Origins and Levels of Monthly and Seasonal Forecast Skill for United States Surface Air Temperatures Determined by Canonical Correlation Analysis

T. P. BARNETT

Climate Research Group, Scripps Institution of Oceanography, La Jolla, CA 92093

R. PREISENDORFER

Pacific Marine Environment Laboratory, National Oceanographic and Atmospheric Administration, Seattle, WA 98115

(Manuscript received 2 July 1986, in final form 4 February 1987)

ABSTRACT

Statistical techniques have been used to study the ability of SLP, SST and a form of persistence to forecast cold/warm season air temperatures over the United States and to determine the space–time evolution of these fields that give rise to forecast skill.

It was found that virtually all forecast skill was due to three climatological features: a decadal scale change in Northern Hemisphere temperature, ENSO-related phenomena, and the occurrence of two distinct short-lived, but large-scale, coherent structures in the atmospheric field of the Northern Hemisphere. The physical mechanisms responsible for the first two signals are currently unknown. One of the large-scale, coherent features seems largely independent of the ENSO phenomena, while the second is at least partially related to ENSO and may be part of a recently discovered global mode of SLP variation. Both features resemble various combinations of known teleconnection patterns. These large-scale coherent structures are essentially stationary patterns of SLP variation that grow in place over two to three months. The structures decay more rapidly, typically in 1 month, leading to a highly asymmetric temporal life cycle.

The average forecast skills found in this study are generally low, except in January and February, and are always much lower than expected from studies of potential predictability. Increase in the average skills will require new information uncorrelated with any of the data used in this study and/or prediction schemes that are highly nonlinear. However, the concept of an average skill may be misleading. A forecast quality index is developed and it is shown that one can say in advance that some years will be highly predictable and others not. Use of the classical definition of "winter" in forecast work may not be advisable since each of the months that make up winter are largely uncorrelated and predicted by different atmospheric features.

1. Introduction

This paper uses advanced statistical techniques to address two questions concerning short-term climate prediction. Given information on the past space–time history of climate fluctuations, how much of the future variation (predictand data) can be skillfully forecast? More interesting and important is the question, What is the space–time evolution of the climate system that leads to high forecast skill? The answers to these questions, which are the goals of this paper, may provide clues to the physical processes that are responsible for short-term predictability.

The most interesting question addressed by this paper, namely, What key features of the space–time evolution of the climate system lead to a successful forecast? has not been well studied. Statistical forecast studies have offered qualitative "scenarios" based on the statistical models to explain the results (e.g., Barnett, 1981a, 1977, 1981b). Such studies have provided only a partial description of how the climate system changed to give the observed forecast. Other statistical studies often just ignore the question of why skill was obtained. General circulation model (GCM) studies have, to date, been only modestly helpful in delineating the physical processes responsible for predictive skill. They have simulated observed climate features in what might loosely be called a "specification" mode. For example, given a particular distribution of SST, compute the associated distribution of climate anomalies in the atmosphere (e.g., Rowntree, 1972; Blackmon et al., 1983; Shukla and Wallace, 1983; and many others). The time evolution of these anomalies and their relation to subsequent forecast skill is only now beginning to be studied (e.g., Chervin, 1986; Lau, 1985). Finally, qualitative forecast techniques, e.g., physical-synoptic methods, offer descriptive scenarios to explain forecast success (cf. Namias, 1975) but lack the quantitative rigor needed to be definitive.

The second question, regarding levels of possible predictive skill, has traditionally been approached in two ways. One tack has been to estimate the amount of climate variability (variance) associated with non-random fluctuations of some particular variable, e.g.,

surface air temperature. This non-noise variance is often called "potential predictability" and is obtained by assuming some type of model to describe the observed time series (e.g., Madden, 1976; Madden and Shea, 1978; Trenberth, 1984a, 1984b). This method generally leads to larger estimates of possible predictive skill than are actually observed in climate forecast experiments (see below). This may be due, at least partially, to the substantial assumptions attending this approach. For example, the "non-noise" variance is assumed to be completely predictable. Also, there may have been neglect, in some studies, of the effect of artificial skill associated with model fitting. Finally, the forecast aspect of the problem is treated only implicitly.

