Long-Range Temperature Prediction Using a Simple Analog Approach[1]

ROBERT E. BERGEN AND ROBERT P. HARNACK

Department of Meteorology and Physical Oceanography, Rutgers—The State University of New Jersey, New Brunswick 08903

(Manuscript received 23 October 1981, in final form 20 April 1982)

ABSTRACT

A study has been made to assess the level of predictive skill associated with the application of a simple analog methodology to long-range temperature prediction over the continental United States in the period 1948–78. This approach relies solely on the pattern correlation statistic to select analogs from which monthly and seasonal temperature category forecasts for 68 climatic divisions (CD's) are subsequently derived. Numerous analog model trials were attempted, employing various combinations of predictor type and domain, forecast period, and forecast lead time. Predictor types include North Pacific sea surface temperature (SST), 700 mb height and 1000–700 mb thickness. The mean percent correct statistic was used to assess the spatial and temporal variations of skill for each analog model trial, as well as for persistence forecasts.

Principal conclusions include:

1) Overall mean percent correct scores for both monthly and seasonal analog models (using three categories) were, for the most part, slightly better than random chance and occasionally better than persistence. Highest overall scores were 45% correct for February forecasts using January 700 mb heights, and 40% correct for winter forecasts using November and fall SST. Counts of significant local skill exceeded chance expectation for many analog models tested.

2) Monthly analog models generally performed best during the period January–June, outscoring persistence and chance in many instances.

3) Seasonal analog models did best for the winter and summer seasons. Winter forecasts were most successful using Pacific SST, while similar results were obtained for summer, using SST or 1000–700 mb thickness. Seasonal analog models also performed well for spring, relative to random chance and persistence, particularly those using 700 mb heights. Thickness models using a forecast lag of one season appeared to be the best overall, with some combinations of domain and lag beating persistence for each season.

1. Introduction

It has long been observed that the gross features of synoptic-scale weather patterns over the Northern Hemisphere do, from time to time, repeat themselves. With this in mind, a variety of forecasting schemes have been developed using the analog as their basis. In this regard, two atmospheric states which are observed to resemble one another, with respect to one or more meteorological elements, are termed analogs. The analog method of forecasting employs the philosophy that weather behaves in such a way that the present initial conditions, if found to be similar to a past situation, will evolve in a similar fashion. Therefore, once two similar patterns are found on weather chart records, the assumption made is that their future development will be similar. This means that if a good analog can be found for the current atmosphere, a forecast could be obtained by using the sequence of previously observed atmospheric states as a reference, since the past evolution is already known.

The use of analogs is by no means a new concept in meteorology. In the past, a variety of analog schemes have been formulated, employing various predictors and analog selection criteria. Sea-level pressure distributions, air-mass movements, frontal speed and direction, and characteristics of the centers of action have all been used to provide the basis for analog selection (Radinovic, 1975). Although analogs have a long history in meteorology, this approach, until fairly recently, had been applied only to shorter-term predictions. It was believed that the size of the data sample for the necessary predictors was too severely limited for monthly or seasonal predictions.

In recent years, considerable study of the use of analogs in long-range forecasting has been carried out, particularly for the British Isles. The British Meteorological Office employed the analog methodology, as described by Bowen (1976), from 1963 until recently in the preparation of thirty-day forecasts. Distributions of mean monthly temperature and precipitation anomalies over the Northern Hemisphere were two of the factors considered. Another factor was the classification of weather types over the British Isles. In addition, the analog selec-

---

[1] Paper of the Journal Series, New Jersey Agricultural Experiment Station.

0027-0644/82/081083-17$08.25

tion process included North Atlantic sea surface temperature (SST), upper-air data from the 500 mb level, as well as indices summarizing broadscale synoptic characteristics. Each selection method yielded a list of potential analog years, which became selectively narrowed down after the poorer matches were rejected. The sequels in these years were then examined by a team of meteorologists and, if good agreement was found between the current and analog charts, a forecast was issued.

A verification scheme indicated that the British long-range forecasting project produced "good" or "excellent" forecast agreement more than 50% of the time for temperature predictions, while precipitation scores were somewhat lower (Bowen, 1976). Even so, 30-day forecasting in Great Britain has been recently discontinued, presumably due to insufficient skill.

Although the analog approach appears to be straightforward, this method is not without its pitfalls. Since the atmospheric state over large areas is never exactly the same, it is often difficult to find "good" analogs (Lorenz, 1969). Therefore, from a theoretical standpoint, the analog method has limited possibilities, since the analog used will still differ, in some ways, from the current observed condition (Namias, 1978). A long record of historic synoptic analyses is obviously desirable and will greatly increase the chances of finding a suitable analog. Therefore, it appears that the degree of success attainable using the analog method, may be dependent upon the extent of available meteorological records. Unfortunately, comprehensive records of the predictor types thought to be useful for applying the method to long-range forecasting, namely SST and upper-air data, exist only for the past thirty years or so. In view of the fact that one is dealing with a fluid atmosphere which demonstrates seemingly infinite variation, such a limited historical file, from which possible analogs are chosen, presents a difficult problem. This is an obvious disadvantage and one which may, presently, be the biggest drawback to the analog approach. Even assuming that this is not a problem, there is no guarantee that the evolution of even very similar weather patterns will be similar.

Despite these limitations, however, there are several distinct advantages enjoyed by the analog method, as pointed out by Namias (1951). Primarily, the analog approach can be applied objectively and, therefore, does not rely upon the complex and subtle reasonings inherent in physical/statistical methods which, today, are still far from perfected. Secondly, its mode of operation is such that forecasts can be prepared quickly, especially when high speed computers are employed for analog selection. In addition, microclimatological influences in an area (e.g., terrain influences) are automatically built into the analog method since past observed values are used to make forecasts. When considering potential analogs for selection, it would be unrealistic to anticipate an exact repetition of large-scale mean weather patterns since this occurrence would, indeed, be rare. However, enough similarity may exist between recurring patterns to perhaps obtain some useful skill for long-range forecasts. An objective approach to the use of analogs in long-range forecasting was introduced by Radinovic (1975), who used the method to project surface temperature and precipitation for periods of up to one month based on analogs selected from the 500–1000 mb thickness field over the Northern Hemisphere. When comparing chart records, the one having the same anomaly sign at the greatest number of points, as well as having the highest correlation coefficient with the current chart, was selected as the analog. His results indicate that this analog methodology realizes a forecasting skill "significantly superior" to that of persistence, for both temperature and precipitation, although this study was limited to only seven stations in Yugoslavia.

A unique approach to the analog theory, developed by Barnett and Preisendorfer (1978), utilizes a "climatic state vector" as its basic element. This vector traces the path of a "climate particle" through a multi-dimensional space and represents the time evolution of the climate system. This "particle" includes properties describing the regional covariability of various geophysical fields thought to be important when considering climatic variation. It is this multi-field approach which provides increased dimensionality to the analog selection process. By applying an empirical orthogonal function (EOF) analysis to each individual data field (i.e., 700 mb height, 1000–700 mb thickness, SST, etc.), the authors were able to identify, in physical space, certain "key regions" of active variability. These regions represent patterns selected from the EOF's which appear to differ significantly from those expected, using a field of randomly distributed variables. A weighting system is then applied, such that the data in each of these regions are weighted in proportion to the amount of their explained variance. Using three different methods of analog selection and employing forecast lead times of one to four seasons, seasonal temperature category predictions were made for thirty-three stations over North America. Statistically significant skill was obtained for various combinations of locale, season, forecast lag time, and analog selection method, although similar score comparisons with a competitive forecasting method, such as persistence, were not made. Generally, predictive skill was greatest for winter and summer, one season lag and when only the immediate past position of the climatic state vector was used for selection of the analog.

The main objective of this study is to assess the level of predictive skill associated with the application of a simple analog methodology to the prediction of long-range temperature anomalies in the continental United States. In this regard, "long-range" refers to lead times ranging from one month to several seasons. This analog method relies solely on the use of the pattern correlation statistic to select the closest match between the "current" data field and other similarly defined data fields. Lund (1963) used this statistic to classify sea level pressure maps for eastern North America.

Numerous analog model trials have been attempted by varying different aspects of the predictor field, such as type, spatial domain, amount of time-averaging, and lag time from period of predictor to that of predictand. Several verification scores have been computed in order to determine the level of skill relative to control forecasts, such as persistence and random selection, for each analog model.

Within the framework of the objectives outlined thus far, it was decided to employ the pattern correlation for the selection of analogs, since this parameter allows for a conceptually and statistically simple application of analog methodology. Since computational costs are relatively small when only pattern correlations are being computed for each analog selection, many experimental trials can be made by varying each aspect of this method. In addition, verification statistics of forecasts based upon simple analog selection criteria can then be compared to those produced using a more complex method, like that of Barnett and Preisendorfer (1978), in order to determine if the added complexity is beneficial. A variety of atmospheric and oceanic variables provided the basis for analog selection. These predictors, discussed in the next section, have been used previously in long-range forecasting attempts.

