Intercomparison of Circulation Similarity Measures

ZOLTAN TOTH*

Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado, Boulder, Colorado

(Manuscript received 25 January 1990, in final form 25 July 1990)

ABSTRACT

Most circulation studies use the root-mean-square difference (RMSD) or correlation (COR) (or both) as a tool for comparing different (observed or forecast) circulation patterns. However, there are some other measures introduced into the literature (e.g., mean absolute error, rms vector error, S1 score) and one might easily construct other measures of circulation similarity. The question of the appropriate choice among possible similarity measures rarely arises. In this paper an objective intercomparison of nine different similarity measures (also called distance functions) is presented. The similarity measures were evaluated through the 700-hPa hemispheric analog forecasts obtained by them. In the indirect evaluation, the analogs to the base cases found by each individual distance function were checked whether they were identical with the best analogs (selected by all nine functions) to the circulation pattern that actually followed. The number of coincidences is an indication of the quality of the similarity measures and is found, both for daily and pentad data, to be largest for a dynamically oriented distance function that measure the difference between the gradient of height of two maps. For daily data, RMSD also appears to be significantly better than COR. However, in a direct assessment, where analog forecasts by each distance function were compared to the analysis fields using one of the distance functions to measure the difference, practically no performance differences were found among the functions that performed differently in the indirect evaluation.

It should be noted that the results of both intercomparison methods are, in a strict sense, valid only for forecast situations. For other purposes, other distance functions might be more appropriate. However, there are some indications that the similarity measure that performed best in the forecast experiments (difference in the gradient of height) remains superior in other applications, too.

1. Introduction

The continually expanding literature on analogs can be divided into two groups. In the real forecasting application, similar atmospheric cases are followed into the future and are used to make predictions some time steps ahead. Studies in this group are concentrating on extending the limit of practical predictability beyond that attainable presently with numerical weather prediction (NWP) models. An example of these long-range forecasting papers is that of Livezey and Barnston (1988), which provides a useful section of recent references. The other group includes papers like that of Kruizinga and Murphy (1983). Here a forecast is derived from an independent source (usually from NWP models) and the analogs are used only to establish simultaneous relationships between different successfully predicted (e.g., geopotential height) and not directly well predicted (e.g., surface temperature at different locations) variables.

An essential part of the studies in either group is the way they measure the similarity between different atmospheric states. But it is not only in the area of analogs that a definition of similarity is needed. In verifying any forecast (most exclusively NWP) map the difference between the predicted and observed situation should also be measured. As Murphy and Winkler (1987) pointed out, forecast verification should be based on the joint distribution of forecasts and observations. Following their ideas, any single score to measure the skill is necessarily incomplete. However, there are a lot of situations (especially with many variables as is the case in map verification) where practical reasons necessitate the choice and use of one measure of forecast performance among possible candidates. It is known from the theory of statistics that in an idealized case, when there is an infinite number of potential analogs to choose from and when the number of independent analog predictions also goes to infinity, it is irrelevant which distance function is used in selecting analogs. But neither of these assumptions are satisfied in a real situation. The number of past realizations is indeed sharply limited, representing the biggest drawback to analog prediction (see e.g., Ruosteenoja 1988);

* Formerly at Cooperative Institute for Climate Studies, Department of Meteorology, University of Maryland, College Park, Maryland.

Corresponding author address: Dr. Zoltan Toth, NOAA/National Meteorological Center, W/NMC, WWB, Room 204, Washington, DC 20233.

© 1991 American Meteorological Society

Unauthenticated | Downloaded 05/10/21 11:01 PM UTC

also one is making or evaluating a finite number of forecasts. Therefore it might be important to make a good choice among possible similarity measures.

There are a few papers in which the choice of an appropriate similarity measure (or, in other words, distance function) is considered; this review does not include studies that use more elaborate pattern recognition techniques for forecast verification, as e.g., Somerville (1977). In their study of 500-mb wintertime analogs, Gutzler and Shukla (1984) examined three different distance functions, defined through the root-mean-square difference (RMSD) and the correlation (COR) and covariance functions. Although they stated that the three distance functions all emphasize different aspects of the similarity between two circulation patterns, they found that the different methods led to many common analogs as best matches to their base cases. No conclusion was reached regarding a preferable matching system, except that other distance functions, e.g., quantifying the strength of the jet stream, might produce more useful analogs. In another article, Shabbar and Knox (1986) compared the analogs obtained by the RMSD and COR functions. Their subjective conclusion was that the latter better reflects the differences found by visual synoptic inspection. Conversely, in his pioneering work on limited area analogs, van den Dool (1989) used the RMSD function, while mentioning that other functions should also be considered. In a pilot verification study by Arpe (1978) the rms phase error, COR, and RMSD were compared subjectively. The rms phase error (where "for all wavenumbers the forecast data are changed so that they keep the phase of the original forecast data and take the amplitude of the analysis fields") and COR were found to be superior to the simple RMSD function.

In an interesting new development, Murphy (1988) and Murphy and Epstein (1989) provide the analytical relationship between COR and RMSD. They show that COR, compared to RMSD, is an incomplete measure of similarity of forecast and verification maps. In the usual way COR is used, different biases are ignored, and, as supported also by some verification data, its square (and not the coefficient itself) measures rather the potential skill of a forecast and not the skill itself.

It is clear from this short review of the literature that researchers differed on the most appropriate way to compare different atmospheric states. No objective intercomparison of various distance functions has been performed, and there are potential similarity measures that have not been tested at all. Also, in an analog forecast situation the theoretically derived superiority of RMSD to COR in forecast verification has not been confirmed by empirical data.

