the target point (an observational site in that case).

For a given Base there are about 1800 (twice a day × 60 days × 15 years) Candidate Analogues. So in all 1800 Q in (1) have to be calculated for a given target point n and given r. The resulting Q are subsequently ranked from low to high. The lowest Q corresponds to the best analogue (A_1) etc. We realize that techniques other than rms difference should also be considered for matching two height fields, but for the sake of simplicity we will present in this paper results based on (1) only. In Eq. (1) we permit analogues to be found in the base-year as well as in earlier and later years (in research mode this can be done). Note also that we calculate the rms difference over geographically matching gridpoints. The idea of shifting the area covered by CA relative to the Base could have some merit, but is not yet considered. By choosing a circle rather than some oval shape around the target point we ignore, for the time being, the fact that midlatitude weather generally travels from west-to-east; flow dependent refinements are possible but are not considered.

It should be emphasized once more that even though we search for analogues over a circular area, we verify the forecast only at the target gridpoint. This vital aspect distinguishes our effort from, for example, Gutzler and Shukla's (1984) approach to AF on "small" (in their case still continental sized) areas. They verified 1-day forecasts produced by AF over an area identical to the analogue search area, thus allowing large errors from the unmatched exterior to propagate into the verification area.

We check for analogy in 500 mb height only. The least we can hope for is that analogy in heights guarantees the advection of vorticity to be similar. Since

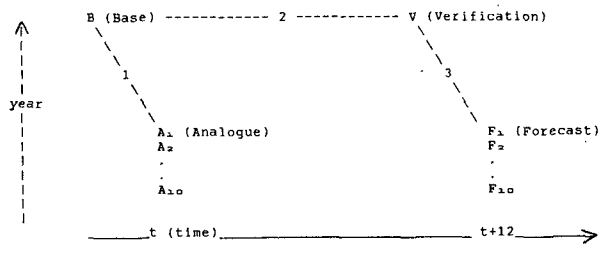

FIG. 1. An outline of the terminology in the Analogue Forecasting process. B, A, V and F represent Base, Analogue, Verification and Forecast. For further description, see section 2.

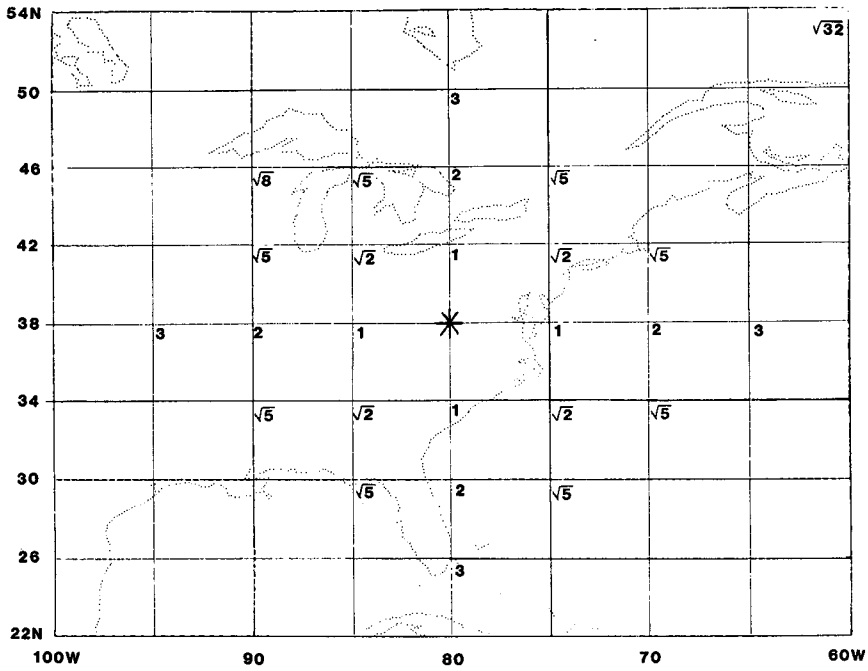

FIG. 2. A 9 × 9 portion of a 4° latitude by 5° longitude Northern Hemisphere grid. The target point near Charleston, West Virginia, is indicated by an asterisk. Surrounding gridpoints are marked by their distance r (in units of grid distance) from the target point. The interior 5 × 5 grid used later on for forecast maps consists of all gridpoints with $r \leq \sqrt{8}$.

advection is known to be a major component of the tendencies at 500 mb (where divergence is normally at a minimum) we should be able to make "barotropic" AF with some degree of success (see Appendix A for discussion on this point).

Finally, we point out that analogy over a small area does not imply that we have matched small spatial scales only. From many studies (example: Blackmon 1976) we know that the bulk of the height variance resides in the long waves. Therefore the position of the long waves is critically important.

3. Results

a. *Finding analogues*

We will first discuss an example here. The target point is 38°N, 80°W (see Fig. 2) and we choose an arbitrary date 0000 UTC 30 January 1966 which we call the *B*ase (see Fig. 1). Analogues were then selected from all cases with a date between 0000 UTC 5 January and 0000 UTC 25 February from the 15-year dataset. Table 1 shows a list of the 12 best analogues for $r = 4$, 3, 2. In this particular case we do not find truly good analogues when we take a large circle, $r = 4$. It has been the experience of many in the past that if there are no truly good analogues, the neighbors in time to the base date (i.e., 0000 UTC 29 January 1966 and 1200 UTC 30 January 1966) will show up highest on the list. Indeed this is the case for $r = 4$ and larger r. If for $r = 4$, one does not remove the neighbors in time from consideration, the forecast automatically becomes persistence, i.e., B itself is the forecast. If for $r = 4$ one does remove neighbors in time (as is usually done), AF automatically degrades to a forecast system worse than persistence simply because the initial error of the next best analogue is larger than the 12 h persistence error. However, as can be seen in Table 1, for $r = 3$ there are already several analogues better than the neighbors in time and for $r = 2$ the neighbors in time have disappeared from the top-12 list. In fact the 12 best analogues are from a variety of years other than 1966, the base year, and they are not clustering in time as they would for larger r (another well-known problem of AF so far). Hence for $r = 2$ we may have a chance to beat persistence. Persistence may seem a weak competitor, but as a control it is of fundamental importance because only forecasts better than persistence have skill in the forecast of the time derivative—the essence of forecasting.

For $r = 2$ the best analogue in Table 1 has an rms difference $(A - B)$ of 47 meters, probably about twice the uncertainty in analyzed gridded 500 mb height fields over the eastern United States in the 1963–77 period. Given that in this area and in this season the standard deviation of 500 mb height fields is about 170 meters (varying in space), one would expect that two randomly chosen cases would be about $170\sqrt{2}$ (=239)

TABLE 1. A list of the 12 best analogues for $r = 4$ (top), $r = 3$ (middle) and $r = 2$ (bottom). The base date is 30 January 1966 at 0000 UTC and the target point is 38°N, 80°W. The unit for rms difference is geopotential meters.

	Year	Month	Day	Hour	Rms
	1966	1	30	0	0.0
	1966	1	29	12	80.1
	1964	2	19	12	84.3
	1969	2	12	0	91.1
	1970	1	7	12	92.1
	1963	2	13	12	93.6
$r = 4$	1966	1	30	12	94.2
	1972	2	19	0	94.3
	1964	2	19	0	94.4
	1966	1	23	12	100.0
	1977	1	25	12	102.9
	1963	2	13	0	103.9
	1964	2	8	0	104.6
	1966	1	30	0	0.0
	1971	2	9	12	61.2
	1970	1	7	12	67.8
	1977	1	25	12	71.7
	1977	1	10	12	73.9
	1975	1	26	0	78.1
$r = 3$	1964	2	8	0	79.7
	1977	2	20	12	82.9
	1971	2	9	0	86.8
	1963	2	13	12	87.2
	1963	2	13	0	87.6
	1972	2	19	0	90.7
	1966	1	29	12	91.0
	1966	1	30	0	0.0
	1970	1	7	12	47.2
	1975	1	26	0	58.1
	1977	1	25	12	66.0
	1971	2	9	12	67.0
	1977	1	10	12	70.9
$r = 2$	1964	2	8	0	72.1
	1969	2	12	0	78.3
	1969	1	7	0	78.4
	1971	1	26	12	79.7
	1967	2	23	12	79.8
	1977	2	20	12	80.1
	1970	1	8	0	80.9

meters apart on the average. So a 47 meter difference does indeed imply a great deal of similarity. Measured by the anomaly correlation, the $B - C$ and $A - C$ fields ($r \leqslant 2$) are correlated by 0.986, where C is the climatological 500 mb height field. [Later we shall see that the arbitrarily chosen 0000 UTC 30 January 1966 case is not exceptionally good with regard to initial error ($B - A$).] A 47 meter difference in this area of high variability is much much better than Ruosteenoja's very best case, i.e., his initial error of 65 gpm against the background of a hemispheric saturation difference of 140 gpm. There is, in fact, rarely a lack of moderate to good analogues for $r = 2$.