2. Data description

Since this study was aimed at the monthly and seasonal prediction of temperature anomalies, temperature statistics from climatic divisions (CD's), rather than individual stations, were employed as the predictand. Each state has previously been divided into several of these climatic divisions, the number depending on state size, with each division consisting of numerous (greater than ten) individual stations. CD's were used since CD temperatures are similar to area averages and should be more representative of the macroclimate than individual station data. In this study 68 CD's were selected for use from a total number of 345 within the conterminous United States (Fig. 1). Their selection was somewhat arbitrary since even-spatial distribution and climatic representation around the country were the main concerns. This data set consisted of monthly means for the entire period 1931–78 and was obtained from the National Climatic Center.

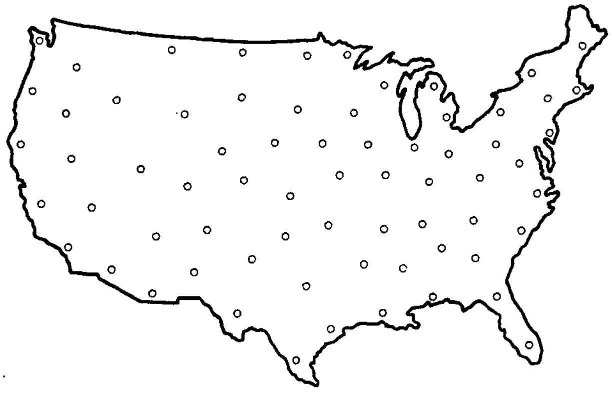

FIG. 1. Location of 68 CD's used to provide historical temperature data.

As previously stated, numerous analog forecast models were formulated, varying in terms of both the predictor type being used for analog selection or the spatial domain used to compute the pattern correlation. The following types of data were used to define predictor variables:

1) Pacific SST data, in the form of monthly means, were staggered 5 × 5° latitude–longitude grid point values for the period 1947–78. Data points chosen for use in this study were located in an area of the eastern North Pacific bounded by 120 and 170°W longitude, 25 and 55°N latitude. This data set was obtained from the Scripps Institution of Oceanography.

2) Mean monthly height values for the 700 mb level were also obtained from Scripps. This data set covers the period 1947–77 and corresponds to 10° latitude–longitude intersections over the Northern Hemisphere. The spatial domains of the 700 mb height field used here consisted of a North American (N.A.) grid, bounded by 20 and 70°N latitude, 50 and 150°W longitude; and a hemispheric (N.H.) grid, bounded by 20 and 80°N, 0 and 150°E.

3) Mean monthly values of sea-level pressure (SLP), for the period 1947–77, were obtained from a data set prepared by the National Center for Atmospheric Research (NCAR) for the period 1899–1977. The data points correspond to the same grid domains used for 700 mb heights.

4) Thickness values for the 1000–700 mb level were derived using a combination of SLP and 700 mb height data. Before determining the thickness field by computing the height difference between the 1000 and 700 mb levels, it was first necessary to convert the SLP data to 1000 mb heights. This was accomplished, following the same procedure employed by the Climate Analysis Center, using the formula

$$\text{height}_{(1000)} = -58.6(\bar{T})\frac{(1000 - \text{SLP})}{(1000 + \text{SLP})},$$

where $\bar{T}$ is defined as the long-term (1931–1960) monthly mean surface temperature for a specific grid point. The temperature data involved in these calculations were obtained from the Climate Analysis Center. Data points for the thickness field correspond to the same North American and hemispheric grids used for 700 mb heights.

3. Procedure

a. Data preparation

Before the analog forecast procedure could be tested, it was first necessary to prepare the original CD temperature data such that it would be compatible with a conventional forecasting format. A thirty-year period (1941–70) was used to establish long-term monthly means for each CD. These means were then subtracted from the actual temperature for each month in the period 1931–78, thereby converting the original raw data into a set of monthly anomalies. For the purpose of this study it was assumed that observed temperature values follow the Gaussian frequency distribution. As a result, all observed monthly values were then assigned to one of three temperature categories; either above, below, or near normal, based upon the appropriate statistical considerations. Temperature intervals were established such that each temperature observation would fall into one of the three temperature classes with an equal probability. This procedure was carried out for every CD so that each would have its own representative set of monthly temperature classes. Seasonal anomalies and classifications were also calculated in the same manner. It should be noted that seasonal in this context pertains to three-month averages as follows: winter (December, January, February); spring (March, April, May); summer (June, July, August); and autumn (September, October, November).

b. Analog selection methodology

As previously stated, the pattern correlation statistic, defined here as the correlation coefficient between two anomaly fields for a given predictor, provided the basis for analog selection. This value was used because it provides a good measure of the large-scale phase agreement between two fields of data. When considering two anomaly fields, a high positive correlation will result when anomalous areas of the same sign spatially coincide. Conversely, little or no coincidence of anomalous areas within the predictor field will yield lower, perhaps even negative values. The pattern correlation was calculated as

$$R_{12} = \frac{(N \sum X_1X_2) - (\sum X_1X_2)}{[(N \sum X_1{}^2 - \sum X_1 \sum X_1)]^{1/2}[(N \sum X_2{}^2 - \sum X_2 \sum X_2)]^{1/2}},$$

where X_1 and X_2 represent the predictor anomaly field at times t_1 and t_2, respectively, and N is equal to the number of grid points within the predictor field. Anomaly fields are used when computing this value to eliminate spurious correlation due to climatological variations of a parameter with latitude.

The initial step in the analog selection process involves scanning forward and backward in time, comparing a particular month or season type to all other similar periods within the time series. For example, if February forecasts are desired, with a lead time of one month, then the January predictor field for the first year (target year) in the record is correlated with that of every remaining January. Note here that "target" refers to any period for which a forecast is desired. This will yield a set of correlation values from which the analog year is ultimately selected. This same procedure is eventually repeated for every other January remaining in the data set, as new "target" Februaries are considered.

This study followed two basic procedures for making temperature forecasts based on analog selection. The simplest approach used the *observed* condition for the appropriate period in a *single* year (analog year) as the *forecast* for the same period in the target year. The analog year is selected on the basis of that year having the highest positive pattern correlation out of the thirty or so that have been computed for the given predictor and month/season types, regardless of the correlation magnitude. For example, assume that CD temperature forecasts are required for the spring of year *A*, the target year, and winter SST's are to be used as the basis for analog selection. After correlating the SST field for the winter of year *A* with that of every other winter in the record, the specific winter having the highest correlation value is selected as the analog. If the analog winter is year *B*, spring temperature *forecasts* for each CD in the target year (*A*) are taken from the *observed* temperatures for the same CD in the spring of year *B*.

The second approach considered the observed conditions in *three* years to formulate the forecast for the target year. In this case, the years corresponding to the three highest pattern correlation values are selected as the analogs. If analog winters are years *B*, *C*, and *D*, spring forecasts for year *A* are based upon a concensus of spring conditions observed in the analog years. This means that, for each CD, an observed spring temperature category, common to any of the three analog years, becomes the forecast for the spring of year *A*. If no similarity exists, the condition observed in the spring of that year having the highest pattern correlation, for example year *B*, becomes the forecast for the spring of *A*. Since resultant forecasts are generally based upon the consideration of more than one analog, the "three best" approach may possibly be more reliable than the previously described "one best" approach.

c. *Forecast verification*

To determine whether analogs selected on the basis of simple pattern correlations can effectively compete against those obtained using a more complex selection criteria, like that of Barnett and Preisendorfer (1978), some similar skill scores were used. Therefore, both "local" and "global" skill scores, like those used by Barnett and Preisendorfer, are computed for each analog trial. Local skill denotes the percentage of correct temperature category forecasts for a particular station or region when considering all forecast cases. This score is not a function of time and, therefore, demonstrates the spatial variability of skill. Global skill denotes the percentage of correct forecasts for a particular year, considering all stations or regions being forecast. This score is a function of time, allowing the temporal variability of skill to be determined. Each of these types of skill measure were applied to monthly or seasonally stratified forecast cases.

These scores, although describing either local or global skill, offer no indication of significance. Additionally, the interpretation of obtained skill scores is generally accomplished by comparing them to some standard. For these reasons the analog model forecast scores were compared with similar scores obtained using various "blind" forecasts, such as persistence and random chance. Persistence forecasts were obtained for all analog forecast periods.