In this study, an objective intercomparison of different similarity measures is presented. Northern Hemispheric (NH) circulation analogs were selected by nine different distance functions. The analogs then were used to make circulation forecasts. The quality of the different similarity measures was measured in two ways. First, a check was made on how many of the analogs obtained by the different similarity measures still remained the best analogs in the forecast period (i.e., the best analogs to the target circulation data to be forecast). (In the verification mode, the best analogs were defined by using all nine similarity measures without any a priori assumption.) Another way to compare the quality of different distance functions used in this study was to employ one of them (e.g., RMSD) to verify the circulation forecasts obtained through analogs selected by every individual distance function. It should be emphasized that both intercomparison methods are based on the quality of analog forecasts. Consequently the results, in a strict sense, will be valid only for forecast situations. There is a possibility that for other purposes other distance functions might be more appropriate.

The database and the methods used are presented in section 2. Section 3 contains the results of the intercomparison study while section 4 gives some conclusions.

2. Data and methods

a. *Database*

This study uses the twice daily Northern Hemisphere 700-hPa height series of the National Meteorological Center (NMC). Instead of the original 541-point grid, the 358-point nearly-equal-area grid (designed and created by Barnston and Livezey 1987) was used. This grid covers the Northern Hemisphere from 20°N, excluding some areas over Africa and Eurasia. (When computing gradients the data over the North Pole were also used.) Because changes in the mean height over the whole hemisphere are supposed to be related mainly to interactions with the remainder of the global atmosphere (and hence not forecast) the data were converted into map anomalies. Each map's mean was removed and only the anomalies from the map mean were kept. In the first part of the intercomparison study, daily averaged data were used from 1–25 January of each year between 1950 and 1986. In the other part of the intercomparison, 5-day averaged (pentad) data were used from the same years, from 1–5 January through 27–31 March (26–30 March in leap years; 18 pentads for each year). If z_{ij} stands for the height at gridpoint i ($i = 1, \ldots, 358$) and at time point j ($j = 1, \ldots, 925$ for daily and $j = 1, \ldots, 481$ for pentad data) then the map anomaly is

$$a_{i,j} = z_{i,j} - \sum_{i=1}^{358} z_{i,j}/358.$$

Unauthenticated | Downloaded 05/10/21 11:01 PM UTC

b. Methods

1) CIRCULATION SIMILARITY MEASURES

Beyond the traditional and generally used distance functions; such as the absolute difference in height, the RMSD, and the COR, the quality of six other functions was investigated. These, following the suggestions of Gutzler and Shukla (1984), contain some information about the difference in the strength of the flows to be compared. To quantify the strength of a flow, the zonal ($\Delta_x a_{ij}$) and meridional ($\Delta_y a_{ij}$) gradients for each grid point were estimated. Whenever it was feasible the difference in height between the two adjacent grid points on the same (or very close) latitude or longitude circle, respectively, was used as an estimate of the gradient. Close to the pole the grid system loses some of its symmetry features in order to maintain a quasi-equal area representation (see Fig. 1 of Barnston and Livezey 1987). Hence, at some locations it was necessary to use three or four neighboring points to compute the gradients. Note that the meridional gradients are computed for the same finite base distance at each grid point. However, due to the geometry of the earth the latitudinal distance for which $\Delta_x a_{ij}$ was computed is not the same at each grid point. As the gradients of different maps are compared grid point by grid point, gradients from *different* locations are not compared to each other directly; no correction was introduced to make the latitudinal base distance uniform over the whole hemisphere. Instead, it was assumed that its changes have no influence on the relative performance of different similarity measures.

The different similarity measures tested in this study are the following:

mean absolute difference in height,

$$D_1 = \left(\sum_{i=1}^{358} |a_{i,j} - a_{i,l}|\right)/358, \tag{1}$$

where, as in the following equations, $j \neq l$. This, probably the simplest way of comparing different circulation patterns, can be altered to put the emphasis on areas where both circulation patterns have large gradients. The mean absolute difference weighted by the gradients is:

$$D_2 = \frac{\sum_{i=1}^{358} |a_{i,j} - a_{i,l}|[(\Delta_x a_{i,j})^2 + (\Delta_y a_{i,j})^2]^{1/2} \times [(\Delta_x a_{i,l})^2 + (\Delta_y a_{i,l})^2]^{1/2}}{358}. \tag{2}$$

The effect of the gradients is even more pronounced when the squares of the gradients are used. This measure is the mean absolute difference weighted by the square of the gradients:

$$D_3 = \frac{\sum_{i=1}^{358} |a_{i,j} - a_{i,l}|[(\Delta_x a_{i,j})^2 + (\Delta_y a_{i,j})^2] \times [(\Delta_x a_{i,l})^2 + (\Delta_y a_{i,l})^2]}{358}. \tag{3}$$

Instead of the mean absolute difference in height most studies prefer to use RMSD:

$$D_4 = \left[\frac{\sum_{i=1}^{358} (a_{i,j} - a_{i,l})^2}{358}\right]^{1/2}. \tag{4}$$

The above measures are a function of the difference in height between two maps. Replacing the height by the height gradient in (1) one obtains the expression of the difference in gradient of height:

$$D_5 = \frac{\sum_{i=1}^{358} [(\Delta_x a_{i,j} - \Delta_x a_{i,l})^2 + (\Delta_y a_{i,j} - \Delta_y a_{i,l})^2]^{1/2}}{358}. \tag{5}$$

This formula is a modification of the rms vector error used for verification purposes at NMC. (They use the square root of the summed squares of differences between the forecast and observed winds, which is an analog to the rms height difference in a sense that the same function is applied here to the wind field as to the height field in the case of RMSD.) Without taking the square root of the quantity in the numerator of (5) we get:

$$D_6 = \frac{\sum_{i=1}^{358} [(\Delta_x a_{i,j} - \Delta_x a_{i,l})^2 + (\Delta_y a_{i,j} - \Delta_y a_{i,l})^2]}{358}. \tag{6}$$

This expression, at least in our experiment, is equivalent to the rms vector error, since taking the root of quantities in a series does not change the increasing (or decreasing) order of the elements in the series. Hence, for the sake of brevity we call the distance function (D_6) rms vector error.