Collectively the listings in Table 1 for $r = 3$ and $r = 2$ display increasingly better analogues relative to those found for $r = 4$ (the 12th best analogue for $r = 2$ is as good as the very best analogue for $r = 4$). Of course in itself this does not prove that 12-hour forecasts through AF are feasible for $r = 2$. But on the other hand, large-scale weather (fronts, troughs and ridges at 500 mb) rarely travels faster than 900 km in 12 hours. Thus we anticipate that $r = 2$ can nearly always be used and in cases of weak advection, perhaps $r = \sqrt{2}$ or even $r = 1$ [i.e., just five grid-points used to determine Q in (1)]. It is extremely important to have data twice a day (at least) as it allows us to use smaller circles than would be required for data once a day only.

b. Pilot 12-hour 500 mb height point forecast using AF

Figure 3 shows the results of a 12-hour forecast verifying at 1200 UTC 30 January 1966 at the target point (38°N, 80°W). As a function of 'r' we show: 1) the initial error ($B - A$), 2) the forecast error ($F - V$) and 3) persistence ($V - B$). The latter is obviously the same for all values of r. The results in this example are obtained by combining verification scores over the best 10 analogues (excluding neighbors in time for $r = 4$ and 3) in the following manner:

$$\text{initial error} = \left\{\frac{1}{10}\sum_{i=1}^{10}(B - A_i)^2\right\}^{1/2}$$

where A_i represents the height of the ith analogue at the target point, and

$$\text{forecast error} = \left\{\frac{1}{10}\sum_{i=1}^{10}(F_i - V)^2\right\}^{1/2}$$

where F_i is the forecast height according to the ith best analogue at the target point.

Figure 3 contains qualitatively almost the entire intent of this paper. It shows that:

(i) when only poor quality analogues are available ($r \geqslant 4$; large initial error), the 12-h forecast is equal or worse than persistence in accuracy.

(ii) when the circle around the gridpoint is chosen too small, say $r = 0$ (i.e., the target point only), perfect analogues can be found, but the forecasts are a disaster,

(iii) somewhere in between, miraculously, there is an area ($r = 1$ to 3) where good (but by no means perfect) analogues can be found, and where 500 mb forecasts better than persistence can be made in the first 12-hour time step. Within the range of intermediate r values the error growth is considerable though. In 12 hours the initial mean-square error has grown in this case by a factor of 2.5 to become the forecast error.

Note that the "initial error" in Fig. 3, evaluated at the target point only, is smaller than the values shown in Table 1 which are rms differences over the circular area. Apparently the similarity is automatically better at the center of the circle. There is an ambiguity whether we should calculate the initial error at the target point only, or over the circle used in the analogue-search

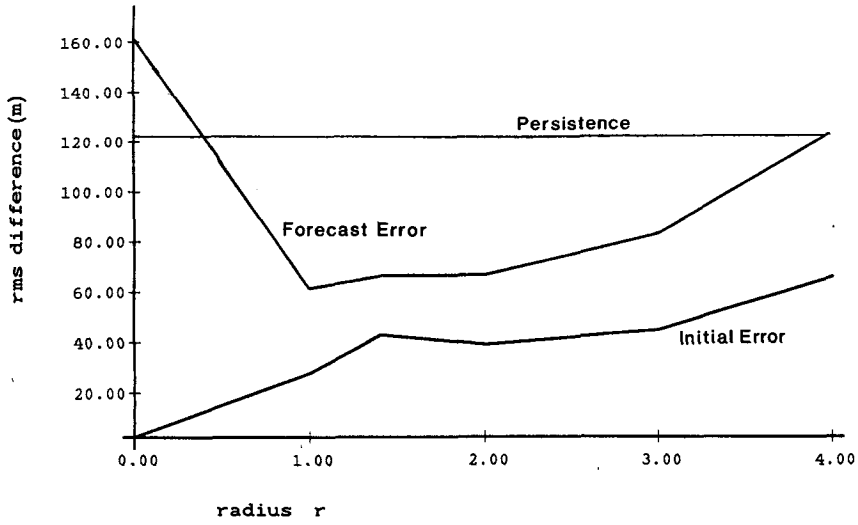

FIG. 3. Verification scores of 12 hour forecasts valid for 1200 UTC 30 January 1966 at 38°N, 80°W as a function of the radius r. The root-mean-square (rms) values for forecast and initial error are taken over the ten best analogues.

[i.e., Q as in (1)]. We will use the initial error at the target point in this paper.

Figure 4 is the same as Fig. 3 but for two other arbitrarily chosen dates. In both cases we see analogues of better quality than in Fig. 3, i.e., much smaller initial error for $r = 2$. Forecasts are again considerably better than persistence for the 1 February 1974 case, and the error growth is much smaller now than in Fig. 3. However, the range of r values over which we have prediction capability is narrower, which could be caused by excessive advective winds in two out of the ten best analogues. The very best forecasts, in absolute terms, are for 0000 UTC 31 January 1964 (Fig. 4b). However, in this case the flow did not change much in 12 hours, a no-win situation for any forecast model (or forecaster) in terms of outperforming persistence.

c. *Extensive verification of point forecasts at 38°N, 80°W*

We proceeded to make forecasts verifying 12 hours after the following 20 base dates: 0000 and 1200 UTC 27 January . . . 1200 UTC 5 February for all 15 years of the dataset. In all, 300 base cases were used, covering a large spectrum of synoptic flows during winter. In each of the 300 cases, we made 10 forecasts (based on the 10 best analogues, excluding neighbors in time), so that the statistics presented here are based on 3000 instantaneous flow forecasts. We kept $r = 2$ in all cases, the target point was always at 38°N, 80°W and for all bases, the window was from 0000 UTC 27 December through 0000 UTC 5 March. The 3000 cases were stratified according to their quality, see Eq. (1) i.e., $0 < Q < 10$, $10 < Q < 20$, etc. Within each of these bins we calculated the initial error at the target point as

$$\text{initial error} = \left\{\frac{1}{N}\sum_i (B_i - A_i)^2\right\}^{1/2}$$

and similarly the forecast error as

$$\text{forecast error} = \left\{\frac{1}{N}\sum_i (F_i - V_i)^2\right\}^{1/2}$$

In Fig. 5a, we show the forecast error as a function of the initial error. The horizontal line at 80.3 meters signifies the mean error of a 12 hour persistence forecast at 38°N, 80°W, an area of potent high frequency variability. The rms error of forecasting always climatology (C) is 170.4 gpm. The number of cases (N) in each bin is indicated below the small squares.

The first obvious conclusion derived from Fig. 5a is that in order to beat persistence (on average at the target point) we need to find an analogue for which the initial error $B - A$ is about 40 meters or less (quality about 50 gpm or less). This condition is met 85% of the time. The best analogue nearly always has an initial error less than 40 gpm. (We can now see that our randomly chosen example in Table 1, features a rather large initial error.) So there is potential for forecasts of some use at the target point nearly all the time, a gratifying result, that differs strongly from the literature on AF.

The second conclusion to be drawn from Fig. 5a is that the smaller the initial error the better the forecast. Although this was to be expected, it is nice to see this very basic assumption about forecasting empirically verified for reasonably small and medium sized initial errors.

Putting the observational/analysis error in B and A at 20 meters, we found 83 pairs of A and B that resem-

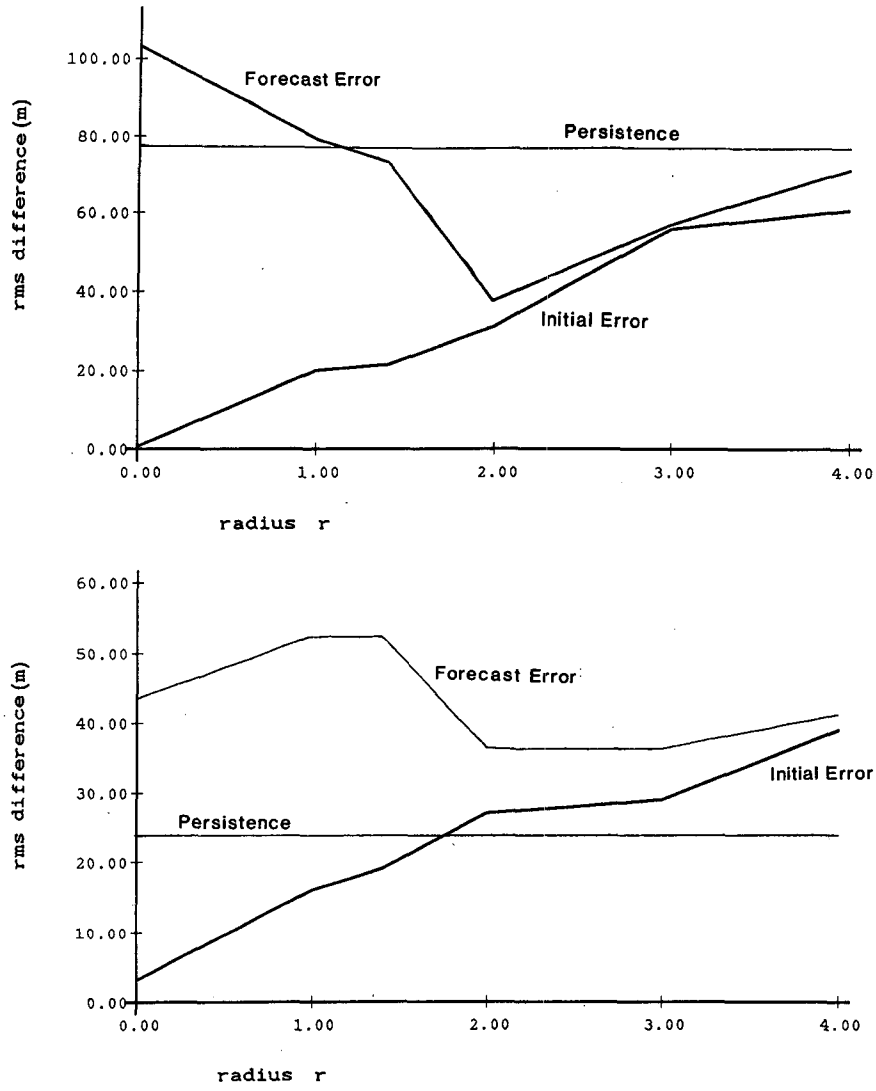

FIG. 4. As in Fig. 3, but now the base dates at 1200 UTC are 1 February 1974 (top); and 31 January 1964 (bottom).

ble each other to within about the observational error, a result that is in stark contrast to Lorenz (1969). For these 83 cases the forecasts are very good indeed.