For any forecast scheme to be of value it must significantly outperform random chance in forecasting skill. Barnett and Preisendorfer employ the use of a "stochaster", a forecaster making predictions purely by chance, to reach into the historical file and randomly select analogs. Using the binomial distribution, it becomes possible to ascertain the probability $p(X)$ of the stochaster attaining X correct forecasts out of N trials. This was applied as

$$p(X) = \frac{N!}{X!(N-X)!}\alpha(1-\alpha)^{N-X}, \quad (1)$$

where $\alpha = (N/3 - 1)/(N - 1)$, and is defined as the probability of a successful prediction by the stochaster. To find the X value that has significance at the 95% level, as defined by the above distribution, the cumulative probability is computed such that

$$q(X) = \sum_{X=0}^{X_1} p(X) \approx 0.95, \quad (2)$$

and then spotting the integer X_1 that coincides most closely to the 95% confidence level. Significance at this level means that the probability of a purely random process attaining X_1 correct forecasts is only 5%. If the forecaster attains an X_1 value that lies at, or above this level, then it may be concluded that the forecaster has significant local skill. For instance, by applying Eqs. (1) and (2) to a set of thirty forecasts that were made for a particular month/season type and CD, a minimum score of 44% correct would be necessary for the local skill at that CD to be deemed significant. Additionally, in order to determine if a set of forecasts, made over the entire field (i.e., 68 CD's), is significantly better than chance, the binomial distribution was used again.[2] Due to sampling variation, the number of CD's required to achieve "significant" local skill, in order to declare "field significance" at the 5% level, is >0.05 × 68 (or 3.4). Also, since there is considerable spatial coherence of monthly/seasonal temperature, the effective number of degrees of freedom (DoF) is <68. Based on the work of Diaz and Fulbright (1981) and Diaz (1981), the number of DoF for United States surface temperature was set at five. This is approximately the number of empirical orthogonal functions needed to explain about 95% of the variance of *seasonal* temperature. Normally this number would vary somewhat by time of year and by averaging period length. Since the exact number of DoF for each situation encountered in this study could not be obtained from the studies cited or from the other available literature (to the best of the author's knowledge), DoF equal to five was *estimated* and applied uniformly to each analog trial. It is believed that this is the minimum number of DoF and, therefore, would make significance testing for the "field" quite stringent. In Eq. (1), N was set at five and α was set at 0.05 (to be consistent with the 5% level used to assess local skill). In Eq. (2), $q(X)$ remained at 0.95. The integer X_1 that most closely coincides with the 95% confidence level was determined to be 12 after using Eqs. (1) and (2). Therefore, at least 12 CD's out of 68 are required to achieve significance in local scores, so that field significance be achieved.

Significance testing was also performed on the overall global scores, which were scores for all forecasts (~30) and CD's (68) combined for a given analog trial. Again the binomial distribution was used [Eqs. (1) and (2)]. The effective number of DoF, N, was set at 150 and $q(X)$ at 0.95. The N = 150 was based on DoF equal to five for each year and 30 years of forecasts for each analog trial. Independence between years was assumed. Given this,

---

[2] Based on an unpublished manuscript and presentation by Robert Livezey and W. Chen (Climate Analysis Center), at the Sixth Climate Diagnostics Workshop, Palisades, NY, 14–16 October 1981.

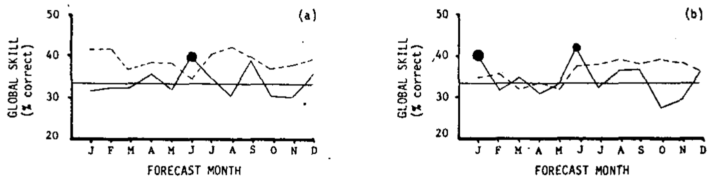

FIG. 2. Overall global skill obtained by persistence (dashed), and analog models (solid) using Pacific SST with forecast lead times of (a) one month and (b) two months. Dots denote significance at 5% level.

an overall global score of 39% was needed for significance at the 5% level.

Other verification techniques, such as the application of a conventional skill score and the reduction of error score, as well as a two-category forecast verification based on anomaly sign, were also performed. Additional measures of skill computed include a skill score based on the distribution of class errors, percentage of two class errors given a forecast "extreme," and the percentage of correct class forecasts given an observed "extreme." Only the mean percent correct statistic for the three-category forecasts are presented in this paper.

4. Results

The results presented here are divided into monthly and seasonal forecast cases, with several types of verification statistics shown for each. Overall global skill and counts of significant local skill are examined for each analog model, comparing these results to both random chance and persistence. Distributions of local skill are also examined to determine areas of relative predictability as demonstrated by the analog method. Numerous analog models have been formulated, employing various combinations of predictor type and domain, forecast period and forecast lead time. However, computational costs would have become excessive if every possible combination had been attempted, despite the fact that the analog approach presented here is relatively inexpensive to test. It was therefore necessary to limit analog model trials by using just three types of predictors (1000–700 mb thickness, 700 mb heights, and Pacific SST), two spatial domains for the predictors, and two types of forecast lags. Preliminary results suggested that the "three-best" approach would yield superior forecasting skill to the "one-best" approach, in the majority of cases. As a result, the "three-best" approach was employed exclusively in analog trials reported here.

a. *Monthly predictions*

1) OVERALL PERFORMANCE

Monthly global skill scores are presented in Figs. 2–5 for the various analog models used in this study. Global skill scores obtained using the persistence method are also shown for comparison. These scores represent the mean percent correct for all CD's and years combined. A score of 33% correct has been highlighted to demonstrate how well both forecasting methods perform relative to random chance. Although the analog scores fluctuate about the random score, it appears that the analog method, on the whole, performed slightly better than random chance. Significance at the 5% level was achieved for only a few months for each analog model. The number of such months varied, with the N.H. heights model (one-month lead) having four such months (February, April, June and October), while the two-month lead thickness models had none. The other models had one or two months whose overall global score exceeded 39%, and therefore could be deemed significant at the 5% level given the assumed number of DoF. It should be emphasized again that the number of DoF was unknown, so that if DoF equalled, say 270, then the required percent correct would drop about 1%, which would have made several more

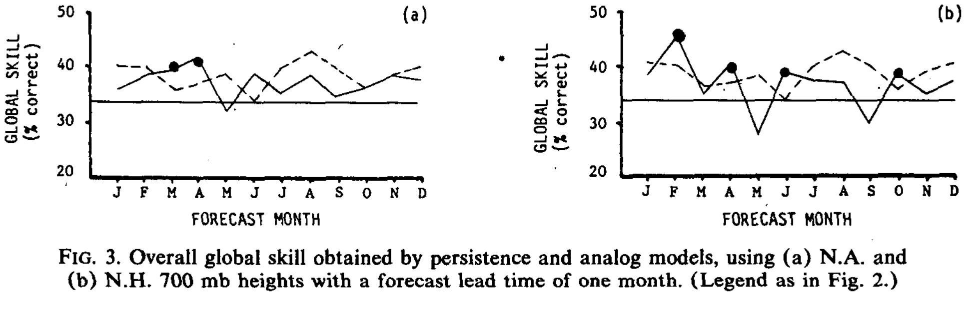

FIG. 3. Overall global skill obtained by persistence and analog models, using (a) N.A. and (b) N.H. 700 mb heights with a forecast lead time of one month. (Legend as in Fig. 2.)

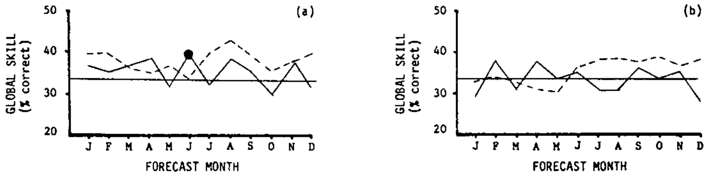

FIG. 4. Overall global skill obtained by persistence and analog models, using N.A. 1000–700 mb thickness with forecast lead times of (a) one month and (b) two months. (Legend as in Fig. 2.)

months significant for the various models. Persistence scores are better than the chance scores for most months and generally superior to the analog scores. For both methods, large fluctuations in terms of percent correct often occur from one month to the next with no apparent physical explanation in most cases. Persistence scores were generally highest during the summer months and lowest during the spring. For one-month lead forecasts, those analog models using 1000–700 mb thickness and 700 mb heights as predictors generally outperform those using SST. Although overall skill levels are modest (i.e., generally <40%), thickness and height models scored well, relative to persistence and chance, for the period February–June. It is interesting to note that June temperature forecasts tend to be the best overall, regardless of the predictor type employed, when a comparison to persistence is made. This results from the fact that persistence performed poorly for June while the analog models generally performed above their average.