A second approach to estimating expected forecast skill levels is to select a group of predictors a priori and see how well they can be used to predict future climate change. The actual selection of predictors is crucial to this approach. Barnett (1981a) used hypothesis testing and a physically reasoned approach to select predictor variables for estimating the predictive skill of surface air temperature over the United States and Europe using a number of regional indices of sea surface temperatures (SST). None of these experiments used all of the available predictor information and so may have underestimated forecast skill. Harnack (1979, 1982), Harnack and Lansberg (1978) and Harnack and Lanzante (1984) used full fields of predictors, e.g., SST, to estimate the predictive skill. However, they generally offered the prediction model the choice of a large set of predictor information from which to choose and form a hindcast model. This "predictive" approach carries some penalty by overestimating hindcast skill at the expense of real forecast skill, the levels of overestimation depending critically on statistical methodology used. There is an additional problem associated with the above "prediction" approaches in that they are rarely tested in an extensive set of truly independent forecast experiments (for an exception, see Dixon and Harnack, 1986). Nevertheless, the skill levels found by this approach are generally considerably less than found by the time series modeling approach, as noted above.

Subsequent sections of this paper discuss first the data and then a statistical method for obtaining maximum predictive skill from a set of predictor data. Advanced methods of significance testing and model interpretation are also included. The mathematical details of these methods are given in the Appendix. The general levels and distributions of skill in forecasting surface air temperature over the United States on a monthly and seasonal basis are discussed next. Final sections provide insightful views of the changes in the Northern Hemisphere oceans and global atmosphere that lead to the estimated forecast skill.

2. Data

The data field to be predicted is the surface air temperature over the United States. Several versions of this field have been used in the numerical experiments described below. The simplest approach was to represent the field by 33 widely separated stations (Fig. 1) of monthly data for the period 1931–80. These data come from the Monthly Climatic Data for the World and were obtained from R. Jenne at NCAR. Another possible representation of the temperature field is to form district averages for the areas immediately surrounding the stations shown in Fig. 1. These data are described by Diaz and Quayle (1978) and have the advantage of minimizing effects of urban growth, station movement, etc. (cf. Douglas et al., 1982; Cayan and Douglas, 1984). Forecasts were done with both datasets and yielded essentially identical results. Station data for the period 1900–29 were also used in an independent test of model validity.

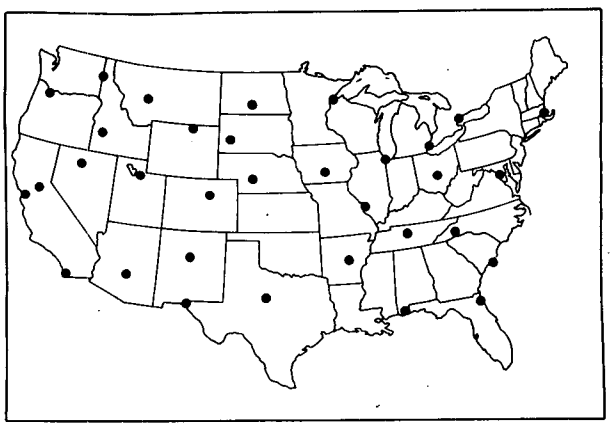

FIG. 1. Locations of stations/districts providing surface air temperature predictand data.

Four predictor data fields were used. The first was the sea level pressure (SLP) field for the region 140°E to the Greenwich meridian and 20°–70°N. The data were on a 5° latitude by 10° longitude diamond grid. There were a total of 280 grid points of data in this region. The gridded data were obtained from NCAR and had been corrected insofar as possible for the problems noted by Trenberth and Paolino (1980). Based on studies by the latter authors, the monthly SLP data for the period 1931–80 were used in the subsequent analysis. However, data for the period 1900–29 were used for an independent test of model validity. A more planetary view of the results obtained from the limited region SLP experiments was obtained using the near-global SLP field described by Barnett (1985). These data cover the region from 42.5°S to 72.5°N and the time span 1950–85 at monthly intervals on a 5° × 10° grid. Regionally averaged SST data (Fig. 2) for the North Pacific and the North Atlantic were also used as a predictor set. The data were obtained by processing individual ship observations in the Marine Deck. Monthly values for the period 1931–80 were available for all 21 regions (see Barnett 1981, 1984, for additional details). Finally, the air temperature data