The correlation between two maps was computed in the standard way:

$$D_7 = \left(\sum_{i=1}^{358} a_{i,j}a_{i,l}\right)\left[\sum_{i=1}^{358} (a_{i,j})^2 \sum_{i=1}^{358} (a_{i,l})^2\right]^{-1/2}. \tag{7}$$

To compute the difference between two vorticity fields, the vorticity at each grid point ($V_{i,j}$) was estimated in a similar way as the gradients. Here, instead of the height values, the finite differences in the u and

Unauthenticated | Downloaded 05/10/21 11:01 PM UTC

v component of the geostrophic wind were determined along a longitudinal and latitudinal circle, respectively, over a finite base distance. Then the change in $u\,(\Delta_y a_{i,j})$ along a longitudinal circle was subtracted from the change in $v\,(\Delta_x a_{i,j})$ along a latitudinal circle,

$$V_{i,j} = \Delta_x \Delta_y a_{i,j} - \Delta_y \Delta_x a_{i,j},$$

where Δ_x and Δ_y stand for the appropriate finite differences. The difference in vorticity between two maps is:

$$D_8 = \Big(\sum_{i=1}^{354} |V_{i,j} - V_{i,l}|\Big)/354. \tag{8}$$

No vorticity was determined for the four grid points around the pole as no gradient values at the pole were present. Note that the same comments about the different base distances in latitudinal direction apply here as for the computation of gradient values above.

Finally, a modified version of the S1 score was tested:

$$D_9 = \left\{ \left[\frac{\sum_{i=1}^{358} |\Delta_x a_{i,j} - \Delta_x a_{i,l}|}{\sum_{i=1}^{358} \max(\Delta_x a_{i,j}, \Delta_x a_{i,l})} \right]^2 + \left[\frac{\sum_{i=1}^{358} |\Delta_y a_{i,j} - \Delta_y a_{i,l}|}{\sum_{i=1}^{358} \max(\Delta_y a_{i,j}, \Delta_y a_{i,l})} \right]^2 \right\}^{1/2}. \tag{9}$$

The original S1 score was found to perform better in a regional domain on station observation data in a subjective intercomparison study by Teweles and Wobus (1954).

It can be said in advance that there were two or three systems that usually performed relatively poorly. Nevertheless, they are included here because in the intercomparison the performance of each system was tested against the performance of all others and hence partly against these less reliable similarity measures.

2) Intercomparison

To compare the usefulness of different distance functions, analog forecasts through each of the nine functions defined above were produced. The database was divided into two parts: a 33-yr "training" period (1950–1982) and a 4-yr independent testing period (1983–1986). Analogs were selected only from the training period in a research mode, i.e., the base cases (for which an analog was sought) could either precede or follow the analogs in time. For base cases in the training period, candidate analogs only from the other 32 years of the training period were considered. In the case of the daily data, 20 base cases in every year beginning each 1 January were considered; in the pentad data base, 12 base cases per year beginning 1–5 January were considered. The analogs were always selected from the same 20-day or 12-pentad window of the year as the base cases. Therefore, for a daily base case in the training period 32 years × 20 days = 640 possible analogs were to be considered.

For every base case and for all the nine similarity measures the 10 most favorable analogs (with the smallest distance functions or the largest COR values) were recorded, and forecasts five time steps (days or pentads) ahead were made.

The success of the different similarity measures was evaluated by means of verifying these forecasts. This was done in two ways. As the question here is the proper way of comparing two circulation patterns, it is not adequate to evaluate the analog height forecasts obtained by the different distance functions in terms of the height difference between the forecast map(s) and the appropriate analysis map(s) following the base cases, measured by a particular distance function. First, an indirect evaluation was chosen. As the 10 top-ranking analogs for every base case were selected, the same was done for the first, second, . . . , fifth map *following* the base case, i.e., for the verifying analysis of the daily or pentad forecasts. The same analog selection rules applied above that were used for the base cases, when considering analogs to the analysis field for the first, second, . . . , fifth time step ahead, except for appropriately moving the analog selection window (20 days or 12 pentads) one, two, . . . , five days or pentads ahead. So when looking for the best analogs to the analysis field following a base case by, for example, one day (two days), potential analogs were considered in the period 2–21 (3–22) January from all the other years. The analogs to the analysis fields obtained this way are apparently the *best analogs* that can be found in our archive as potential analog forecasts. Consequently, the better the distance function (similarity measure), the more analogs to the *base cases* are gained through it, which remain the best analogs to the *analysis fields* (maps subsequent to the base cases) as well. To avoid any a priori preference for one or another matching system, the list of 10 best analogs was combined from all the nine distance functions (see Fig. 1). In each base case and at each forecast lead time (1–5 time steps ahead), 90 best analogs (which are not all necessarily different since with the different matching systems we may get the same analogs) need to be compared with the list of 10 top-ranking analogs to the base cases found by applying one analog type in particular. The measure of analog type performance employed here is just the number of coincidences found in the lists of the 90 best analogs and 10 top-ranking analogs to the base cases, summed over a part or the whole of the training period or the test period. The number of coincidences for different analog types (that