The performance of models using thickness and heights were simlar for lead times of one month, generally outperforming SST models with the same lag. When considering 1000–700 mb thickness and 700 mb heights, there is no obvious advantage for using either predictor domain. However, for lead times of one month, the North American domain yields slightly better results. Results for thickness and SST models using a two-month lag were generally poor. Therefore, a model using heights at the longer lag was not attempted. In summary, these graphs suggest that overall global skill is small for the analog method and only occasionally better than persistence for monthly forecasts. Additionally, no large difference in predictor performance is noted.

The number of CD's for which monthly local skill was significant at the 95% level for each combination of month and analog model type is given in Table 1. These are simply counts indicating how many CD's satisfy the criteria for significance, as established by the binomial distribution. These counts are given for each analog model, as well as for persistence, using forecast lags of one and two months. It should be mentioned that persistence counts for the same lag vary with each predictor since slightly varying time periods were covered by the respective data sets. As a result, slightly different significance criteria were established for each predictor used.

These results seem to indicate that both the analog and persistence methods demonstrated some real forecasting skill relative to chance since a purely random distribution of skill would dictate a 5% chance that 12 out of 68 CD's would meet the significance criteria as discussed earlier. A large portion of the model type–month combinations are significantly better than random chance, but a much smaller number are both statistically significant and better than simple persistence. The latter cases are underlined in Table 1. In this regard, models did best in the period January–June; especially March, April and June. The particularly "poor" months were May and October.

With the exception of a few months, SST models employing a one-month forecast lag did not perform well, with June being the only month to beat persistence. Model performance improved greatly for SST when the forecast lag was increased to two months,

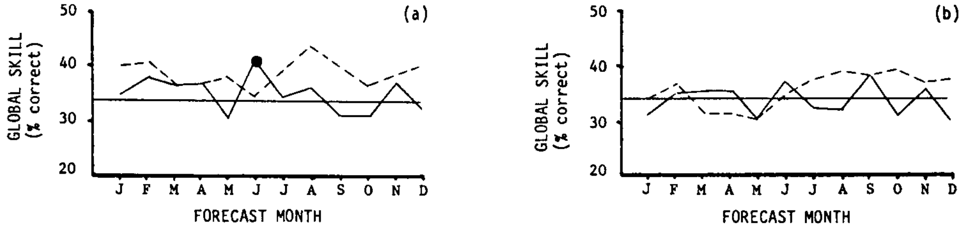

FIG. 5. Overall global skill obtained by persistence and analog models, using N.H. 1000–700 mb thickness with forecast lead times of (a) one month and (b) two months. (Legend as in Fig. 2.)

TABLE 1. Number of CD's out of 68 for which monthly local skill was significant at the 95% confidence level. Values in italic denote those models which are both significantly better than random chance (5% level) and better than persistence.

Model type	Lag	D	J	F	M	A	M	J	J	A	S	O	N
Pacific SST	1 Month	21	4	12	9	15	7	*26*	16	9	25	6	7
Persistence	1 Month	26	29	32	15	19	18	12	29	38	26	15	21
Pacific SST	2 Month	15	*37*	*18*	*15*	9	11	*35*	15	16	*26*	3	5
Persistence	2 Month	18	11	15	4	13	5	13	23	24	20	21	12
N.A. 1000–700 mb thick.	1 Month	13	18	22	*18*	*29*	10	*32*	17	28	20	8	22
N.H. 1000–700 mb thick.	1 Month	11	17	23	*24*	*23*	7	*33*	13	23	11	10	16
Persistence	1 Month	27	25	34	17	18	17	13	29	41	35	19	23
N.A. 1000–700 mb thick.	2 Month	5	3	19	*14*	*25*	*13*	*16*	12	10	17	17	*16*
N.H. 1000–700 mb thick.	2 Month	9	8	13	*15*	*16*	8	*21*	14	12	20	12	*16*
Persistence	2 Month	20	11	19	4	11	5	13	26	27	23	23	12
N.A. 700 mb heights	1 Month	*23*	18	21	*20*	*27*	10	*24*	16	20	6	8	19
N.H. 700 mb heights	1 Month	13	17	*42*	*14*	*30*	3	*26*	18	19	10	*15*	9
Persistence	1 Month	20	29	28	11	17	15	10	22	35	25	14	19

generally beating persistence for the period January–March, as well as June and September. The improvement of skill with lead time is, of course, unusual and is undoubtedly related somewhat to the thermal inertia of the oceans.

Analog models using 1000–700 mb thickness and forecast lags of one month performed well during the spring and early summer period, particularly April and June, beating persistence in most cases. Furthermore, a North American predictor domain seemed to be more successful for thickness model types. Even using a forecast lag of two months, thickness models outperformed persistence for the period March–June, although overall model performance eroded for the longer lag.

Analog models using 700 mb heights and a forecast lag of one month also generally performed well for the period December–June, beating persistence in the majority of cases. In fact, the N.H. heights model for February achieved the single highest count for local significance, beating the highest persistence count found for any month. Similar models were not attempted using a two month lag based on the poor showing of thickness models at the same lag. There was no clear advantage for using either predictor domain when 700 mb heights provide the basis for analog selection.

Considering all predictor types on a seasonal basis, analog models using 700 mb heights produced the best overall results for the winter months, beating persistence in December and February. Model performance using SST and a forecast lag of two months was exceptional for January predictions, easily outscoring persistence. Both atmospheric predictors produced good results for the period March–June using a one-month lag. June forecast models performed best during this period, with the analog method beating persistence in each trial. The SST model (two-month lag) performed particularly well for June. Otherwise, analog performance was generally poor, relative to persistence, for the remainder of the year.

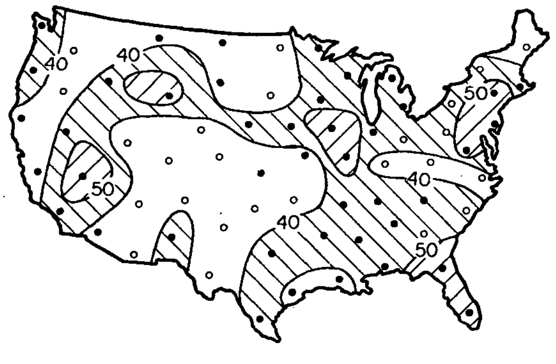

FIG. 6. Distribution of local skill obtained for January by analog model using November Pacific SST. Filled circles indicate CD locations where local skill exceeds that of persistence and random chance. (Note: isopleths for <40% not shown)

2) GEOGRAPHICAL VARIABILITY OF SKILL

The results presented in this section are displayed as maps indicating the spatial variability of local skill. Since global scores represent an area average encompassing the entire forecast domain, important areas of predictability may be disguised. Therefore, it is necessary to examine the spatial distribution of local skill for possible coherent areas of higher skill (i.e., >40%). Approximately 30 forecasts are verified to produce each local score. These maps, presented in Figs. 6–13, clearly demonstrate that the forecasting abilities of the analog method are spatially dependent, as might be expected from the results presented in previous studies. (In areas of no skill, isopleths for <40% are not shown.) For purposes of presentation, maps of local skill were selected and

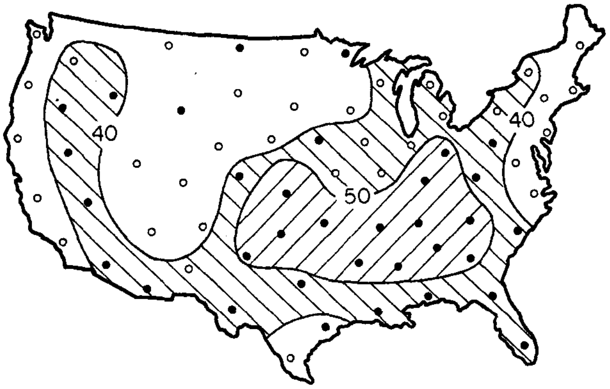

FIG. 7. Distribution of local skill obtained for April by analog model using March N.H. 700 mb heights. (Legend as in Fig. 6.)

FIG. 9. Distribution of local skill obtained for October by analog model using September N.H. 700 mb heights. (Legend as in Fig. 6.)

are shown for the best analog model in terms of forecast performance for the months of January, April, July and October. In addition, maps are shown for the overall best case (February using N.H. 700 mb heights in Fig. 10) and for several June models. Map contours are labeled in terms of local skill (percent correct). Locations whose local score exceeds both that of persistence and chance are denoted on each map.

The most notable feature of the maps presented here are the large coherent area of >40% correct obtained for April by the analog model using March N.H. 700 mb heights as the predictor (Fig. 7). This is in direct contrast to the study by Madden and Shea (1978), which suggests that the potential predictability for April is low for most of the United States. However, the areas of highest potential predictability for July and October, as indicated by this same study, have good agreement with the areas of greatest local skill obtained by analog models using 700 mb heights as the predictor (Figs. 8 and 9). Good agreement was also found to exist for January using Pacific SST and a two month lag (Fig. 6). This particular model was the best at forecasting January temperatures, regardless of lag, producing a large area of >40% correct over the eastern half of the country.