One obtains a complementary impression by calculating the rms error of the 12-h persistence forecast for each bin separately:

$$\text{persistence rms error} = \left\{\frac{1}{N}\sum_{i=1}^{N}(B - V_i)^2\right\}^{1/2}.$$

Figure 5b is identical to Fig. 5a but now the persistence rms error per bin has been added to the figure. We now see that AF gains over persistence by a somewhat constant amount irrespective of the magnitude of the initial error. (Discard the two points on the right which are based on a small number of cases). Apparently there is a relationship between persistence and analogue quality, such that it is easier (harder) to find good analogues for flows that prove to be persistent (transient) over the next 12 hours. In NWP a similar dilemma occurs in that predictability seems highest when the flow persists. It is not obvious why quality and persistence are related. We also note (not shown) that small (large) initial errors ($B - A$) seem to correlate with medium (large) anomalies ($B - C$). Some of these properties are related to the use of the rms measure for matching analogues and verifying forecasts.

The data in Fig. 5 can also be used to study error growth, the principal goal in Lorenz (1969). We will come back to this point later in section 4a of this paper.

The results in Fig. 5 are negatively influenced by a small percentage of "busts," i.e., analogues that lead

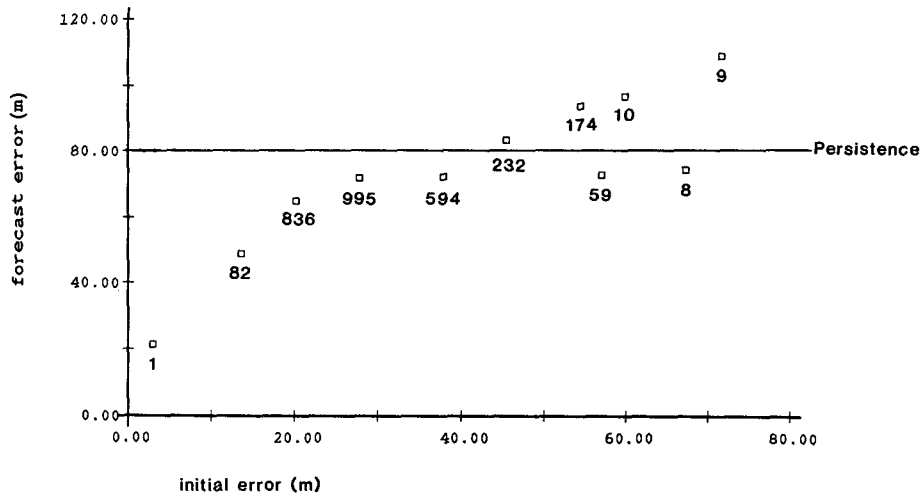

FIG. 5a. The 12-h 500 mb point height forecast error in geopotential meters (gpm) as a function of the initial error (also in gpm) based on 3000 winter forecasts at 38°N, 80°W. Each case consists of ten forecasts based on the ten best analogues. The results are summarized for each bin along the horizontal axis. The numbers just below the small squares indicate how many initial errors fell into that bin.

to very bad forecasts. Under some conditions (e.g., fast moving flow and strong baroclinic development), a circle with $r = 2$ and analogy at 500 mb height alone would not suffice for making a good 12-hour forecast. Since these conditions are known beforehand we probably could do a better job with a flow-dependent r and requiring other levels to be analogous too. However, these aspects have not yet been investigated.

One way to improve the forecast would be to combine analogues. After all, we are essentially making one control forecast (the one based on the best analogue) + 9 Monte Carlo cases around it. Analogues 2 to 10 can be interpreted as identical twin experiments, i.e., everything is the same as the control except for naturally perturbed initial conditions. We constructed a weighted mean over the 5 best analogues by

$$F^* = \frac{1}{W} \sum_{i=1}^{5} F_i/Q_i$$

where Q_i is quality, F_i is the height forecast according to the ith analogue and W is the sum of the $1/Q_i$. In Table 2 we provide extensive summary statistics for all 3000 point forecasts verified individually (column labeled "all"), the best analogues only (300, column "anal") and the weighted forecast (300, column

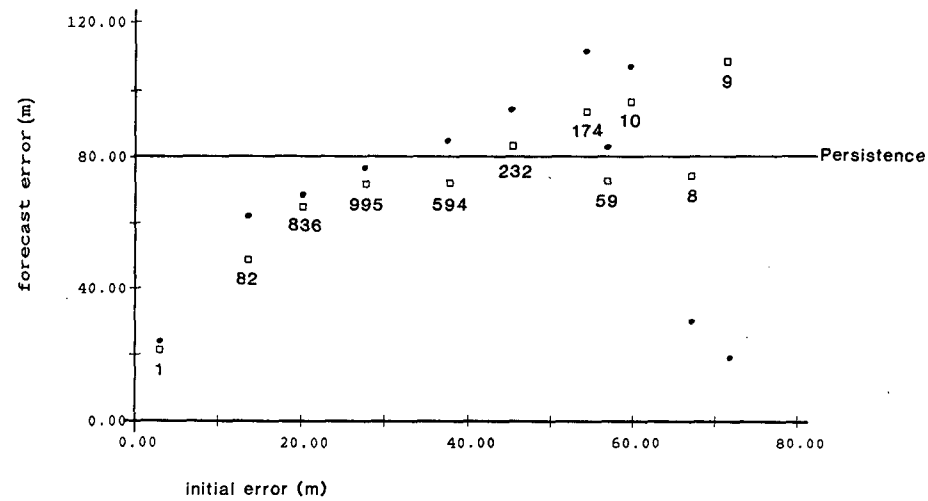

FIG. 5b. Identical to Fig. 5a, but now the persistence errors, evaluated for each bin, have been added.

"ana5"). In Table 2 several additional verification statistics are presented. We define a local anomaly correlation as

$$\mathrm{AC} = \frac{\sum_{i=1}^{N} (F_i - C)(V_i - C)}{\{\sum_{i=1}^{N} (F_i - C)^2 \sum_{i=1}^{N} (V_i - C)^2\}^{1/2}} \quad (2)$$

where N is the number of cases (3000 or 300 in Table 2). The tendency correlation (TC) is calculated similar to (2), the arguments being $(V_i - B_i)$ and $(F_i - B_i)$. It is clear from Table 2 that combining five analogues gives a dramatic improvement in skill (as measured by rms error, AC and tendency correlation) over the single best analogue. Importantly enough, this is achieved without damping the forecast towards climatology (see $(F - C)$ statistics) as we would expect to happen if we combined five forecasts that had little in common. Averaged over all 300 cases the prominent (ana5) scores are a 51.6 gpm rms error, a 0.95 AC and a 0.77 TC. (The tendency correlation being larger than 0.5 is consistent with an rms error smaller than persistence.) In Appendix B we will compare these statistics to verification scores of NWP. In that context it should be pointed out that the initial error $(B - A)$ averaged over all 3000 cases is 33 gpm, which is certainly more than the analysis/observational error in B (B − truth), which would be relevant to the error growth in a subsequent run with a NWP model. In Appendix B we will therefore consider AF verification statistics based on two subsets which satisfy the condition that the Q for the best analogue is smaller than 20 or 15 gpm.

TABLE 2. Verification statistics (RMSE, AC and TC) of the 300(0) 12-hour 500 mb point height forecasts, verifying at 38°N, 80°W. "All" refers to the case where we verify the ten best analogues individually, ana1 when we consider the best analogue only and ana5 refers to a weighted mean over the first five analogues. The symbols B, A, F and V are explained in Fig. 1; C is climatology. The units for the rms scores are gpm.