Unauthenticated | Downloaded 05/10/21 11:01 PM UTC

Distance Function	Analogs to the base case				Analogs to the verifying analysis field (following the base case)			
1	${}_0a_1^1$,	${}_0a_2^1$,	$\cdots$,	${}_0a_{10}^1$	${}_1a_1^1$,	${}_1a_2^1$,	$\cdots$,	${}_1a_{10}^1$
2	${}_0a_1^2$,	${}_0a_2^2$,	$\cdots$,	${}_0a_{10}^2$	${}_1a_1^2$,	${}_1a_2^2$,	$\cdots$,	${}_1a_{10}^2$
3	$\vdots$				$\vdots$			
$\vdots$	$\vdots$				$\vdots$			
7	$\vdots$				$\vdots$			
8	${}_0a_1^8$,	${}_0a_2^8$,	$\cdots$,	${}_0a_{10}^8$	${}_1a_1^8$,	${}_1a_2^8$,	$\cdots$,	${}_1a_{10}^8$
9	${}_0a_1^9$,	${}_0a_2^9$,	$\cdots$,	${}_0a_{10}^9$	${}_1a_1^9$,	${}_1a_2^9$,	$\cdots$,	${}_1a_{10}^9$

FIG. 1. Indirect intercomparison of different distance functions ($d = 1, \cdots, 9$). ${}_la_n^d$ stands for the nth ($n \leq 10$) best analog to a particular base case ($l = 0$, left hand part of the figure) or to the first, . . . , fifth analysis field ($l = 1, \cdots, 5$, shown as $l = 1$ in the right hand part) following in time the base case. The 90 best analogs in the left-hand part were compared to the 10 best analogs in each row of the right-hand part. The performance of each distance function was measured by the number of coincidences (same cases) in these two groups of analogs.

includes the double or multiple coincidences, too) were statistically compared to each other by a chi-square test with one degree of freedom.

Another way of comparing the quality of analogs obtained by different distance functions is a direct assessment of the numerical difference between the analysis fields and forecast fields by one or another matching system. Results obtained this way will also be reported in the next section.

3. Results

The results of the indirect intercomparisons of different similarity measures on daily data are presented in Figs. 2 and 3 for the 33-yr training period and the 4-yr testing period, respectively. One-, two-, three-, four-, and five-day forecasts were made with analogs obtained through each similarity measure and verified as described in section 2b.2. The number of coincidences between the analogs to the base cases and the best analogs to the verifying analysis (following the base cases) theoretically can reach the product of all trials and the 9 distance functions (33 years $\times$ 20 days $\times$ 10 cases $\times$ 9 distance functions = 59400 for the training period and 4 years $\times$ 20 days $\times$ 10 cases $\times$ 9 distance functions = 7200 for the test period). Of course, since the 10 best analogs selected by the 9 different distance functions are not necessarily identical, this number is considerably smaller in practice. Consequently, the results for the first day (around 16 000 hits for the most favorable analog types) are quite impressive and show that those analogs that were selected as the best analogs to the base situations are, in many cases, the best analogs on the following day as well. The quality of the analogs, as expected, decreases with increasing lead time; but even for the less successful distance functions the number of coincidences, without any question, is significantly larger than the chance expectation (which is 6600 trials $\times$ 90 target analog cases/32 years/20 days $\approx$ 928 hits for the training period and 800 trials $\times$ 90 target analog cases/33 years/20 days $\approx$ 109 hits for the test period).

As regarding the differences in the performance of individual similarity measures, first note that there are five out of the nine candidate measures that were ranked in one of the first three positions with any lead time for the training or test period. These better distance functions are: the mean absolute difference in height (1), the mean absolute difference weighted by the gradients (2), the RMSD (4), the difference in gradient of height (5) and its modification, the rms vector error (6). An interesting finding is that the absolute error (1) usually has a poorer performance than the RMSD (4). This shows that it is more important to have a relatively evenly distributed spatial error field than a lower average absolute error, but with outstanding discrepancies at some locations. It also supports the preference of most authors for RMSD rather than absolute error.

Another remarkable feature is that the relative performance of the mean absolute difference modified by the gradients (2) improves with increasing lead time while the opposite is true for RMSD (4) (see upper portion of Figs. 2 and 3). Apparently the energy-related