As previously stated, the performance of monthly analog models was most outstanding for June. Large coherent areas of local skill were demonstrated by models using each predictor type. Forecasting skill for June seemed to be particularly good in the central portions of the country. These results are shown in Figs. 11-13.

b. Seasonal predictions

1) OVERALL PERFORMANCE

Global skill scores obtained by seasonal analog models are graphed with similar scores for persistence in Figs. 14-18. The relative overall abilities of both forecasting methods were similar to those obtained for monthly predictions, with the analog method performing only slightly better than chance

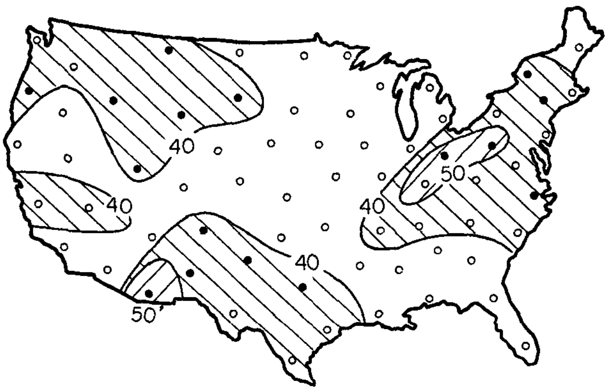

FIG. 8. Distribution of local skill obtained for July by analog model using June N.H. 700 mb heights. (Legend as in Fig. 6.)

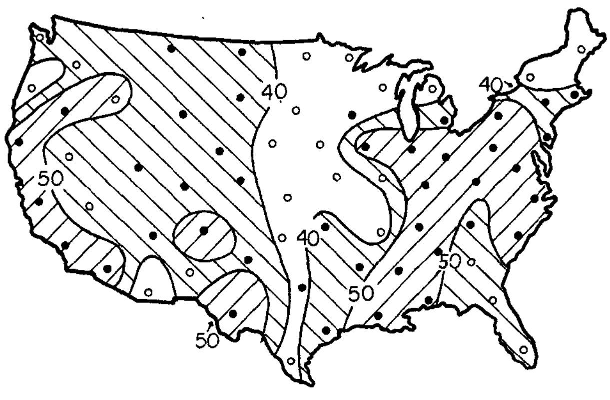

FIG. 10. Distribution of local skill obtained for February by analog model using January N.H. 700 mb heights. (Legend as in Fig. 6.)

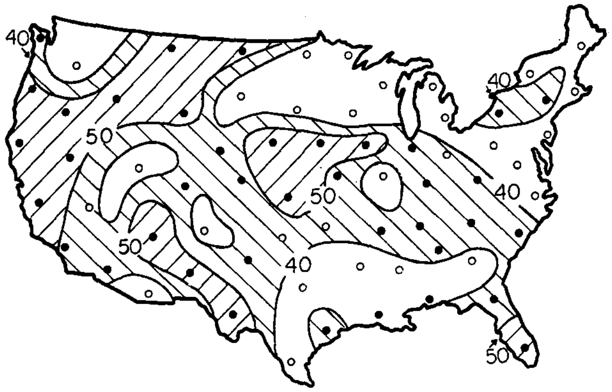

FIG. 11. Distribution of local skill obtained for June by analog using April Pacific SST. (Legend as in Fig. 6.)

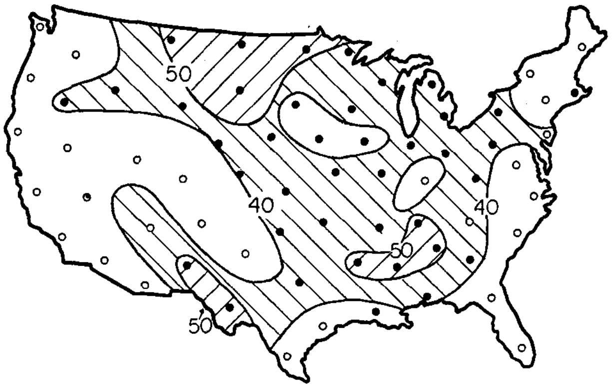

FIG. 13. Distribution of local skill obtained for June by analog model using May N.A. 1000–700 mb thickness. (Legend as in Fig. 6.)

and generally not as well as persistence. Model type-season combinations which had scores significantly better than ramdom chance include: Pacific SST (one month and one season lead) and N.H. thickness (two season lead) for winter; N.A. 700 mb heights (one month lead) for spring; N.H. thickness (one season lead) for summer; and N.A. thickness (one season lead) for autumn. In all of these instances, and in a few others, especially in spring, the analog model beat persistence. As might be expected, persistence scores were somewhat lower on a seasonal time scale than on a monthly time scale, particularly during the winter and summer seasons. The lowest scores, for either method, were generally observed during the spring. Models using analogs based on 700 mb heights were the important exception (Figs. 15 and 16).

The highest overall global scores obtained by the analog method were observed for the winter season using SST's as the analog selector and forecast lead times of one month or one season (Fig. 14). The overall mean percent correct was 40% for winter temperature forecasts based on November or fall SST. In both cases, the SST models beat persistence. For SST models, however, analog performance declined considerably when a forecast lag of two seasons was employed. With the exception of summer forecasts using lags of two seasons, all other SST models performed rather poorly when compared to persistence. Thickness models also performed well for winter with forecast lead times of one and two seasons, equalling or beating persistence in most cases. Note that model performance fell sharply when only a one month forecast lag was considered (Figs. 17 and 18).

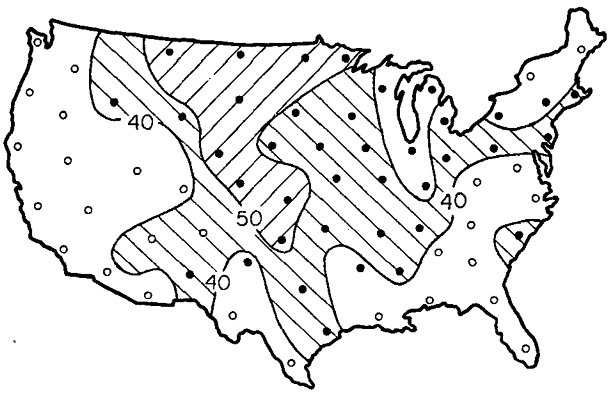

FIG. 12. Distribution of local skill obtained for June by analog model using May N.H. 700 mb heights. (Legend as in Fig. 6.)

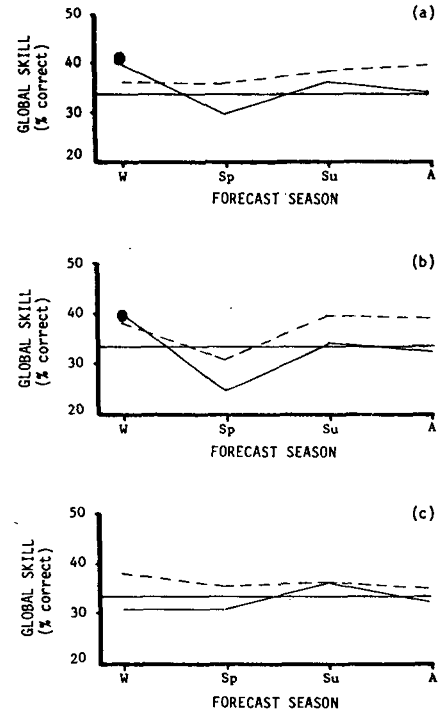

FIG. 14. Overall global skill obtained by persistence and analog models, using Pacific SST with forecast lead times of (a) one month, (b) one season and (c) two seasons. (Legend as in Fig. 2.)

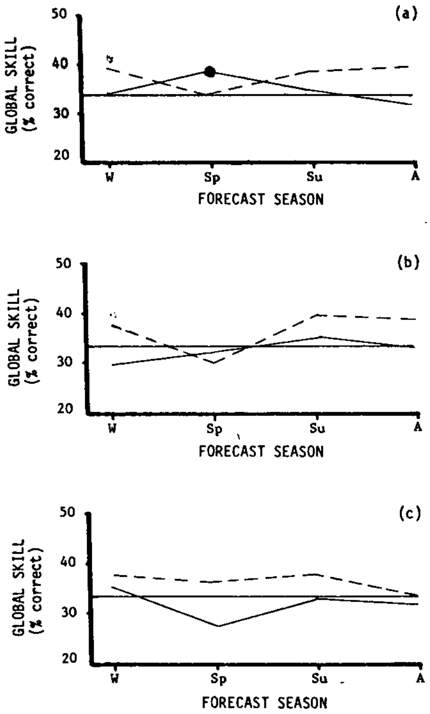

FIG. 15. Overall global skill obtained by persistence and analog models, using N.A. 700 mb heights with forecast lead times of (a) one month, (b) one season and (c) two seasons. (Legend as in Fig. 2.)