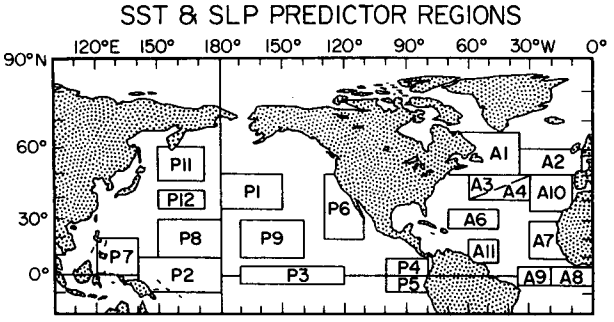

FIG. 2. SST from the large averaging areas shown above were used as predictor information. SLP predictor data came from the region 20°–70°N, 140°E to the Greenwich Meridian.

were used to predict themselves, i.e., the predictand data were represented as an autoregressive process.

3. Methods

The general approaches to model building, significance testing and interpretation are summarized in this section. Since most of the techniques are not well known in oceanography/meteorology, the present discussion will be entirely qualitative, attempting to give the reader a feel for the methods in the context of more familiar techniques. A mathematical description of the methodology is given in the Appendix to assist those interested in rigor and/or a more precise description of the approaches.

a. Model building

Canonical correlation analysis (CCA) is at the top of the hierarchy of regression modeling approaches. The first description of the method appeared in Hotelling (1936), and more recent descriptions appeared in Anderson (1984), among others. The method has been used in meteorology only sparingly (e.g., Glahn, 1963; Davis, 1977; Barnett, 1981a; Nicholls, 1987). The approach may be understood as follows. The simplest approach to statistical model building is to regress one variable upon another. Multiple regression, the next higher step, attempts to relate a vector of predictor data to a single predictand variable. Stepwise multiple regression attempts to select from a large set the most predictively important variables to explain a single predictand. Canonical analysis is the generalization of all these approaches. It finds the optimum linear combination of the predictor data vector that will explain the most variance in the predictand data vector. Both predictor and predictands are now full multidimensional vectors of information.

CCA may also be understood by analogy to standard empirical orthogonal function (EOF) analysis. The EOF approach defines a new orthogonal coordinate system that optimally describes variance in a single dataset. This coordinate system is based on the eigenvalue/eigenvectors of a covariance matrix computed from the single dataset. CCA defines coordinate systems that optimally describe the cross covariance between two different datasets [ref. Eqs. (A4)–(A8)]. Again, this is expressed as an eigenvalue problem. Now, however, eigenstructure is obtained from the product of the cross covariance matrix between two datasets and its transpose [ref. (A7)]. Since this product matrix describes the hindcast skill (regression coefficients squared), its eigenstructure maximizes this skill. Again, by analogy, the eigenvalues in an EOF analysis represent the relative variance associated with each dimension in the new coordinate system. In CCA, the resulting eigenvalues, denoted by μ_k, are called canonical correlation coefficients and represent the levels of correlation between patterns of predictor variables and patterns of predictand variables (A11). The sum over all k of the μ_k^2 is the hindcast skill of the model. In standard EOF analysis, the same sum is just the total variance of the dataset.

The strengths of CCA are its ability to operate on full fields of information and to objectively define the most highly related patterns of predictors (Y) and predictands (T). By including both space and time lag information in the predictor field [e.g., (A1)], it is thus possible to define both the space and time evolution of the predictor data that best predicts an associated pattern of T-variability. These abilities of CCA allow us to fulfill one of the goals of this paper.

There are several potential drawbacks to CCA. In highly intercorrelated data fields the estimation of the inverse matrices needed in CCA may be impossible since the matrices may be degenerate. However, one can overcome this problem by first orthogonalizing the Y and T data [Eq. (A2)] and then using the orthogonal variates as input to the CCA analysis. This step also allows one to prefilter the data to eliminate noise and invoke the principal of parsimony so vital to statistical modeling.