Unauthenticated | Downloaded 05/10/21 11:01 PM UTC

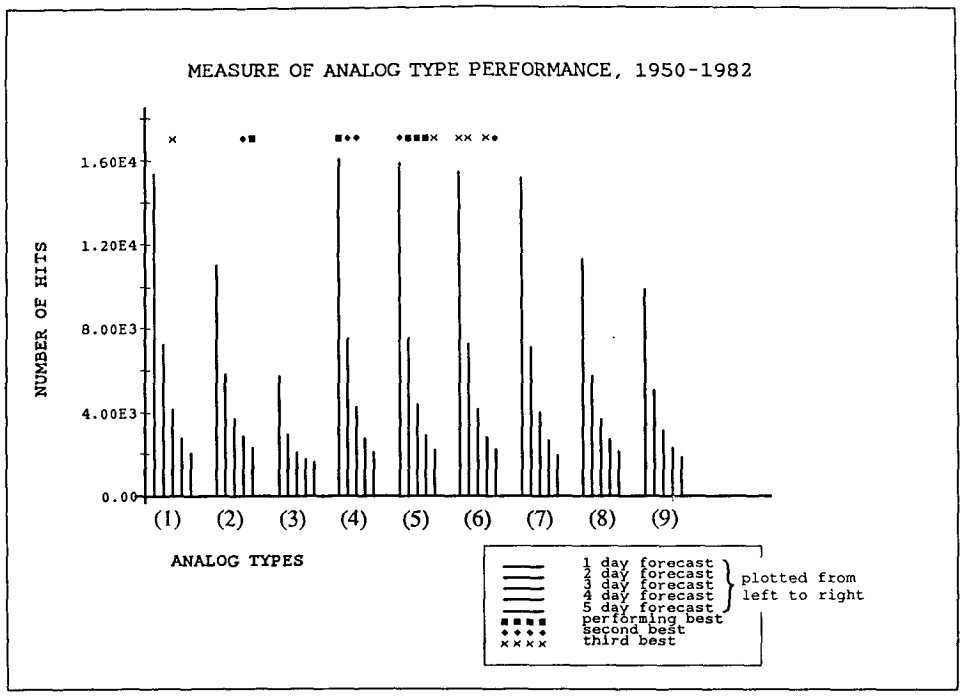

FIG. 2. Evaluation of distance function performance on daily data, 1950–1982. Number of analogs found to the base cases that are the best analogs for the verifying analysis (of a 1-, 2-, 3-, 4- or 5-day forecast) too. (1) Mean absolute difference in height, (2) mean absolute difference modified by the gradients, (3) mean absolute difference modified by the square of the gradients, (4) RMSE, (5) difference in gradient of height, (6) rms vector error, (7) COR, (8) difference in vorticity, and (9) modified version of S1 score. For further details, see text.

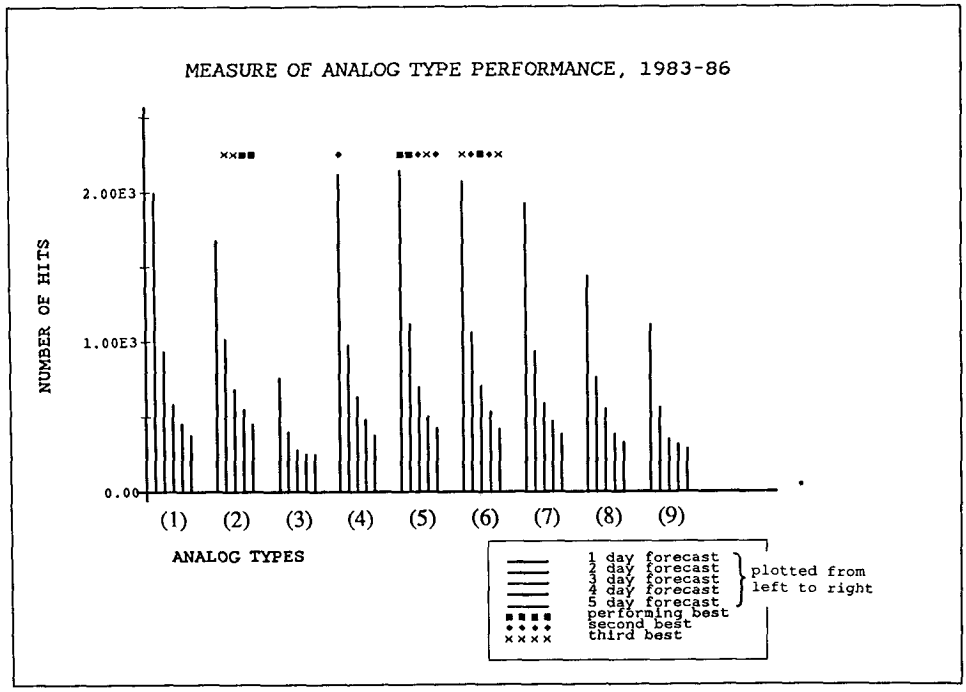

FIG. 3. Same as Fig. 1 except for 1983–1986.

Unauthenticated | Downloaded 05/10/21 11:01 PM UTC

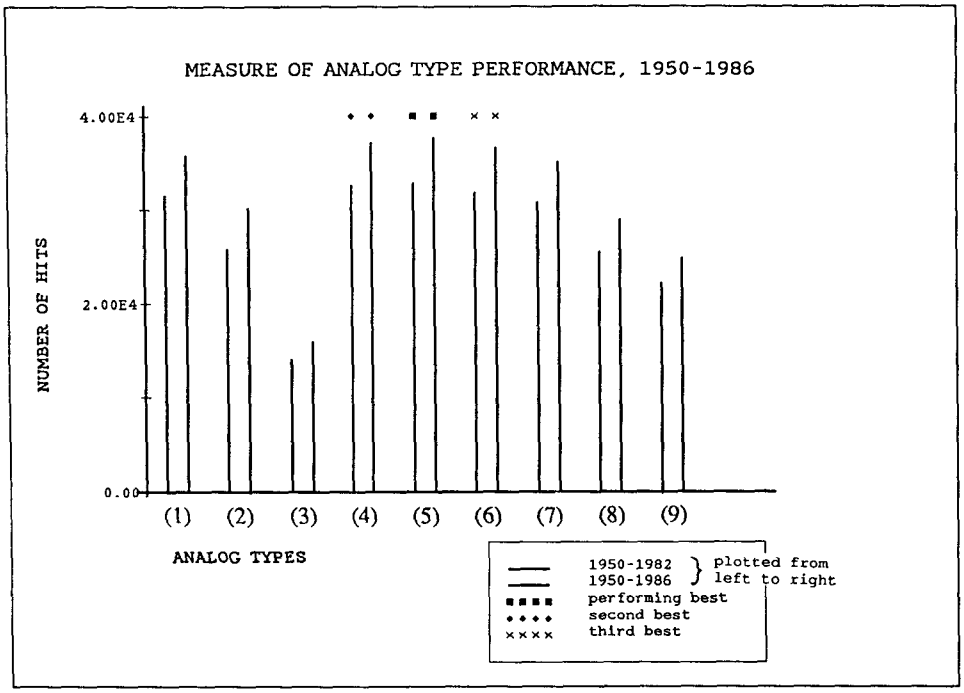

FIG. 4. Same as Figs. 1 and 2 except for accumulating the results of 1-, 2-, 3-, 4- and 5-day forecasts for 1950–1982 and 1950–1986.