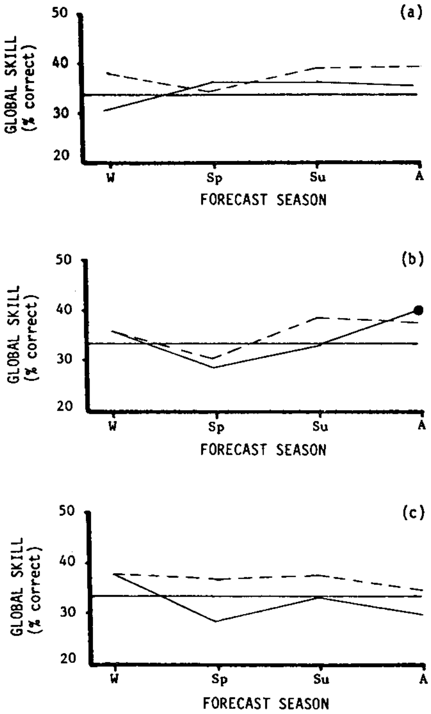

FIG. 17. Overall global skill obtained by persistence and analog models, using N.A. 1000–700 mb thickness with forecast lead times of (a) one month, (b) one season and (c) two seasons. (Legend as in Fig. 2.)

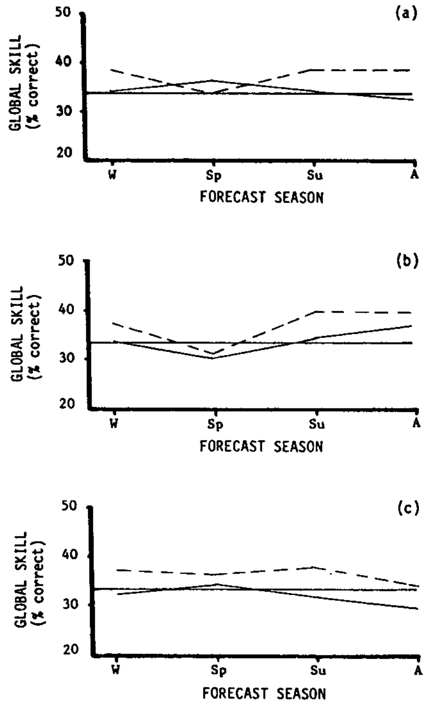

FIG. 16. Overall global skill obtained by persistence and analog models, using N.H. 700 mb heights with forecast lead times of (a) one month, (b) one season and (c) two seasons. (Legend as in Fig. 2.)

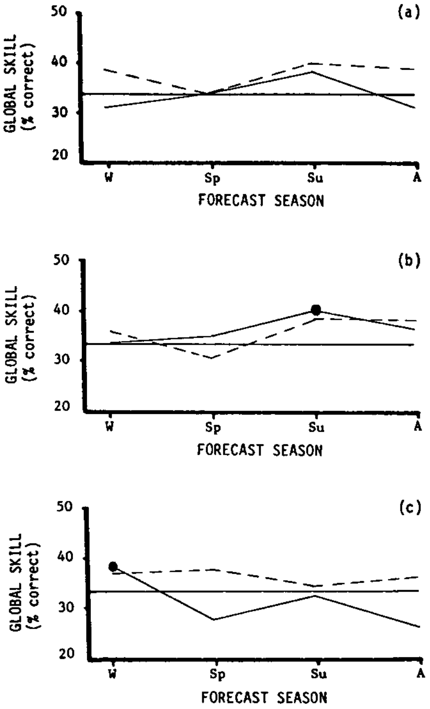

FIG. 18. Overall global skill obtained by persistence and analog models, using N.H. 1000–700 mb thickness with forecast lead times of (a) one month, (b) one season and (c) two seasons. (Legend as in Fig. 2.)

TABLE 2. Number of CD's out of 68 for which seasonal local skill was significant at the 95% confidence level. Values in italics denote those models which are both significantly better than random chance (5% level) and better than persistence.

Model type	Lag	Winter	Spring	Summer	Autumn
Pacific SST	1 Month	*21*	8	*32*	12
Persistence	1 Month	17	10	21	27
Pacific SST	1 Season	*29*	1	9	12
Persistence	1 Season	16	10	21	27
Pacific SST	2 Season	10	13	*19*	*13*
Persistence	2 Season	23	20	13	10
N.A. 700 mb heights	1 Month	14	*24*	17	6
N.H. 700 mb heights	1 Month	12	*14*	12	11
Persistence	1 Month	26	6	18	24
N.A. 700 mb heights	1 Season	7	11	15	11
N.H. 700 mb heights	1 Season	16	9	12	17
Persistence	1 Season	19	3	18	22
N.A. 700 mb heights	2 Season	19	6	10	11
N.H. 700 mb heights	2 Season	11	16	9	6
Persistence	2 Season	24	20	15	10
N.A. 1000–700 mb thick.	1 Month	11	*16*	21	21
N.H. 1000–700 mb thick.	1 Month	13	*12*	25	10
Persistence	1 Month	26	8	25	32
N.A. 1000–700 mb thick.	1 Season	*17*	9	13	*29*
N.H. 1000–700 mb thick.	1 Season	*15*	*18*	*31*	19
Persistence	1 Season	14	7	26	25
N.A. 1000–700 mb thick.	2 Season	*21*	6	14	11
N.H. 1000–700 mb thick.	2 Season	*20*	7	14	6
Persistence	2 Season	18	20	26	12

Skill scores obtained for the spring, using either the analog or persistence method, were generally the lowest observed for all seasons. However, the analog model did demonstrate an ability to skillfully forecast spring temperatures using analogs based on February 700 mb heights (Fig. 16, top). When considering all predictor types, the scores using 700 mb heights for seasons other than spring were generally the lowest obtained. Analog models based on the thickness field from the previous season also outscored persistence on a global basis for the summer and autumn, using hemispheric and North American predictor domains, respectively (Figs. 25 and 26, middle).

Counts indicating the number of CD's for which seasonal local skill was significant at the 95% level are given in Table 2. These results were formulated in the same manner as those given in Table 1 for each analog model and persistence. Both methods again demonstrate a degree of forecasting skill relative to chance, since counts of local significance are higher, in a large number of cases, than those expected from a purely random process. In a much smaller number of instances models had both significant forecasting skill *and* were superior to simple persistence. These are denoted in Table 2.

As was the case for monthly predictions, persistence often outperformed the analog method in terms of the number of CD's achieving significant local skill, although significance counts for persistence are somewhat lower on a seasonal basis. Again, the different persistence counts for a particular lag are a reflection of the different significance criteria established by the binomial distribution for each predictor. The lowest counts for either method were observed for spring season forecasts, although several analog models performed well relative to persistence for spring. Results also indicate that generally more CD's demonstrate significant local skill using persistence when seasonal forecasts are based on temperatures from the previous month rather than the previous season. It is notable that significance counts obtained for persistence using a two season lag for winter and spring forecasts are much higher than those obtained using a one season lag.

Considering forecasts with lead times of one month and one season only, SST models performed well for the winter, particularly when forecasts were made one season in advance. Both of these models using SST beat persistence. Summer forecasts improved greatly when May SST was used in place of spring SST and again beat persistence. SST model performance was generally poor for the spring and fall, using either lag.

Spring forecasts using February 700 mb heights were, by far, the best for this predictor type and considerably better than persistence. Otherwise,

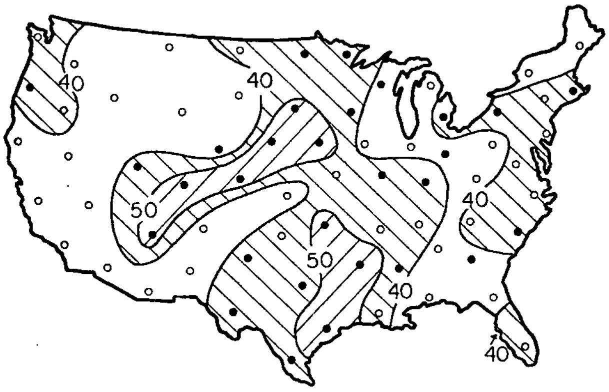

FIG. 19. Distribution of local skill obtained for winter by analog model using autumn Pacific SST. (Legend as in Fig. 6.)