The limitation of the number of predictors is vital to CCA. Since the method is largely a posteriori, it will always find the best possible relation between Y and T. Given enough predictors, CCA, like any other regression scheme, will build a model capable of accounting for large amounts of variance in T, but this apparent skill would be largely artificial. Unfortunately, limiting the number of predictors/predictands in the analysis can exclude information that is potentially useful. An example of this problem will appear later in the text where our objective prefiltering eliminated an apparently predictable signal. All of this means that testing the significance of CCA models is not particularly amenable to standard methodology, e.g., a test against the "null hypothesis."

b. Significance testing

The procedures of section 3a place a heavy burden on a significance test since the analysis has "stacked

the deck" to give the best possible relation between Y and T, i.e., there is nothing a priori about the analysis (Lawley, 1959). The procedure we have chosen to test the skill of the models is often called "cross validation" (Stone, 1974, 1977; Efron, 1982; Tukey, 1958). This idea is an old one and goes as follows: Withdraw from Y the predictor data associated with a discrete time t_ν, $1 \leqslant \nu \leqslant n + 1$, and also withdraw the contemporary predictand datum from T, and denote them by $Y(x, t_\nu)$ and $T(x', t_\nu)$. The remaining (Y, T) data now have n time values and can be relabeled (Y', T'). The (Y', T') datasets, which are often called the training sets, are used to construct a model, of the form (A14), for estimating $T(x', t_\nu)$.

An independent test of the model, subject to the conditions noted below, is obtained by using $Y(x, t_\nu)$ to predict $T(x', t_\nu)$ via (A14). Denote this estimate of T by $\hat{T}_\nu$. Note that the $(Y, T)_\nu$ values used to obtain $\hat{T}_\nu$ must in no way enter the model building process. These values are often called the testing set. One can proceed through the available data with $t_\nu = 1, 2, \cdots, n + 1$, thus obtaining $n + 1$ independent predictions from the model (which itself will change somewhat with every realization). The result of this repetitive action will be a time series of predicted fields $\hat{T}_\nu$ and a set of corresponding observations, denoted by T_ν. Note the apparent similarity of this approach to "jackknifing," where the same "leave one out" strategy also is used. However, this latter approach generally deals with estimating the significance of some statistical moment of $\hat{T}_\nu$. By contrast, the "bootstrap" method (e.g., Inoue and O'Brien, 1984) approaches the resampling/significance problem in the same philosophical way except it differs by replacing values selected for its testing set, i.e., a test set is constructed in which a particular pair of variates may appear several or more times. These and other more subtle differences in resampling significance test methodology are discussed by Efron (1983).

Cross validation is meaningful only if the predictors are serially uncorrelated. Thus, in the current context, the T should not be correlated from one year to the next. These year-lag correlations were computed for all of the individual months/stations and season/stations used in this study. The vast majority (88%) of the data exhibited a nonsignificant correlation value according to standard testing procedures, a result in agreement with that of Madden (1976).

The major exception to the above statement occurred during the summer season and summer months, where significant one year lag correlations were found, a result found previously by Namias (1978), van den Dool et al. (1986) and others. However, at lags of two and three years the number of significant correlations dropped to the number expected by chance. The cross validation was modified for the summer experiments to accommodate these facts as follows: Again, a sample denoted by $(Y, T)_\nu$ was withdrawn from the data for eventual testing. The data values on either side of t_ν, i.e., $(Y, T)_{\nu-1}$ and $(Y, T)_{\nu+1}$, were also withdrawn but discarded completely from both the model building and testing procedures. Hence, the test pair $(Y, T)_\nu$ was separated by two years from the training set and so was largely independent of the model building/training set. Under this circumstance, the cross-validation procedure should be reliable.

The predicted and observed data fields were next decomposed into terciles, T_ν being used to estimate its tercile limits and $\hat{T}_\nu$ being used to estimate its own limits (see Appendix). Care was taken to insure that $n + 1$ was an even multiple of 3. The percent of correct forecasts at a particular location (x'), e.g., "Above" predicted and "Above" observed, is called the *local* skill, S_L [e.g., (A15)]. The significance of S_L may be determined in the normal manner from a binomial distribution. The percent of stations across the United States that showed significant local skill was used to estimate the *global* skill [$\hat{S}_G$; Eq. (A16) was also estimated]. The significance of $\hat{S}_G$ again was obtained from the binomial distribution after accounting for the correlation between adjacent stations (see Appendix, and Livezey and Chen, 1983).