information contained in the gradients becomes more and more important with time; consequently, the simple measure of the difference between the two height fields (RMSD) loses its effectiveness. Not so the difference in gradient of height (5) and the rms vector error (6), which are the two best distance functions. Moreover, they have the most stable results with different lead times or verification periods. Note that the difference in gradient of height (5) was in all cases among the best three similarity measures. At the other end of the spectrum is the mean absolute difference modified by the square of the gradient (3), which is burdened by an apparently excessive emphasis on the gradients. Neither the modification of the S1 score (9) nor the difference in vorticity (8) are successful. In the latter case the poor performance may stem from the double derivative that may cause noise to dominate in the measure at least in some cases. The correlation (7), which is widely used both in analog studies and in model verification, is not among the best similarity measures either. That does not come as a surprise in view of the theoretical findings of Murphy and Epstein (1989), discussed in the Introduction.

To summarize the results for daily data, Fig. 4 presents the total number of hits accumulated over all lead times for the training period and for the whole data series including the test period as well. Here again, for both periods considered, the best analog type is the difference in the gradient of height (5). The second best is the RMSD (4), since the large number of hits with short lead times (without any standardization) dominate the results. The third best analog type is the rms vector error (6). Note that a modification of the indirect evaluation method, dropping the less reliable analog type(s) in the reference list of best analogs to the analysis fields, does not change the main features of Fig. 4. To quantify the statistical significance of the differences in the performance of individual analog types, the chi-square values were computed for each possible pair of distance functions.[1] In Table 1 the results for the whole data series are shown. It is clear that the difference between all possible pairs except (4)–(5) and (4)–(6) are significant at the probability level of 0.01 or, in most cases, at a much lower level. The difference between (4) and (5) and between (4) and (6) are significant at 0.05.

It can be concluded that for daily data the best similarity measure is probably the difference in the gradient

[1] The null hypothesis in this test is that the number of coincidences for two different similarity measures comes from the same probability distribution. If we have two samples with $n_1 = v_h + v_n$ and $n_2 = w_h + w_n$ cases, where v_h and w_h are the number of hits and v_n and w_n are the number of nonhits respectively, it can be proved that the variable

$$n_1 n_2 \sum_{i=1}^{2} \frac{1}{v_i + w_i} \left(\frac{v_i}{n_1} - \frac{w_i}{n_2} \right)^2$$

has a chi-square distribution with one degree of freedom, providing $n_1 \to \infty$ and $n_2 \to \infty$.

Unauthenticated | Downloaded 05/10/21 11:01 PM UTC

TABLE 1. Chi-square values for the comparison of results for daily analog forecast performance by the nine different distance functions (see Fig. 4), for each pair and for the period 1950–1986. (1) Mean absolute difference in height, (2) mean absolute difference weighted by the gradients, (3) mean absolute difference weighted by the square of the gradients, (4) RMSD, (5) difference in gradient of height, (6) rms vector error, (7) COR, (8) difference in vorticity, and (9) modified version of S1 score. An asterisk indicates values above the 0.001 probability level (10.83). The values of the chi-square distribution at 0.05 and 0.01 probability levels with one degree of freedom are 3.84 and 6.63, respectively. For further details, see text.

Distance functions	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	0	*	*	*	*	9.34	9.37	*	*
(2)		0	*	*	*	*	*	*	*
(3)			0	*	*	*	*	*	*
(4)				0	4.17	4.52	*	*	*
(5)					0	*	*	*	*
(6)						0	*	*	*
(7)							0	*	*
(8)								0	*
(9)									0

of height. Although RMSD and rms vector error perform almost as well, they are significantly less reliable in our intercomparison than the difference in the gradient of height. All the other similarity measures, including the widely used COR, without any doubt, perform significantly worse than the difference in the gradient of height (or, in most cases, than RMSD or rms vector error as well). However, one should use caution when drawing conclusions from the above, since the other, direct intercomparison method gave no evidence supporting the indirect method's results. In the direct intercomparison, either of the three distance functions that performed best in the indirect intercomparison were used to measure the numerical difference between the analysis maps and forecast maps obtained by the 9 individual distance functions. The difference in the error of the different forecasts obtained through the individual distance functions was usually small but favorable to the best distance functions if present at all during the first forecast days, but this vanished rapidly. With longer lead times practically no difference was evident. However, this negative outcome with the direct intercomparison might simply result from the inherent poorness of analog forecast skill and hence do not necessarily cast doubt on the indirect intercomparison results.