Number of cases	all 3000	ana1 300	ana5 300
Initial rms error $(B - A)$	33.0	24.9	19.1
12-h forecast rms error $(F - V)$	72.6	68.8	51.6
12-h persistence rms error $(B - V)$	80.3	80.3	80.3
12-h climatology rms error $(V - C)$	170.4	170.4	170.4
12-h rms forecast magnitude $(F - C)$	165.1	169.1	161.0
Initial AC $(A - C, B - C)$	0.98	0.99	0.99
12-h forecast AC $(F - C, V - C)$	0.91	0.92	0.95
12-h persistence AC $(B - C, V - C)$	0.89	0.89	0.89
Tendency correlation $(F - B, V - B)$	0.58	0.63	0.77

d. An example forecast map valid at 1200 UTC 30 January 1966

By moving the target point around (and with it the circle over which analogues are sought) we can make forecasts at any gridpoint and make a map of the resultant field—our forecast map. In Fig. 2 we showed a 9 × 9 grid. Taking 0000 UTC 30 January 1966 as the Base date again, we made 12-hour forecasts for each of the target points on the interior 5 × 5 grid. Obviously we need data on the entire 9 × 9 grid to do so. At each of the 25 gridpoints we took $r = 2$ and a window from 5 January through 25 February. To reduce the noise in the forecast somewhat, we decided a priori to combine five analogues (gridpoint by gridpoint) into one forecast in the manner described in section 3c. No explicit spatial smoothing is applied.

Figure 6 has six maps: (a) Base map (valid at 0000 UTC 30 January 1966); (b) Verification map (valid 12 hours later at 1200 UTC 30 January); (c) Forecast map (valid at 1200 UTC 30 January 1966); (d) Observed tendency map $(V - B)$; (e) Forecast tendency map $(F - B)$; and (f) Forecast error map $(F - V)$.

To judge the case further, verification statistics compiled over the 25 gridpoints are presented in Table 3. Figures 6a and 6b feature a severe cold outbreak over the southeastern part of the United States. A 500 mb low moved eastward and developed an intense jet to the south of the low. The observed tendency, Fig. 6d, showed a drop in heights of 24 decameters in only 12 hours off the coast of the Carolinas. The forecast tendency, Fig. 6e, has the right pattern (Figs. 6d and 6e have a pattern correlation of 0.77, see Table 3), but misses the intensity of the development, particularly the strong jet south of the lowest heights. Both in rms error (less than persistence) and tendency correlation (higher than 0.50) some real skill in AF is indicated. Also the anomaly correlation is respectable: 0.97, which is 0.07 higher than persistence.

Given the maps in Fig. 6, and the verification statistics in Table 3, we call this forecast moderately successful. The map as a whole (Fig. 6c) also looks meteorological, so the tendencies produced pointwise by AF are reasonably coherent in space. This point will be further supported by the favorable outcome of the verification of gradients in section 3e. What seems missing (in this case) is the baroclinic development, which is understandable because we required analogy in 500 mb heights only. So our result here and for the 3000 point forecasts would probably have been better over midoceans and western parts of the continents (where the flow is more barotropic) provided that good quality analogues can be found there too. Requiring analogy for temperature, as well, is obviously desirable, but the subsequent increase in the degrees of freedom

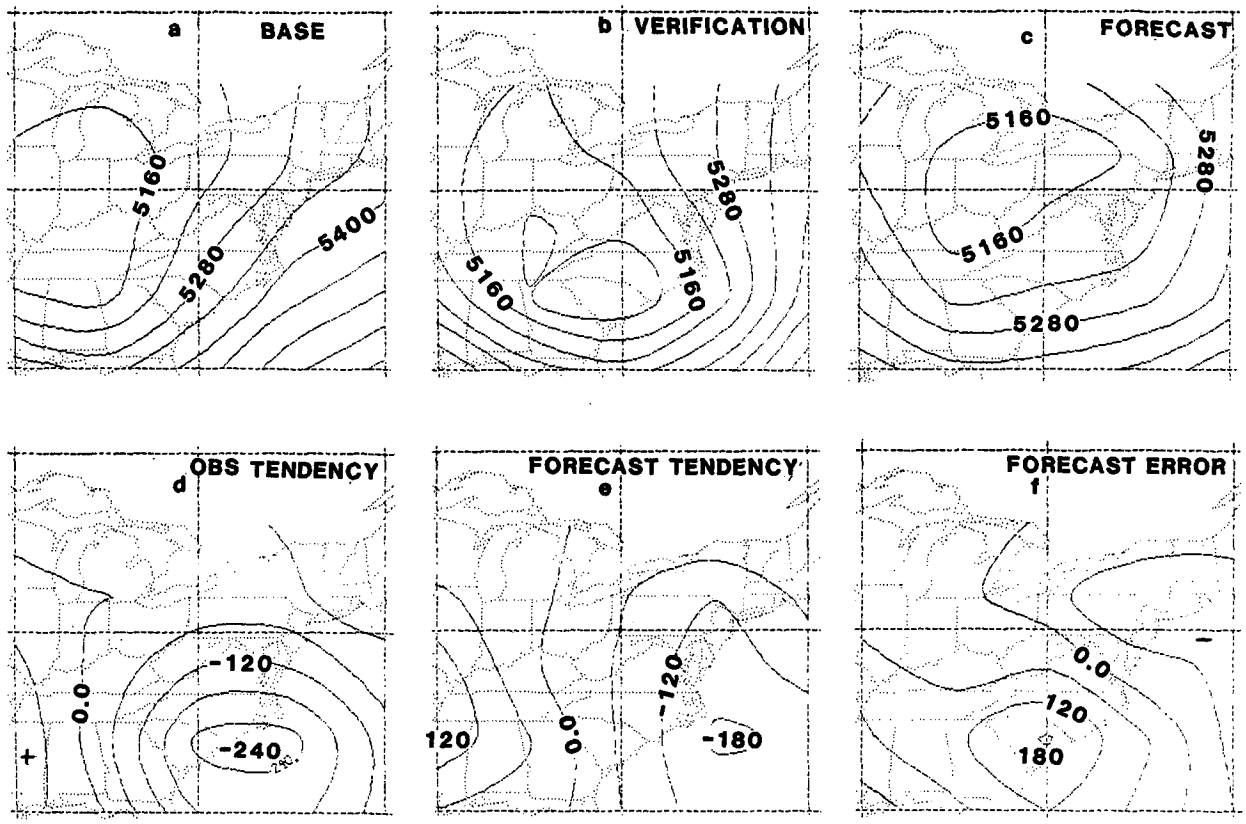

FIG. 6. Six maps pertaining to the 12-hour 500 mb height forecast verifying at 1200 UTC 30 January 1966. (a) the Base (B); (b) the Verification (V) 12 hours later; (c) the Analogue Forecast (F); (d) the Observed tendency ($V - B$); (e) the Forecast tendency ($F - B$); (f) the Forecast error ($F - V$). The units are in geopotential meters and contours are at every 60 gpm in all the maps.

would make it harder to find better-than-mediocre analogues.

Although the methodology allows different historical analogues to be used at different gridpoints, we do occasionally see the same analogues appear at adjacent gridpoints (i.e., they are high on the list like Table 1). Out of a maximum of 25 (i.e., the number of target gridpoints), only 11 distinct dates are seen in the best analogues used to construct Fig. 6c. The large scale nature of the data is therefore seen to reduce the spatial independence of the forecast tendencies. Nevertheless, we use a lot of analogues, many more than a single one. So how much did we gain (in Table 3) over the traditional AF situation where the same analogue (and only a single analogue) would have been used at all gridpoints? The latter case can easily be investigated by choosing $r = \sqrt{8}$ (a 5 × 5 grid centered around 38°N, 80°W), and using the best analogue to make a forecast at all 25 gridpoints. The result of this experiment is given in Table 4 and as can be seen this is not

TABLE 3. Verification statistics (RMSE, AC and TC) of the 500 mb height forecast map, verifying at January 30, 1966 at 1200 UTC, calculated over the 25 gridpoints of the interior 5 × 5 grid (see Fig. 2). The units for the rms scores are gpm.

Initial rms error ($B - A$)	28.5
12-h forecast rms error ($F - V$)	71.1
12-h persistence rms error ($F - V$)	107.5
12-h climatology rms error ($V - C$)	238.6
Initial AC ($A - C$, $B - C$)	0.99
12-h forecast AC ($F - C$, $V - C$)	0.97
12-h persistence AC ($B - C$, $V - C$)	0.90
Tendency correlation ($F - B$, $V - B$)	0.77

TABLE 4. As Table 3, but only the initial RMSE and 12-h forecast RMSE (in gpm) and TC, for experiments using the best analogue (anal), weighting over the first 5 analogues (ana5) and different sized areas for analogue searching. In all 6 cases the verification is over the same interior 5 × 5 grid. The results in the sixth column (ana5, moving $r = 2$) are identical to those reported in Table 3. The $r = \sqrt{8}$ (anal) experiment is as equal as possible to previous literature on AF.

	$r = \sqrt{8}$		$r = \sqrt{32}$		moving $r = 2$	
	anal	ana5	anal	ana5	anal	ana5
Initial RMSE	61.2	53.4	94.7	89.7	24.7	28.5
12-h RMSE	132.2	118.2	116.8	107.9	72.4	71.1
TC	0.28	0.37	0.57	0.54	0.80	0.77

a very good forecast. The $r = \sqrt{8}$ (anal) case is, in experimental design, as similar as possible to the experiments discussed by Gutzler and Shukla (1984) [except for the size of the area (bigger in their case) and the forecast lead time (1 day in their case)] and, to no surprise, our results for $r = \sqrt{8}$ are in agreement with the literature, i.e., you cannot beat persistence. Therefore, it appears that changing analogues from point to point is crucial for a successful forecast.