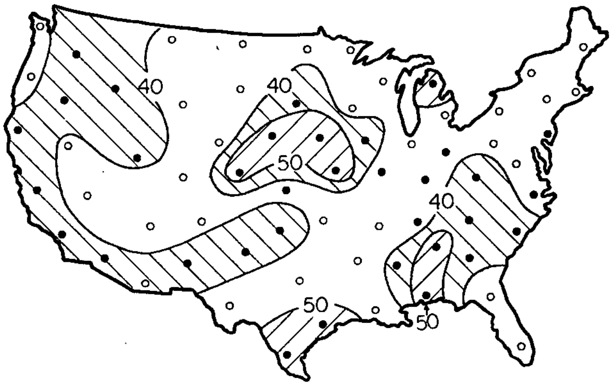

FIG. 20. Distribution of local skill obtained for spring by analog model using February N.A. 700 mb heights. (Legend as in Fig. 6.)

model performance using heights and lags of either one month or one season was generally unimpressive.

Analog models using 1000–700 mb thickness outscored persistence for every season when forecasts were made one *season* in advance. However, relative model performance in terms of predictor domain was not consistent, varying with the season being forecast. Spring forecasts based on thickness with a one *month* lag also beat persistence, although thickness models at this lag generally did not perform as well as those using a one season lag. Of the three predictor types employed, thickness performed best *overall.*

For the shorter lag types (i.e., one month/season), the analog model using autumn SST provided the best results for winter forecasts. The best results for spring were obtained using February 700 mb heights. The best summer forecasts were made by analog models employing SST and thickness as predictors, with forecast lead times of one month and one season, respectively. Autumn forecasts were most successful when the thickness field from the previous season provided the basis for analog selection. Analog models employing the various combinations of predictor and lag presented above produced the highest counts of local significance, relative to those of persistence, on a seasonal basis.

Model performance using SST improved considerably for spring and summer when forecast lead times were increased from one to two seasons. This model beat persistence for summer forecasts. However, model performance fell sharply for the winter using the longer lag. Results using 700 mb heights and a forecast lead time of two seasons were unimpressive for every season. Thickness models for winter using a two season lag showed improvement over similar models with shorter lags and beat persistence, when either predictor domain was used. Model performance for the remainder of the year, however, declined considerably when a two season lag was employed. Considering a lead time of two seasons for winter forecasts, thickness models performed the best of all predictors used and beat persistence. Spring forecasts at this lag were poor relative to persistence, for each predictor. Summer forecasts based on the winter SST field were better for summer than two-season persistence, while similar results were obtained by analog models and persistence for autumn.

2) GEOGRAPHICAL VARIABILITY OF SKILL

As previously stated, global skill scores represent an area average and are not necessarily an accurate gauge of how useful a particular forecasting method may be. Maps displaying the spatial variability of seasonal local skill are presented in Figs. 19–23. The best model for each season is shown. Map contours are again labeled in terms of mean percent correct, with locations outscoring persistence also being indicated. It can be seen that the forecasting abilities of the analog method on a seasonal basis are strongly spatially dependent, as was the case for monthly skill presented earlier.

The best winter model, using autumn Pacific SST,

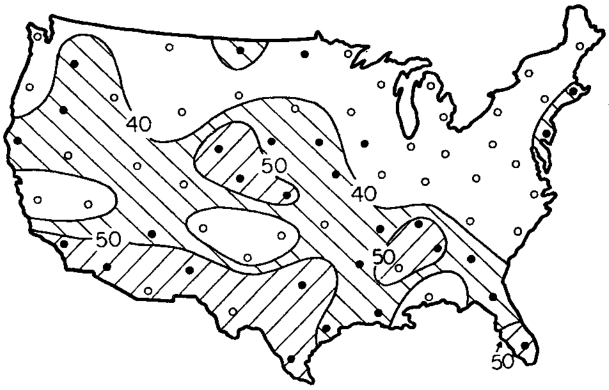

FIG. 21. Distribution of local skill obtained for summer by analog model using May Pacific SST. (Legend as in Fig. 6.)

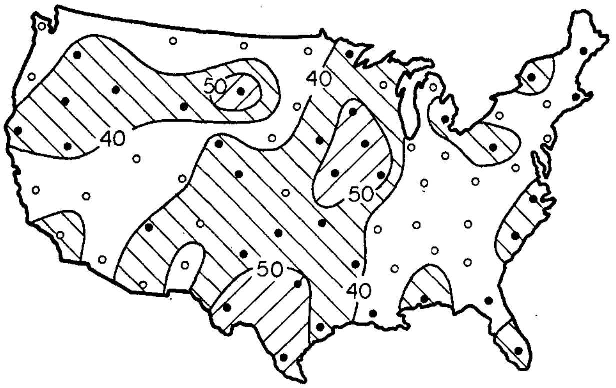

FIG. 22. Distribution of local skill obtained for summer by analog model using spring N.H. 1000–700 mb thickness. (Legend as in Fig. 6.)

produced large coherent areas of higher skill over the greater portion of the central and eastern United States (Fig. 19). In contrast, Barnett and Preisendorfer (1978) found the largest area of high skill for their best winter analog model to be located over the western portion of the country. Both studies suggest, however, that skillful predictions can be made along the east coast. Other model types employed for winter forecasts (not shown) produced smaller areas of skillful predictions over the Central Plains and Great Lakes region. A study of the potential predictability of seasonal temperatures by Barnett (1981) indicates that the areas of highest predictability are found along the west coast, southeastern United States, and the northern tier of states during winter.

Spring forecasts using February N.A. 700 mb heights produced coherent areas of skill over the southeastern portion of the United States (Fig. 20) and over portions of the western states and Central Plains.

For summer, SST and N.H. thickness models using a one month and one season lead, respectively, produced large coherent areas of higher skill over the southern and central states (Figs. 21 and 22). Fig. 21 in particular has a distribution of skill that is suggested by the potential skill portion of the study by Barnett (1981).

Autumn forecasts generated by using N.H. thickness and employing a one *season* lag were most successful over large portions of the Northern Plains and mid-Atlantic states (Fig. 23). Analog models using predictors from the previous *month* (not shown) produced no important areas of skill for autumn.

Maps indicating the spatial distribution of local skill are presented by Barnett and Preisendorfer (1978) for their "best" winter and summer analog models (see their Figs. 8–10). For their best winter model, 46% of the 24 forecast verification points lying within the conterminous United States achieved a minimum local skill of 40% correct. Their best summer model places 58% of the verification points inside the area of 40% correct. It should be pointed out that their local skills were based on 24 forecasts. Of the 68 CD's used here for verification purposes, 56 and 53% achieved local skills of at least 40% correct for the "best" winter and summer analog models, respectively. Based on this comparison, it appears that the simple analog approach can effectively compete against a more complex analog methodology, like that of Barnett and Preisendorfer (1978). A more precise comparison is not possible at this time, due to differences in study methodology and presentation.

3) TEMPORAL VARIABILITY OF SKILL

By definition, global skill is a function of time and represents the average skill over the entire forecast domain. The overall global skill presented earlier represents the skill for a particular month or season type for all forecast years combined. Plots of global skill for individual years, as obtained by the best analog model for each season are shown in Figs. 24–27. It can be seen here that global skill is highly variable in nature, as evidenced by the inconsistency of observed scores. For example, scores obtained for the winter season (Fig. 24) range from a low of 10% correct in 1960 to a high of 72% correct in 1977. Although skill scores for the model types shown are greater than random chance over 60% of the time, the high variability of skill as demonstrated by these particular models makes forecasts based entirely on the analog method somewhat tenuous.

5. Conclusions and discussion

A study has been made to assess the level of predictive skill associated with the application of a simple analog methodology to long-range temperature prediction over the continental United States. This

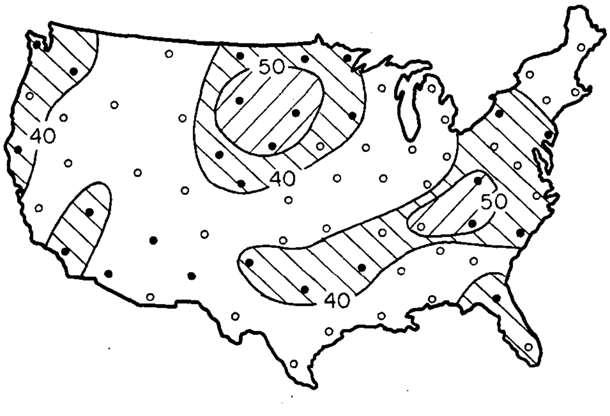

FIG. 23. Distribution of local skill obtained for autumn by analog model using summer N.H. 1000–700 mb thickness (Legend as in Fig. 6.)