In many cases scientists use and compare time-averaged height fields. Therefore, a check was performed on whether the results obtained through daily data are also valid for five-day (pentad) averages (Fig. 5). The verification data were structured differently from the daily data, allowing the stability of the performance of individual distance functions to be seen over three 11-yr subperiods in the design period. First, as with the daily analogs, the number of coincidences for the first forecast pentad is significantly larger than the chance expectation. However, it decreases rapidly with time and, being close to chance expectation, does not help to choose a favorable similarity measure. Hence the outcome for only the first pentad is shown in Fig. 5. Six similarity measures rank among the three best in one or more periods. However, there is only one, the difference in the gradient of height (5), which is so ranked in all four periods. In Fig. 6, which summarizes the results as does Fig. 4, it is even clearer that the best matching system is once more the difference in the gradient of height. However, the second best in this case is its modification, the rms vector error. When considering the 33-yr design period, the third best system is the RMSD (4), but if the 4-yr independent period is added, COR (7) becomes the third best. In contrast to the results for daily data, the difference in vorticity does not perform poorly, especially for the last 15 years of our data series. The better performance on pentad data is evidently due to the smoothing effect of 5-day averaging. Perhaps its superiority on the recent part of the data series results from generally less noise then. An accurately designed time or spatial smoothing might allow the difference in vorticity to perform better over the whole data series (as suggested by van den Dool 1989).

The chi-square values of the differences in the coincidences for each possible distance function pairs for the whole data series are presented in Table 2. All the differences between the best distance function, the difference in the gradient of height, and the others (except its modification, the rms vector error) are significant at or below the probability level of 0.05. However, as was done with the results based on daily data, caution must be used since the direct intercomparison gives no evidence to support the outcome of the indirect method.

5. Conclusion and discussion

A basic procedure in a wide range of circulation studies is the comparison of different circulation maps (either observed or forecast) to each other. This study presented an objective intercomparison of different similarity measures (distance functions) such as the root-mean-square difference, correlation, and other, less frequently (or previously not) used distance functions. An indirect intercomparison was performed by using hemispheric analog forecasts obtained through nine individual distance functions. The underlying idea

Unauthenticated | Downloaded 05/10/21 11:01 PM UTC

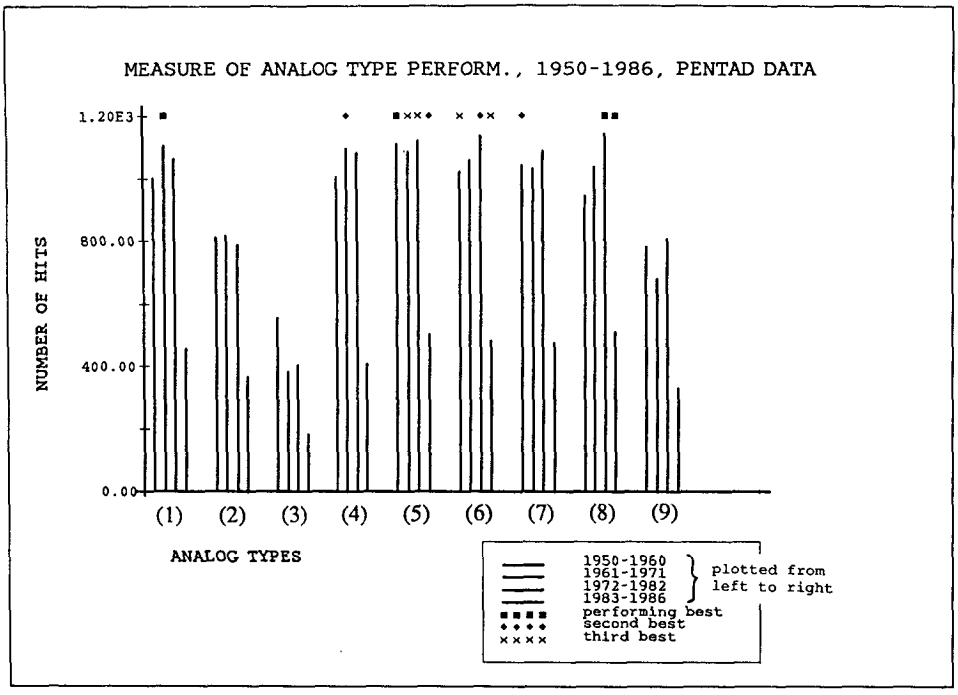

FIG. 5. Evaluation of distance function performance on pentad data, stratified by four subperiods in 1950–1986. For further details, see caption of Fig. 1.

is that a distance functions perform well if the analogs to the base cases obtained through it are (at least in many cases) the same cases as the *best analogs* found to the verifying analysis (that follows the base cases). To produce this reference list of best analogs all the nine distance functions were used without any a priori preference (Fig. 1). The results for the individual distance functions, showing how many of the analogs