There is, however, another reason for poor skill in our $r = \sqrt{8}$ case (as well as in Gutzler and Shukla's experiments on "small" areas). By making the verification and analogue search area the same, one loses protection against contamination from the boundaries and beyond. That is why we also did a $r = \sqrt{32}$ experiment, covering the 9×9 grid in our analogue search (with 38°N, 80°W as the central point), but verifying the 12 hr forecast based on the best analogue over the interior 5×5 grid only. The results of this experiment (see Table 4 under $r = \sqrt{32}$ heading) indicate that although the initial error is larger for larger r, protection against boundary contamination does indeed reduce the error growth spectacularly. However, the $r = 2$ "moving" LAM AF model is far better. Using only the best analogue or a weighted combination of the first five appears to be a minor issue in this example.

e. *Extensive verification of forecast maps*

We went on to make 15 forecast maps, verifying at 1200 UTC 3 February in the years 1963–77, using the same (ana5) procedure as in section 3d. By summing over time (15 years) and space (25 gridpoints) overall verification statistics were obtained in a straightforward manner. As far as rms error and Anomaly and Tendency Correlation are concerned, the results are very much like those for the 300 point forecasts in Table 2 (ana5 column). That is, AF gives a forecast of the correct amplitude with clear gains in accuracy over climatology and persistence, particularly so after weighting over the best five analogues. However, since we have produced forecast *maps* here composed of point forecasts based on different analogues, the more challenging question here pertains to the accuracy of gradients in the forecast height field. On the 5×5 interior grid we calculated all permissible centered differences (18 in all) and verified them against the observed gradients.

The verification of height and height gradients is given in Table 5. Only rms errors are shown as AC and TC lead to the same conclusions. These scores provide us with objective evidence that the gradients of the 12-h 500 mb height forecasts have skill over persistence and climatology. This finding backs up our subjective opinion that the forecast map for 1200 UTC 30 January 1966 (Fig. 6c) looks meteorological. Along the same lines we conclude that the forecast height errors have rather high spatial correlation. If the tendencies and 12-h forecast errors were spatially uncorrelated the errors in the gradients would be $55.2\ \sqrt{2} = 78.1$ gpm, and we find them to be only 59.1. (Note that the *initial* errors seem to be spatially uncorrelated.) One problem, however, is some loss of amplitude in the forecast gradients, i.e., 89 versus 112. Apparently weighting five analogues could be excessive. For the sake of comparison to NWP we also assessed the accuracy of the gradients by calculating the S1 score (see Appendix B).

f. *24-hour point forecasts*

Given a 12-h 500 mb height forecast map on the 5×5 interior grid one can search for historical cases in the 1963–77 dataset that match the 12-h forecast over the $r = 2$ circular area, placing, as before, 38°N, 80°W in the center as the target point for a 24-h 500 mb height forecast. This means that not only do we change analogues from gridpoint to gridpoint to arrive at a 12-h forecast map, we also change analogues from t to $t + \Delta t$, where Δt is 12 hours for practical reasons. We have applied this repeated procedure to the fifteen 12-h forecast maps, discussed in section 3e, so as to obtain a (weighted over five analogues) 24-h 500 mb height point forecast for 0000 UTC 4 February 1963–1977. Since the 24-hour forecasts are for the target point only, we will, for comparison, also verify the initial state and the 12-hour forecast at the same point (38°N, 80°W). The results are given in Fig. 7. The full line is the error growth of AF, while the thin line represents persistence. While at $t = 0$ persistence is error free, and AF has a 26.2 gpm error, AF outperforms persistence at $t = 12$ h, and, by an increasing margin, at $t = 24$ h. [Contrast this to Ruosteenoja (1988) who found the AF initial error to be as large as the persistence error after 2.5 days (his Fig. 5).] The dashed line represents the rms error of persistence of the 12-hour forecast out to 24 hours. Clearly there is information in the time deriv-

TABLE 5. Verification of fifteen 12-h AF maps in terms of rms height and height difference errors. The maps verify on 1200 UTC 3 Feb 963–77. The verification area is the 5×5 interior grid. Units are gpm. In the last column the forecast magnitude is given.

	RMSE				
	Initial $(B - A)$	12-h $(F - V)$	Persist $(B - V)$	Climate $(V - C)$	Forecast magnitude $(F - C)$
Height	18.8	55.2	75.4	120.6	112.7
Height difference	28.5	59.1	80.9	115.0	89.3

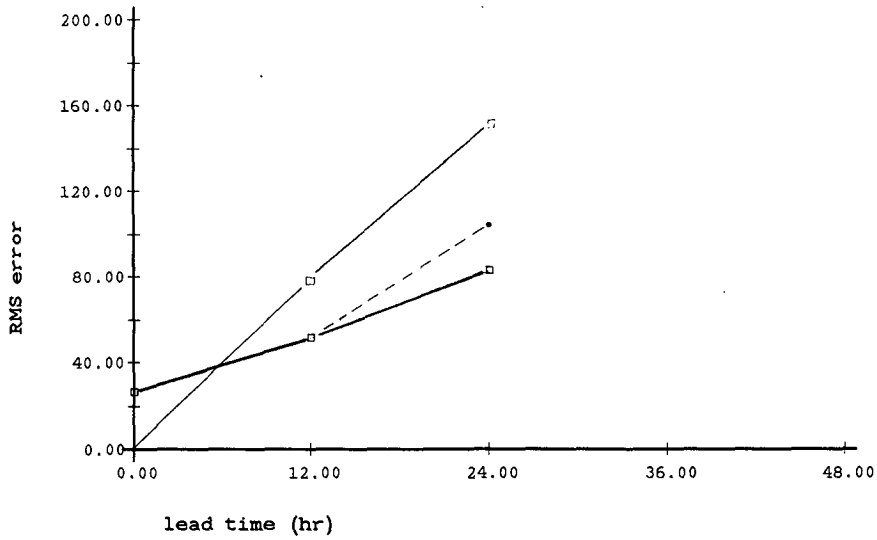

FIG. 7. The rms error of AF at 38°W, 80°N, as a function of lead time (full line). The thin line is for persistence of the initial state, while the dashed line refers to persisting the 12 h forecast. Statistics based on 15 forecasts originating from 0000 UTC 3 February. Forecasts are a weighted mean based on the five best analogues.

ative produced by AF in the second time step. So the procedure of the second Δt timestep seems to work fine on this, admittedly, small sample. Implicitly these results speak well also for the 12-h forecast *maps*. If these maps were unmeteorological, it would have been impossible to find matching historical analogues of acceptable quality, and the second time step would have been a disaster automatically. Note also that the forecast at $t = 24$ h is (as is the 12-h forecast) a linear combination of five previously observed atmospheric states and therefore there is no compounding loss of physical realism of the forecast flow patterns as the AF procedure is extended to 24 hours (or beyond).

4. Applications

We will discuss two applications of the use of limited-area analogues. The first is concerned with predictability and the growth of initially small differences. The second and potentially practical application is an attempt to forecast forecast skill using a collection of analogue forecasts to measure the local stability of the flow. These two applications are not unrelated. In this section we consider individual point forecasts only (i.e., no linear combination such as ana5).

a. Predictability

Lorenz (1969) tried to find naturally occurring analogues so as to empirically study the rate of divergence of two atmospheric states that are initially close. The strength of this approach is that no simplifying assumptions about the physics of the atmosphere are needed. But his attempt was considerably frustrated by the fact that, over the Northern Hemisphere, no two states can be considered good analogues. As argued extensively in the above, very good analogue pairs can be found over a smaller area if one matches 500 mb heights only. Under those conditions, we do find a few percent of the analogue pairs (B and A) to be separated by a distance comparable to the observational error. Hence we can study the doubling time of rather small initial differences. This is done here by reprocessing the 3000 12-h point height forecasts discussed in section 3c. In fact, we use exactly the same data points as used in Fig. 5a, but now we plot the ratio of forecast to initial error as a function of the initial error. The result is given in Fig. 8. On the right-hand-side the vertical axis has been rescaled to indicate the doubling time (in hours) of an initial discrepancy (initial "error" in the NWP context) as a function of the magnitude of the initial error. A ratio of 2 (4) implies a doubling of the rms error in 12 (6) hours. In a flow that "forgets" its initial state after some time, it is to be expected that the error growth increases with decreasing initial error. Indeed, that is what we find in Fig. 8. This is rather different from Lorenz (1982) who postulated an exponential increase of small errors with time (i.e., a constant doubling time). Note that in contrast to Lorenz (1969) the error growth of small errors was obtained without any extrapolation. Note also that our results pertain to one location only. For not-so-small errors of 60 gpm and larger, the error growth in Fig. 8 is visibly smaller, although saturation should occur only beyond 200 gpm. In principle, by relabeling the initial error axis into a forecast lead time axis, we could find the time T at which the error has reached 95% of its