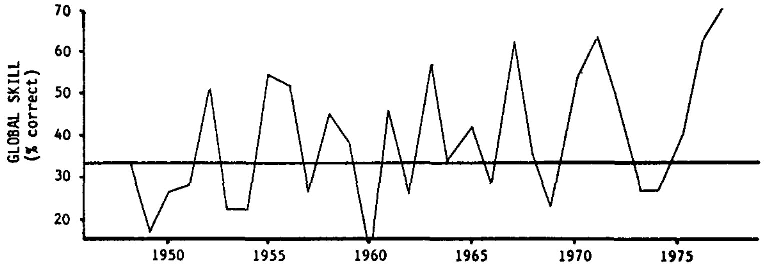

FIG. 24. Annual global skill obtained for winter by analog model using autumn Pacific SST.

simple approach relies solely on the pattern correlation statistic to select the analogs from which temperature forecasts are subsequently derived. Forecasts for approximately thirty cases were made for each trial. Numerous analog trials were attempted employing various combinations of predictor type and domain, forecast period and forecast lead time to generate both monthly and seasonal forecasts. Predictors used in analog models included Pacific SST, 700 mb heights and 1000–700 mb thickness. Domains over North America and the Northern Hemisphere were used for the atmospheric predictors. Several verification statistics have been presented in order to determine the level of skill relative to control forecasts, such as persistence and random selection. Statistical significance tests, using the binomial distribution, were made for the overall global scores and the local scores.

A comparison of the results obtained for the various analog models tested yields the following main conclusions:

1) Overall global skill scores (all years and CD's combined) for both monthly and seasonal analog models were, for the most part, only slightly better than random chance and only occasionally better than persistence. For monthly forecast models, statistical significance was achieved mainly in months of the period January–June, while for seasonal forecasts, significance was achieved for each season by one or more of the analog models. It should be noted that overall global skill is not always an accurate gauge of "usefulness" since these scores represent an average skill over the entire forecast domain.

2) Based on counts of significant local skill, monthly analog models generally performed best during the period of January–June, outscoring persistence and random chance in many instances. For the winter period, analog models using 700 mb heights were the most successful. For the remainder of the year, model performance based on 700 mb heights and 1000–700 mb thickness was similar and generally better than SST models, with no apparent preference to domain. June was the best overall month, with the analog method beating persistence for each model tested, while May and October were generally the least successful months. Forecast lead times of one month were superior to those of two months for analog models employing atmospheric predictors, while SST models were most successful using a two month lag.

3) On a seasonal basis, several analog models performed significantly better than chance and persistence. Winter forecasts were most successful when analogs were based on Pacific SST, while similar results were obtained for summer using SST and N.A. 1000–700 mb thickness. Several analog models also performed well for the spring, relative to persistence, particularly those using 700 mb heights. For autumn, N.A. thickness did well. Thickness models using a forecast lag of one season appeared to be the best overall, beating persistence and random chance for each season, although no clear advantage exists for using either predictor domain. Furthermore, it

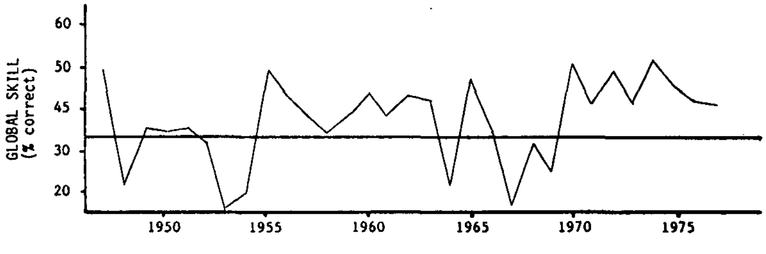

FIG. 25. Annual global skill obtained for summer by analog model using May Pacific SST.

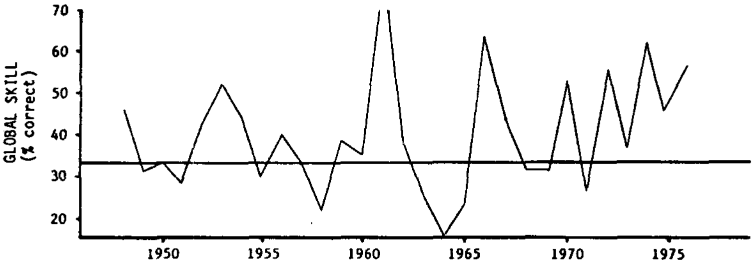

FIG. 26. Annual global skill obtained for autumn by analog model using summer N.H. 1000–700 mb thickness.

is difficult to determine the optimum forecast lead time here, since relative model performance using lags of one month and one season varied with predictor type and forecast period.

4) Although overall global skill scores presented here were generally unimpressive, large coherent areas of high local skill were produced by both monthly and seasonal analog models. However, skill in this sense was found to be highly regional in nature, and its spatial distribution varied considerably with model type and forecast period. Furthermore, the distribution of skill was only partially consistent with predictability studies found in the literature.

Although the work presented here is just a first step, it appears that this simple analog approach holds promise for contributing to the long-range prediction of temperature, if used judiciously, and can effectively compete against a more complex methodology, like that of Barnett and Preisendorfer (1978).

An underlying problem, apparently common to all analog schemes, concerns the dependence of the method upon the extent of available historical data. Since basic analog philosophy utilizes the concept that history does repeat itself, it is likely that the predictive abilities of the method are restricted by limited amounts of data. Better analogs may therefore be found when more data become available, possibly resulting in greater forecasting skill. This idea is also supported by Barnett and Preisendorfer (1978). Considering the method of analog selection presented here, it is possible that at least some correlations between the target and analog years would be higher using longer records of predictor data. In this study, forecasts were made regardless of the degree of matching. However, an examination of the relationship between the pattern correlation values used to select individual analogs and their corresponding global skill scores does not indicate that they are strongly related. This suggests that the precise degree of matching may not be the only aspect to consider when selecting potential analogs. More work in this area is necessary in order to detemine additional factors which may be important for analog selection and to develop optimum analog selection criteria.

In the meantime, the analog method described here is, by design, simple in nature, relatively inexpensive to apply and capable of quickly generating thoroughly objective forecasts. Furthermore, the frequent occurrence of relatively high local skill (i.e., >40%) is encouraging and suggests the potential for more general usefulness if analog selection techniques can be improved. Data sample and predictability problems will, of course, limit skill.

Acknowledgments. This work was performed as part of NJAES Project No. 13504 supported by the New Jersey Agricultural Experiment Station, and the National Oceanographic and Atmospheric Administration, Grant No. NA80AA-D-00004. The authors wish to thank Mr. John Bruno, Mr. Bruce Wyman, Ms. Susan Rinderer, Ms. Cathy Biddulph and Ms. Jeremi Harnack for their assistance in the preparation of tables and figures. The assistance of

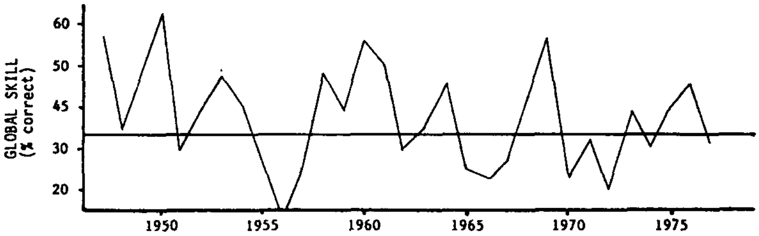

FIG. 27. Annual global skill obtained for spring by analog model using February N.A. 700 mb heights.

Mr. Mark Crane and Mr. John Lanzante in computer-related areas is also appreciated.

REFERENCES

Barnett, T. P., 1981: Statistical prediction of North American air temperatures from Pacific predictors. *Mon. Wea. Rev.*, **109,** 1021–1041.

——, and R. W. Preisendorfer, 1978: Multifield analog prediction of short-term climate fluctuations using a climate state vector. *J. Atmos. Sci.*, **35,** 1771–1787.

Bowen, D., 1976: Long-range weather forecasting. *Water Power Dam Construct.*, July, 31–35.

Diaz, H. F., 1981: Eigenvector analysis of seasonal temperature, precipitation, and synoptic-scale system frequency over the contiguous United States. Part II: Spring, summer, fall, and annual. *Mon. Wea. Rev.*, **109,** 1285–1304.

——, and D. C. Fulbright, 1981: Eigenvector analysis of seasonal temperature, precipitation, and synoptic-scale system frequency over the contiguous United States. Part I: Winter. *Mon. Wea. Rev.*, **109,** 1267–1284.

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

Lund, I. A., 1963: Map-pattern classification by statistical methods. *J. Appl. Meteor.*, **2,** 56–65.

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

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

——, 1978: Long-range weather and climate predictions. *Studies in Geophysics: Geophysical Predictions.* National Academy of Sciences, 103–114.

Radinovic, D., 1975: An analog method for weather forecasting using the 500–1000 mb relative topography. *Mon. Wea. Rev.*, **103,** 639–649.