The significance testing procedure described above is lenient, particularly in allowing T_ν and $\hat{T}_\nu$ to define their individual tercile limits. However, we are trying to test the *forecast* skills of the statistical models, and these were not anticipated to be high. In such a situation a measure of model skill based on, say, variance accounted for would be overly harsh since one large "miss" would doom the model. Similarly, if one terciled $\hat{T}_\nu$ by the tercile limits of T_ν, then models that accounted for small amounts of variance would typically tend to forecast only "normal" conditions. The strategy used here is really designed to measure the accuracy of the phase of the forecasts. The philosophy, then, is similar to GCM climate studies where the model's climatology is used to define the model's anomaly field since use of the observed climatology for this purpose could obscure[1] any ability the model had to simulate variations in climate.

c. *Model interpretation*

One of the prime reasons for using CCA was that it provided two optimally defined diagnostic fields of information that show where forecast skill occurs in the predictand data and the space–time evolution of the predictor field that gives rise to that skill. The information on time evolution is possible due to the concatenation of consecutive months of spatial fields [i.e., (A1)]. These canonical patterns (A12b) are obtained by inversion of (A12a). The equations (A12a) will ap-

[1] This is so because the model climatology and observed climatology are generally rather different (e.g., Barnett, 1986).

pear superficially familiar to EOF users in that the canonical patterns appear to perform the same role in CCA as eigenvectors do in EOF analysis. However, while the canonical components (A4) are orthogonal, the corresponding canonical patterns need not be.

The *canonical predictor patterns* (A12b) are denoted by $g_j(x)$, where x is a dummy variable representing both space and time [ref. (A1)]. We refer to these patterns as g-maps. The g_1 map describes the linear combination of the predictor data that will contribute the largest fraction of hindcast skill in the predictand (T) set, g_2 the next largest skill and so on. The patterns of hindcast skill that accompany the g-patterns are called *canonical predictand patterns* and are denoted by h_k [ref. (A12b)]. Again, the pattern associated with the largest fraction of predictability in the T-field is given by h_1 and so on.

Interpretation of the g- and h-maps is facilitated if they are normalized to unit vectors. This allows ready evaluation of the relative contribution to the associated g-map of each month in a sequence of concatenated spatial fields. The details may be found in the Appendix.

4. Forecast skill

This section describes the results of a series of experiments to estimate the skill associated with prediction of monthly and seasonal surface air temperatures over the United States. The analysis considered only the summer and winter seasons and also the individual months that make up these seasons, i.e., June, July and August and December, January and February, respectively. Also, the results discussed below were made for a forecast lead of one time unit, i.e., one season or one month. Considering that the cross-validation approach to scoring was used, the results should be good approximation to actual forecast skills.

It is important to note here and following that the use of the word "persistence" refers to attempts to forecast using the three prior monthly values of air temperature instead of only the value immediately preceding the forecast time. This nonstandard usage thus refers to the construction of a low order autoregressive model for forecasting future temperatures and obviously includes, as a special case, the more common, simple persistence model.

a. Seasonal forecast skill

Forecasts of winter and summer air temperature were made from (i) persistence, (ii) SST, (iii) SLP and (iv) all three fields combined. The results were as follows:

Persistence. The temperature data for the three individual months preceding the season of interest were selected a priori as predictors. Thus, for example, T-field data for March, April and May were used to predict "summer." The tercile skill scores for this experiment are given in Table 1, while the strongest result is shown in Fig. 3a. Persistence (as defined above) clearly does a moderately good job of predicting summer temperature but not winter values. Numerous prior studies have obtained similar results based on hindcast studies. The current results, based on forecast skills, show that scores in the low 40s are common over nearly 75% of the country during summer (Fig. 3a). Over the lower half of the country the local scores are significant at the 5% level. Using (A17) we found typical values of $q_e = 6$ for the summer, i.e., there were only 6 spatial degrees of freedom in the full 33 station set. However, even after accounting for this fact, the global skill (S_G) was still significant at over 99.9%. These impressive results must be tempered by the fact that the actual local scores themselves are only 10–20 percentage points above those expected by a stochastic forecaster (33%). Note also that the pattern of predictive skill does not correspond well with the main patterns that account for most of the air temperature variance in summer (e.g., Diaz, 1981, Fig. 4), although there is no reason a priori to expect it should.