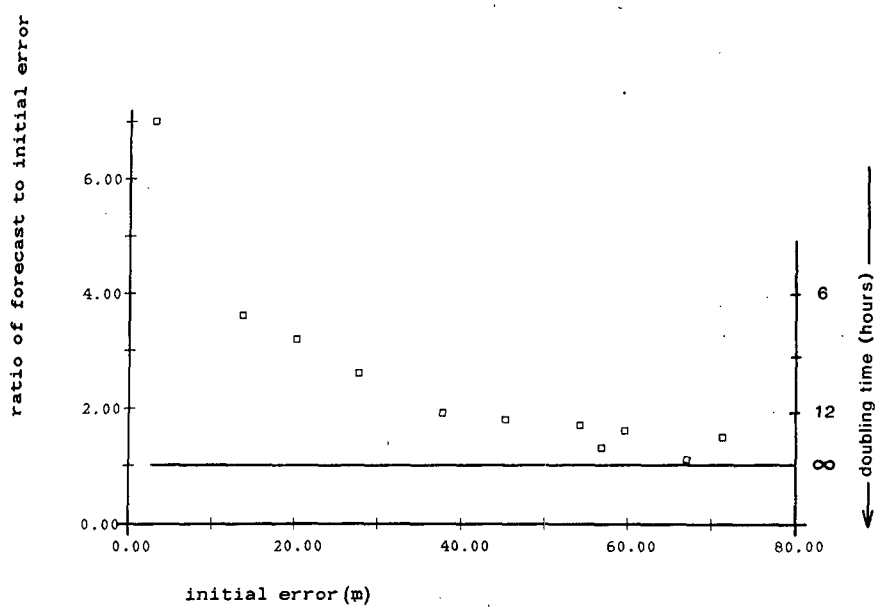

FIG. 8. Error growth as a function of the magnitude of the initial error. This figure is a rearrangement of the data displayed in Fig. 5. A ratio of 2 (4) can be interpreted as a doubling time of the rms error of 12 (6) hours.

final saturation value (239 gpm). For a curve suggested by the small squares in Fig. 8, T is finite, no matter how small the initial error is. We mention three problems and challenges in interpretation.

One has to remember that we have sought analogues for 500 mb height only. The initial error in the thermal field is unspecified and perhaps large. Therefore, the overall level of error growth in Fig. 8 could be an overestimate. Indeed we do not see anything as slow as a doubling times of 2 days (for small errors on a hemispheric domain!) as reported by NWP-identical twin experiments (Dalcher and Kalnay 1987). Averaged over all 3000 cases (Table 2) the initial error (33 gpm) grows in 12 hours to the 72.6 gpm forecast error. That points at doubling in less than 12 hours.

There is a problem associated with the uncertainty of the initial discrepancy $B - A$. For very good analogue pairs the uncertainty (i.e., the observational/analysis error) in B and A becomes an issue. Since A is selected to be nearest to B we may, to some degree, have fitted noise and hence the true value of $B - A$ could be larger than what we have calculated. If that is so the error growth, displayed in Fig. 8 is overestimated, and increasingly so for smaller values of $B - A$. This effect could partly explain the shape of the curve in Fig. 8.

There is yet another challenge to the interpretation of Fig. 8. The small (large) initial errors seem to be associated with persistent (transient) flow. Therefore it may be worthwhile to consider local predictability estimates separately for persistent and transient flows. When combining data for more than one location into a Figure like Fig. 8, this interpretation challenge ultimately disappears because on the globe, as a whole, there are no clear-cut persistent or transient flow patterns.

b. *Forecasting forecast skill*

Since the skill of medium range NWP forecasts varies greatly from day to day, the forecast of skill has attracted considerable attention lately (Kalnay and Dalcher 1987). The underlying question is whether we can assess the stability of the flow at the initial time, the assumption being that stable flow patterns are more predictable. Several strategies have been designed to measure the stability of the flow, for example, by considering an ensemble of Monte Carlo forecasts or lagged operational forecasts (verifying at the same time). Within the context of analogues, the spread of ten independent analogue forecasts (the F_i in Fig. 1) can conveniently be measured by:

$$S = \left\{ \frac{1}{N} \sum_{\substack{i=1,10 \\ j=1,i-1}} (F_i - F_j)^2 \right\}^{1/2}$$

where $N = 45$. In the spirit of Kalnay and Dalcher (1987) we have plotted the 12-hour forecast error as a function of S, see Fig. 9. In this figure we have used the same 3000 point forecasts discussed in section 3c. As before, we have binned the forecast error for bins of 10-gpm width, i.e., $0 < S < 10$, $10 < S < 20$, etc. For a given spread, we have verified ten individual forecasts, i.e., $F_1, \ldots, F_{10}$, so the number of cases in each bin (given below the small squares in Fig. 9) is a multiple of 10.

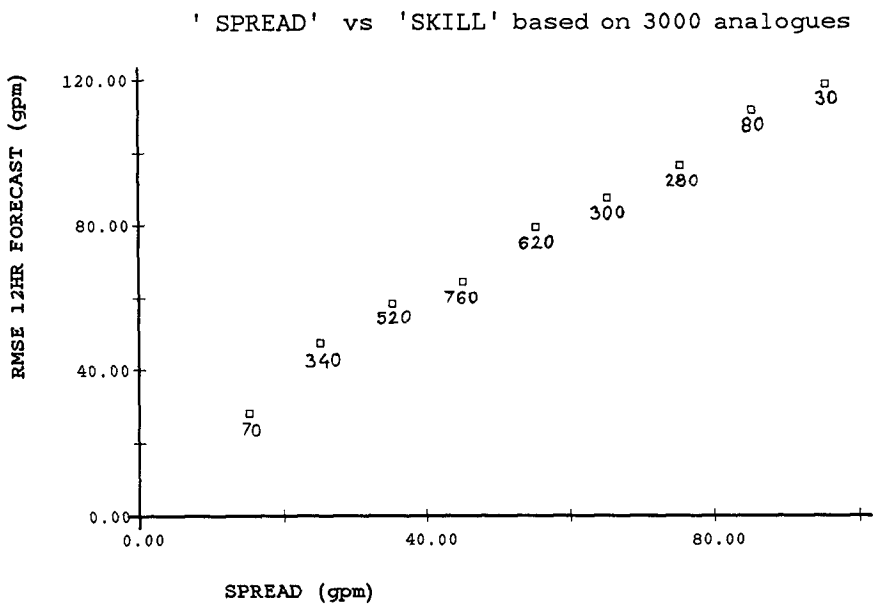

FIG. 9. The 12-h forecast error (gpm) as a function of the spread (gpm) of ten analogue forecasts. Result based on 3000 point forecasts shown in Fig. 5a.

Figure 9 presents rather convincing empirical evidence that spread and forecast skill are related. The relationship is so strong that one wonders whether analogues can be used to forecast forecast skill of an NWP model.

The way we measure spread is identical in methodology to what is used in Monte Carlo forecasting except that, in the latter, the manner in which one perturbs the initial conditions is somewhat arbitrary and debatable. The rate of divergence (i.e., the spread) when using analogues is probably a lot higher (and realistically so) than when using ten Monte Carlo runs of an NWP model. A problem of many NWP models is that the various Monte Carlo solutions often stay close together, closer than any of them is relative to the verification. The analogues have no such problem, since we use the atmosphere as its own model. While the skill in 12-hour AF is largely due to vorticity advection, i.e., the barotropic part, the spread of ten AFs reflects to a large degree baroclinic processes.

5. Conclusions and discussion

The purpose of this paper is to show that even with very limited datasets, there is some potential in analogue forecasting (AF), much more so than suggested in the literature (Lorenz 1969; Gutzler and Shukla 1984; Shabbar and Knox 1986; Ruosteenoja 1988; etc.). The solution lies in limited-area analogue forecasting. In order to make a 500 mb height forecast at a target point, we only need analogy in a circle (around the target point) with a radius of less than 1000 km. Even with a data library of only a few years, many moderate to good analogues can be found over a small area. The circle has to be small enough to find good analogues and yet large enough to prevent boundary and remote effects to travel to the center point in 12 hours. By checking for analogy only in 500 mb height we have made an AF model akin to a barotropic model (see Appendix A). Hence our expectations regarding forecast skill have to be modest.

Using a 15-year data library (1963–77), we verified a multitude of 12-hour 500 mb height *point* analogue forecasts in winter valid at 38°N, 80°W. Conclusions are:

1) For $r > 4$ (circle with 1800 km radius), the analogues found are generally too poor a match of the base to be of much use in practical forecasting (Figs. 3 and 4);

2) For $r < 1$ (radius 0 to 400 km) near-perfect analogues can be found, but the forecasts are disastrous because of rapid advection of information from the unmatched exterior to the target point (Figs. 3 and 4);

3) For $\sqrt{2} < r < 3$, moderate-to-good analogues can be found which provide 12 hour forecasts that are generally able to beat the persistence forecast (Figs. 3 and 4). In other words, we do not have to wait an eternity before we can do something useful with AF.

4) More specifically, in order to beat, on average, "persistence" in a region of high variability, such as that represented by 38°N, 80°W (in January/February), we need to find analogues that are about 50 geopotential meters (or less) separated (over the circle, in rms sense) from the base case. As shown in Fig. 5 this can usually be achieved.