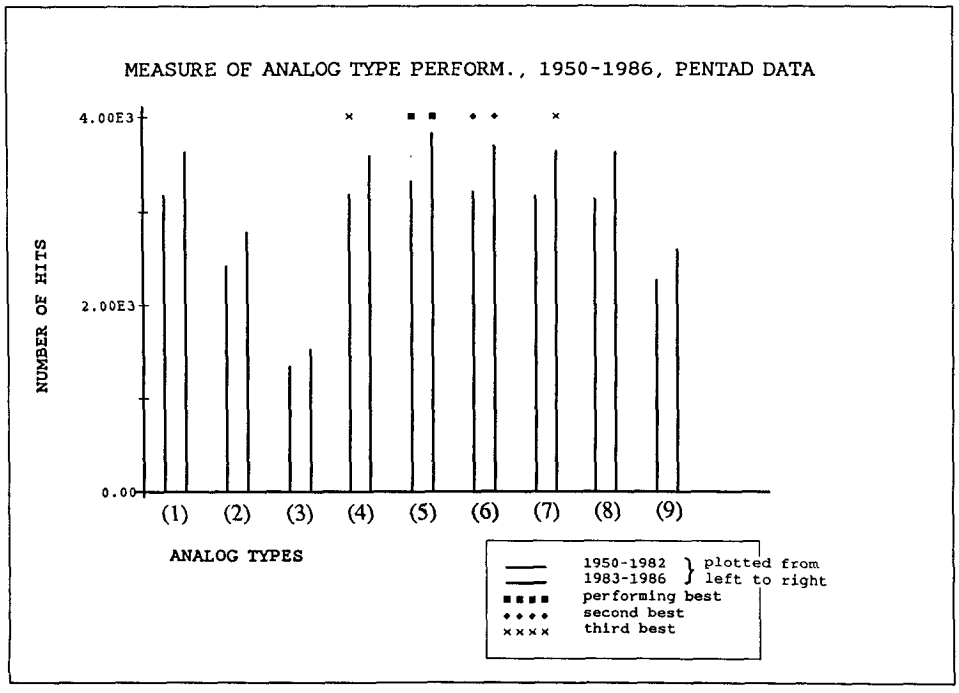

FIG. 6. Same as Fig. 4 except for accumulating the results for 1950–1982 and 1950–1986.

Unauthenticated | Downloaded 05/10/21 11:01 PM UTC

TABLE 2. Same as Table 1 except for pentad analog performance (see Fig. 6).

Distance functions	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	0	*	*	0.20	5.75	0.76	0.01	0.01	*
(2)		0	*	*	*	*	*	*	6.46
(3)			0	*	*	*	*	*	*
(4)				0	8.08	1.73	0.31	0.28	*
(5)					0	2.33	5.23	5.34	*
(6)						0	0.58	0.62	*
(7)							0	0.00	*
(8)								0	*
(9)									0

continue being good analogs in the forecast period as well, were summarized in Figs. 4 and 6 for daily and pentad data, respectively. The best similarity measure in both cases was the difference in the gradient of height (5). This dynamically oriented measure outperforms all other methods, including RMSD and COR. Especially on daily data, the use of RMSD, in accordance with theoretical considerations, is preferred to COR. The differences in the performance of individual similarity measures in this indirect comparison are statistically significant. However, no substantial difference was found between different distance functions when comparing the analog forecasts obtained by them directly to the appropriate analysis fields by using any of the distance functions.

As indicated earlier, the results of the intercomparison method based on analog forecasts in a strict sense show only which is the best distance function to use for producing analog forecasts. For other purposes, different distance functions might be more appropriate. However it seems intuitively right that a similarity measure that finds analogs giving the best forecasts focuses on important features of the flow and probably offers good (if not the best) performance in other applications too. In recent and yet unpublished studies, the author found some independent evidence that actually this might be the case. When comparing the RMSD and the difference in the gradient of height in other than forecast applications (identifying quasi-stationary states with daily data or investigating the phase space structure of time-averaged circulation data), the latter similarity measure seemed always to be more useful. Also, it captures the flow in more dimensions in the phase space than the RMSD does. This might seem confusing at the first look since both similarity measures are a function applied to the very same gridded data. However, the difference in the gradient of height seems to be able to extract more information from the data than the RMSD. This also gives a possible explanation why it performed better than any other similarity measure in this intercomparison study.

Acknowledgments. The data base of this study was kindly provided by the Climate Analysis Center of NMC. Huug van den Dool and Robert Livezey showed supportive interest and supplied helpful comments during the development of this project. Generous remarks by Anthony Barnston, Alan Murphy, Uwe Radok, and Colin Ramage on earlier versions of this manuscript also proved to be very useful. The computations were completed at the Cooperative Institute for Climate Studies, Department of Meteorology, University of Maryland, College Park, Maryland, while the author held a Fulbright Fellowship there. The University's facilities and the help of its staff are gratefully acknowledged. The manuscript was typed by Barbara Sloan, CIRES, University of Colorado.

REFERENCES

Arpe, K., 1978: The use and verification of ECMWF forecast results, a case study. *The Interpretation and Use of Large-Scale Numerical Forecast Products,* ECMWF Seminars, 343–369. [Available from ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, England.

Barnston, A. G., and R. E. Livezey, 1987: Classification, seasonality, and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea. Rev.,* **115,** 1083–1126.

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

Kruizinga, S., and A. H. Murphy, 1983: Use of an analogue procedure to formulate objective probabilistic temperature forecasts in the Netherlands. *Mon. Wea. Rev.,* **111,** 2244–2254.

Livezey, R. E., and A. G. Barnston, 1988: An operational multifield analog/antianalog prediction system for United States seasonal temperatures. Part 1: System design and winter experiments. *J. Geophys. Res.,* **93,** 10 953–10 974.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.,* **116,** 2417–2424.

——, and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

——, and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.,* **117,** 572–581.

Ruosteenoja, K., 1988: Factors affecting the occurence and lifetime of 500-mb height analogs: 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. First WMO Workshop on the Diagnosis and Prediction of Monthly and Seasonal Atmospheric Variations over the Globe,* College Park, Maryland, World Meteor. Org. WMO-TD-No. 87, 672–681.

Somerville, R. C. J., 1977: Pattern recognition techniques for forecast verification. *Contrib. Atmos. Phys.,* **50,** 403–410.

Teweles, S., and H. B. Wobus, 1954: Verification of prognostic charts. *Bull. Amer. Meteor. Soc.,* **35,** 455–463.

van den Dool, H. M., 1989: A new look at weather forecasting through analogs. *Mon. Wea. Rev.,* **117,** 2230–2247.

Unauthenticated | Downloaded 05/10/21 11:01 PM UTC