SLP predictors. The SLP data for the three individual months preceding the season of interest were used as predictors. The lag of three months was selected a priori in order to exceed typical decorrelation times for midlatitude SLP variations. The analysis, via the g-maps, will show if this was an appropriate lag to consider. The tercile skill scores for the experiment are shown in Table 2. The results are unencouraging to say the least. None of the winter scores exceed 38%, a value significant at only the 22% level. The results for summer are only slightly better, with three scores exceeding the 10% significance level; but just that many would have been expected by chance, given that forecasts were made for 33 stations. With typical summer values of $q_e \approx 6$, it is clear that the $\hat{S}_G$ associated with SLP predictor data is not statistically meaningful. These results are in general agreement with those found earlier (e.g., Barnett, 1981a).

SST predictors. In this set of experiments, the SST for the three individual months prior to the predictand season and two seasons prior to predictand season were used as predictors. For example, SST data from summer, September, October and November were used to forecast winter. Again, these lag intervals were selected on an a priori basis largely from the decorrelation times of SST.

The most successful forecasts were for summer (cf. Table 3), with the spatial distribution of skill (Fig. 3b) being different from that expected by chance at the 95% level. The winter forecast skill (not shown) was confined to the southeastern corner of the country and to the northern plains. This was essentially the same result found by Barnett (1981a), although the areal distribution of skill is somewhat smaller than found in

TABLE 1. Skill scores* given as a percentage of correct tercile forecasts, obtained using a measure of persistence to predict station temperatures in the month/season shown. A value of 33% is expected by chance.

Station	Summer	Jun	Jul	Aug	Winter	Dec	Jan	Feb
Jacksonville	37	39	33	41	22	41	27	22
Charleston	35	31	41	33	37	43	25	16
Mobile	43	35	43	45	37	41	37	16
Abilene	45	33	45	37	35	47	35	22
El Paso	41	31	33	41	31	35	33	39
Phoenix	41	47	52	39	22	27	31	33
San Diego	50	39	31	43	41	35	37	25
Asheville	45	39	43	35	41	45	25	31
Nashville	43	33	52	37	33	52	31	22
Little Rock	60	37	43	39	37	41	37	18
Albuquerque	45	37	37	41	22	37	33	43
Washington DC	43	25	39	35	35	50	25	27
Columbus	54	33	41	41	31	47	16	25
St. Louis	39	33	37	25	35	41	33	27
Denver	39	52	29	35	14	10	20	16
Sacramento	37	56	37	33	39	37	37	41
San Francisco	41	41	31	25	45	35	43	31
Blue Hill	37	35	20	18	31	31	20	37
Chicago	20	35	33	33	35	43	22	20
Detroit	20	39	43	37	35	43	16	34
Des Moines	37	37	37	16	22	33	20	29
North Platte	29	47	29	29	12	33	18	27
Salt Lake City	58	50	61	65	69	29	43	27
Winnemucca	54	47	33	43	31	35	43	31
Toronto	31	41	37	25	33	31	22	29
Rapid City	43	39	31	39	31	31	33	50
Sheridan	41	45	29	35	33	39	29	31
Boise	41	35	37	43	39	29	54	22
Portland	31	18	39	31	33	31	45	29
Duluth	37	20	31	27	27	29	39	29
Bismarck	41	37	25	47	25	41	20	29
Helena	43	33	25	35	37	31	45	39
Spokane	48	31	52	37	48	41	45	37
Average	41.2	37.5	37.3	35.9	33.3	37.5	31.5	28.9

* Values greater than 41, 44 and 48 are significant at the 0.10, 0.05 and 0.01 levels, respectively.

that study; a result likely due to the different methodology used here.

The surprising result of this experiment was the unexpected skill found for summer forecasts. This result was robust to changes in the data omission window of the cross-validation testing. Note the magnitude and spatial distribution of skill are quite similar to those found in the persistence experiment. Similar results have not been found by others (e.g., Barnett, 1981a; Harnack, 1979, 1982). The techniques used here and length of datasets are considerably different from those used previously, and this must partially explain the difference. However, the skill seems to originate from a genuine geophysical signal that earlier studies did not detect. A more complete discussion is deferred to section 5.