5) The better the analogue, the better is the 12-hour forecast. This obvious assumption has been confirmed here empirically for small initial errors. Figure 5 features an excellent case which started with an initial discrepancy of less than 10 gpm and led to a 12-hour forecast with an error of only 22 gpm.

6) Creating a weighted average over the five best analogues yields a point forecast of realistic amplitude, which on average over all 300 cases, has a 51.6 gpm rms error, a 0.95 AC and a 0.77 TC (Table 2, ana5 column).

Forecasts at 25 adjacent target points were put together to create 12-hour 500 mb height forecast *maps*. At each gridpoint, one final forecast was produced by weighting the first five analogues by the inverse of their quality (determined by the initial discrepancy over the circle). We conclude that:

7) The resulting forecast maps look meteorological even though the forecasts at adjacent points are based on developments in different historical analogues (Figure 6). A verification of the height gradients in 15 AF forecast maps (Table 5) supports this conclusion.

8) The limited area AF approach appears to be holding up in the 2nd time step leading to a 24-hour forecast (Fig. 7).

We can think of at least four contributing factors as to why we succeeded in making analogue forecasts with some degree of success while previous workers (Gutzler and Shukla 1984; Shabbar and Knox 1986; Ruosteenoja 1988) did not. These factors are 1) the use of limited areas or, more generally, the lowering of the degrees of freedom in finding matching states, 2) protection against contamination by the boundary conditions (and beyond) by verifying at the target point only, 3) the use of data every 12 hours so that the analogue search area is permitted to be reasonably small and 4) taking a weighted average over the first few analogues.

Analogue Forecasting as described here has a lot more skill than envisioned before. In Ruosteenoja (1988) the mean initial error ($B - A$) was as large as the persistence error after 2.5 days, but in our limited-area version AF beats persistence at 12 hours, and with an increasing margin at 24-hours. This is a great improvement in empirical forecasting and there is some reason to believe that we could do better still. Nevertheless, a direct application to short-range 500 mb height forecasting seems unlikely, given that NWP as of today is far better. In Appendix B we show that AF (as described here) performs roughly the same, in terms of 12-hour rms error and S1 score, as NWP in the 1950s and early 1960s. This is to be expected since we match only 500 mb height and therefore AF should be skillful in the barotropic component of the tendency of 500 mb height only. (If we accept a bigger initial error we could match more than one variable and build a baroclinic AF. For longer lead times this baroclinic AF may yield better forecasts.) Even though the verification scores of AF and a barotropic model are somewhat comparable, we have not definitively shown that AF's skill is due to correct vorticity advection.

Many extensions and follow-up efforts suggest themselves, some of which have been outlined below.

1) Experiments with the matching system. This is an almost obvious extension. Rather than rms difference [used in (1)], gradients could be used to find the best analogues. Instead of geopotential height, geostrophic streamfunction could be used. We could try to match vorticity advection. Instead of the $r = 2$ circle we could use different flow dependent shaped areas for analogue search.

2) There is obviously a need to consider other geographical locations, other seasons, and more recent datasets.

3) Given that there are very often several useful analogues, a natural way of Monte Carlo forecasting seems within reach. In this paper we have, in Fig. 6c, combined the first five analogues with weights indicative of their quality. Better weighting schemes should be investigated.

4) Having a concurrent dataset of synoptic weather observations, such as rainfall, surface air temperature, etc., probabilistic forecasts of elements such as rainfall ought to be investigated. Currently such forecasts are based on only realizations of future flow produced by NWP.

5) On the theoretical side we can study error growth (Lorenz 1969), a central concept in questions pertaining to the ultimate predictability of instantaneous weather patterns. One example was worked out in section 4a.

6) As shown in section 4b there seems to be good potential to forecast the skill of AF using the spread of the AFs as a predictor. It would be worthwhile to investigate whether skill of NWP can be forecast by the spread of AF.

7) It is straightforward to check analogy on the thickness field as well, so as to mimic a perfect baroclinic model. However, a larger area or more variables naturally reduce our chances of finding good analogues. We will have to weigh the benefits of reduced error growth against the disadvantage of a larger initial error. Maybe extension to other variables is feasible only in cases of weak flow when $r < 2$ is allowed, or with datasets available every 6 hours.

8) We have investigated the use of negative analogues (antilogues, Van den Dool 1987). Antilogues are as much as possible the opposite of the base, opposite relative to the climatology. We have found that there are many good antilogues in addition to the analogues. Forecasts based on antilogues have 12-hour verification scores comparable to those of analogues. This issue will be pursued in a later article.

Finally, we need to mention a problem in interpretation not encountered in the literature so far. In the

context of AF we would like to interpret the difference $B - A$ to be the initial error. Apart from an ambiguity in measuring $B - A$ (over the circle or at the target point only?) the value of $B - A$ suffers from observational/analysis error (ϵ) in B and A. Because B and A were so far apart in Ruosteenoja (1988) ϵ is a trivial matter, but in our case $B - A$ is numerically sometimes as small as ϵ. An initial error ($B - A$) of 20 gpm may therefore be not much better than one of 30 gpm. So "very good" and "near-perfect" analogues cannot be reliably distinguished. The east coast of the United States is a well observed area with high day-to-day variablility. That probably helped greatly the results presented in this paper, in particular Fig. 5. Over the oceans and some parts of the continents the observations are less accurate and often missing, and it may be that 50 gpm is the lowest meaningful difference between B and A. If that happens to be in an area of low day-to-day variability (i.e., a low persistence error) AF may be a failure in short-range forecasting.

Along the same lines of argument, AF will always be at a disadvantage relative to NWP, even if we have enough data to find analogues such that $B - A$ is on the order of ϵ. In NWP the initial error is $\epsilon_1 = (B - \text{truth})$ while in AF the initial error $\epsilon_2 = (B - \text{truth}) + (A - B)$. Clearly the latter is larger by a factor $\sqrt{2}$ (this is true for very small $B - A$).

Most of the results reported in this paper were, broadly speaking, anticipated before any calculations were done. This is because the AF procedure is very much based on existing knowledge about the dynamics and numerical models. Perhaps empirical NWP would be a better name than AF. There are a few findings that were not at all expected. Prominent among them is that, if the flow is persistent [$(B - V)$ small], it is much easier to find high quality analogues. This is an interesting empirical fact. It seems to indicate that *locally* persistent flows come in only a few varieties while transient flows, although at least as common, come in all shapes and sizes. So a 15-year library is long enough for some flows, and much too short for others.

Acknowledgments. The idea of trying limited-area analogues was born after reading Ruosteenoja (1988). The assistance of Mr. Chuan-Qi Miao with some of the computer work is gratefully acknowledged. All computations were carried out on the IBM 4381 at the University of Maryland. The patient and critical ears of Drs. David Rodenhuis, Owen Thompson, Suranjana Saha, Zoltan Toth, Ake Johansson and Tony Barnston proved invaluable while carrying out this research. Mrs. R. Hirano and C. Vlcak of NMC are acknowledged for helping to find the verification scores of operational forecasts in the 1950s and 1960s. Comments by Drs Eugenia Kalnay, John Lanzante, Phil Arkin and Chris Folland were much appreciated. Funding for this research was provided by a grant from NOAA (NA84-AA-H-00026).

APPENDIX A

The Governing Equations

It is customary, when presenting a model, to write down the governing equations. In the case of AF, all we know is that the equations are perfect, albeit unknown. (In fact, if anthropogenic changes are important, a historical analogue no longer provides a perfect model.) In view of the success of NWP (implying that we do know the governing equations of the atmosphere) we may however speculate what the "equations" of AF are. The AF equations are primarily a function of how we select analogues.

To a high degree of approximation the vorticity equation (in p coordinates) can be written for midlatitudes as:

$$\frac{\partial \zeta_g}{\partial t} = \underbrace{-\mathbf{V}_g \cdot \nabla(\zeta_g + f)}_{\text{(I)}} + \underbrace{f_0 D}_{\text{(II)}} \tag{A1}$$

where vorticity (ζ_g) and horizontal wind ($\mathbf{V}_g$) are geostrophic, D is the divergence and f the Coriolis parameter. Because of the equivalent barotropic nature of the atmosphere, D tends to be minimal at some level near 500 mb (although not everywhere all the time). Therefore, if the atmospheric states have very similar 500 mb height (over a small area) term A has to be very similar for those two states, and hence the tendencies in ζ_g (or height) have to be similar too.

We cannot think of any other equation in which analogy in just one variable at one level would guarantee similarity in tendency. Technically we could repeat all of the above experiments with, say, temperature at some pressure level. But from the thermodynamic equation (geostrophically approximated):

$$\frac{\partial T}{\partial t} = -\mathbf{V}_g \cdot \nabla T + \sigma\omega \tag{A2}$$

where T is temperature, σ is static stability and ω is vertical motion, it follows that similarity in T does not imply that the advection term as a whole is similar, and therefore, even if there is a level where ω is small, temperature analogues need not have the same tendencies.

As far as height fields are concerned our one variable/one level AF works best when D is negligible, which is probably somewhere near 500 mb. Therefore the governing equations of our AF method (sofar) is akin to a perfect version of an equivalent barotropic model with some noise added through the term II in (A1). By combining five analogues this noise will be reduced and F will seem to have evolved from A according to barotropic vorticity advection.

We can only guess about the 'numerics' of the AF model. Since term I is nonconstant in time we cannot assume that:

$$\Delta\zeta_g = -\mathbf{V}_g \cdot \nabla(\zeta_g + f) \times \Delta t \qquad \text{(A3)}$$

where Δt = 12 h and term I is held at its $t = 0$ value. Equation (A3) would be a better approximation for small Δt. So imagine we had a dataset every 30 minutes. In that case we could select analogues on the basis of similarity in vorticity advection and jump Δt ahead at each gridpoint, much the same as in a NWP numerical scheme.

Extending AF to a perfect two level quasi-geostrophic model suggests itself. Demanding similarity in height fields over colocated circles at say 300 and 700 mb implies analogy in both vorticity advection at these two levels and thermal advection at an intermediate level.

APPENDIX B

Comparison NWP and AF

A comparison of NWP and AF cannot be avoided. Here the comparison will be made only in terms of relative skill. Of course two methods yielding equal skill are not necessarily producing the same forecast. Given that our present AF is akin to a barotropic model, it is fair to first compare the verification statistics (Table 2, the ana5 case) to those of barotropic 500 mb height forecasts. A strictly valid comparison cannot be made because that would require a lengthy experiment on an identical set of initial conditions. But we can give a rough assessment nevertheless.

First we discuss rms errors which were sometimes reported in the early literature. Experiments done with the early barotropic models (Staff members 1952) yielded a 12 h rms height error for an European area in winter of 58 gpm. Given the much higher standard deviation of heights over the Eastern United States in the high frequencies (Blackmon 1976) the 12 h AF error of 51.6 gpm (Table 2, ana5) compares very much in favor of AF. Thompson and Gates (1956) report a 64 gpm 24 h 500 mb height error over the United States (as a whole) for barotropic forecasts run from 60 initial conditions in January 1953. Linear interpolation between 64 gpm at 24 h and an initial error of 30 gpm places the 12 h error at 47 gpm, not very different from AF's 52 gpm. It is a known problem that the rms score favors forecasts that lose amplitude over time. We can not verify whether the barotropic forecasts lost amplitude with forecast lead time but they probably did. If so the comparison is a bit more in favor of AF which loses little or no amplitude. NWP has improved dramatically over the years. Several model generations later Dey and Morone (1985) report a 25 gpm rms error (averaged over 102 radiosondes in the Northern Hemisphere) in winter for the 6 h 500 mb height forecasts used in the Global Data Assimilation. Currently that figure is well below 20 gpm. So, in terms of rms error, AF can only compare to the very early models. Even the barotropic forecasts have improved over the years (Bengtsson 1985), primarily because the initial states (produced in conjunction with state-of-the-art baroclinic models) are much better now. It is possible therefore that the AF results (based on 1963–77 data) would likewise improve if we use data over the 15 most recent years.

The AF results in Table 2 indicate that averaged over 3000 cases the initial error ($B - A$) is 33 gpm. That is quite a bit higher than NWP as of the 1980's. Therefore, to place AF and NWP on equal footing as far as initial error is concerned we created a subset of analogue forecasts for verification which satisfied $Q_2 < 20$ and $Q_1 < 15$ gpm respectively. The results are given in Table 6. It is clear that with smaller initial error the 12 h forecast error goes down considerably, not unlike the decrease of forecast error in the barotropic model from the 1950s to the 1980s (Bengtsson 1985). However Table 6 could be a bit too flattering for AF because, as we have seen before, smaller initial errors are associated with flows of higher persistence.

Many of the older studies report forecast accuracy in terms of S1 score which measures the relative error in the gradients. Long records of S1 scores for 500 mb height and sea level pressure are available (Shuman 1972). This forced us to calculate the S1 score of the 15 AF maps discussed in section 3e. We find S1 to be 0.26 for AF, which is very close to the 0.24 value reported by Carstensen (1957) for 12 h 500 mb height forecasts over the United States in the winter of 1955/56. According to Shuman (1972) a value of S1 equal to 0.20 (0.60) indicates near-perfect (worthless) 500 mb forecasts. Assuming an initial S1 = 0.15 in the 1950s and 1960s a linear interpolation using Shuman's (1972) record of 36 h S1 value seems to indicate that AF (that is to say our version of it, using data over 1963–77) would have been competitive with 12 h NWP forecasts in terms of S1 until 1965 or so. Currently S1 is 0.10 in winter.

In the older literature the anomaly correlation was practically never reported. A rare exception is Namias (1958) who mentions a 0.90 AC at 12 h for operational

TABLE 6. As Table 2 but only 4 selected statistics. Column 1 is identical to the ana5 column in Table 2. Columns 2 and 3 are for a subset of the 300 forecasts, satisfying $Q_1 < 20$ and <15 gpm respectively. Units for the rms scores are gpm.

Threshold for Q_1 (number of cases)	ana5 none 300	ana5 20 59	ana5 15 13
Initial rms error ($B - A$)	19.1	10.3	9.2
12-h forecast rms error ($F - V$)	51.6	43.8	33.8
12-h persistence rms error ($B - V$)	80.3	65.6	44.6
12-h forecast AC ($F - C$, $V - C$)	0.95	0.95	0.97

numerical sea level pressure forecasts over North America. (At that time the numerical sea level pressure forecast was considered inferior to the subjective man made forecasts.) Currently the AC at 12 h for 500 mb heights is 0.98 or so, which is better than AF's 0.95 (Table 2, ana5).

A strictly valid comparison between AF and barotropic forecasts would be worthwhile. Not just to compare a variety of skill measures (as we did in the above), but also to investigate what kind of forecast is made by AF. Doing this may give us some clues on model improvement.

REFERENCES

Bengtsson, L., 1985: Medium-range forecasting—The experience of ECMWF. *Bull. Amer. Meteor. Soc.*, **66,** 1133–1146.

Berggren, R., 1958: A comparative study of 700, 500 and 300 mb barotropic forecasts. *Geophysics,* **6,** 147–168.

Blackmon, M. L., 1976: A climatological spectral study of 500 mb geopotential height of the Northern Hemisphere. *J. Atmos. Sci.,* **33,** 1607–1623.

Carstensen, L. P., 1957: Verification data for the 3-level numerical forecasts of 1955–56. Office Note No. 12, p. 5. [National Meteorological Center, Washington, D.C.]

Dalcher, A., and E. Kalnay, 1987: Error growth and predictability in operational ECMWF forecasts. *Tellus,* **39a,** 474–491.

Dey, C. H., and L. L. Morone, 1985: Evolution of the National Meteorological Center global data assimilation system: January 1982–December 1983. *Mon. Wea. Rev.,* **113,** 304–318.

Dunn, G. E., 1951: Short-range weather forecasting. *Compendium of Meteorology.* T. F. Malone, Ed., Amer. Meteor. Soc., 802–813.

Gutzler, D. S., and J. Shukla, 1984: Analogs in the wintertime 500 mb height field. *J. Atmos. Sci.,* **41,** 177–189.

Kalnay, E., and A. Dalcher, 1987: Forecasting forecast skill. *Mon. Wea. Rev.,* **115,** 349–356.

Livezey, R. E., and A. G. Barnston, 1988: An operational multifield analog/anti-analog prediction system for United States seasonal temperatures Part I: System design and winter experiments. *J. Geophys. Res.,* **93,** 10 953–10 974.

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

——, 1982: Atmospheric predictability experiments with a large model. *Tellus,* **34,** 505–513.

Namias, J., 1951: General aspects of extended range forecasting. *Compendium of Meteorology.* T. F. Malone, Ed., Amer. Meteor. Soc., 802–813.

——, and Collaborators, 1958: Application of numerical methods to extended forecasting practices in the U.S. Weather Bureau. *Mon. Wea. Rev.,* **96,** 467–476.

Ruosteenoja, K., 1988: Factors affecting the occurrence and lifetime of 500 mb height analogues: A study based on a large amount of data. *Mon. Wea. Rev.,* **116,** 368–376.

Shabbar, A., and J. L. Knox, 1986: Monthly prediction by the analogue method. *Proc. of the First WMO Workshop on the Diagnosis and Prediction of Monthly and Seasonal Atmospheric Variations over the Globe.* Long-Range Forecasting Res. Rep. Ser. 6, Vol. II, Tech. doc. WMO/TD 87, 672–681. [World Meteorological Organization, Geneva, Switzerland.]

Shuman, F., 1972: The research and development program at the National Meteorological Center. Office Note 72. [National Meteorological Center, Washington D.C.]

Staff Members of the Institute of Meteorology, 1952: Preliminary report on the prognostic value of barotropic models in the forecasting of 500 mb height changes. *Tellus,* **4,** 21–30.

Thompson, P. D., and W. L. Gates, 1956: A test of numerical prediction methods based on the barotropic and two-parameter baroclinic models. *J. Meteor.,* **13,** 127–141.

Van den Dool, H. M., 1987: A bias in skill in forecasts based on analogues and antilogues. *J. Climate App. Meteor.,* **26,** 1278–1281.