Mixed predictors. This experiment combined the predictors from the previous three experiments in an attempt to see if forecast skill was due to a single climatic signal common to all predictor fields or to signals that were unique to a specific field or fields. If the former case were true, then the mixed predictor experiment would score no better than an experiment using only one field as a predictor. If the latter situation were true, then the mixed predictor experiments would have appreciably higher skill than any single predictor experiment.

The mixed experiment was set up as follows: In each single predictor experiment the input data field was decomposed according to (A2) and p principal components used as predictors. The principal components so retained from each of the prior experiments (persistence, SLP and SST) were then considered equivalent to a raw predictor field (Y) and the procedures of the Appendix invoked. Thus, the mixed predictor Y-field was composed of 5 SLP, 5 T and 6 SST principal components so the dummy index $x = 1, 2, \cdots, 16$ [cf. (A1)]. These 16 series embody all the lag information associated with each individual input field (see preceding subsection for definitions).

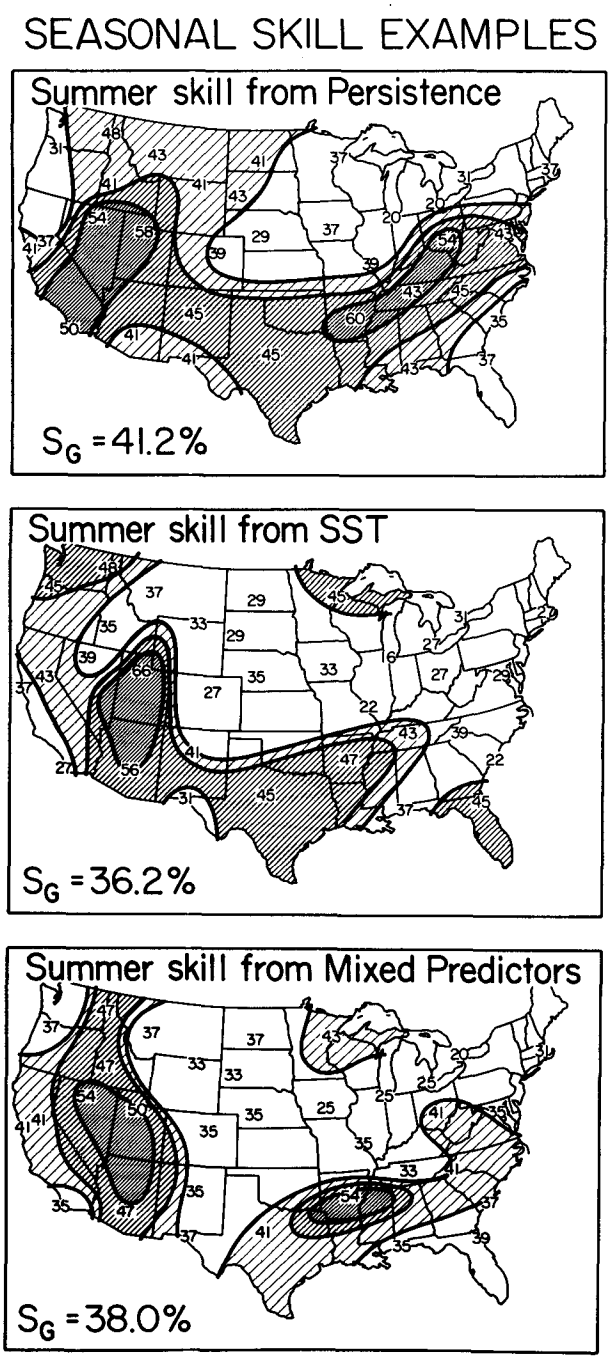

FIG. 3. The upper panel shows the summer surface air temperature forecast skill obtained from persistence. The small numbers refer to the percent of correct forecasts. The hatching indicates statistical significance at the 10%, 5% and 1% levels, the significance increasing with the density of the hatching. The contour intervals of local skill associated with these levels are 41, 44 and 48, respectively. The center and lower panels show the forecast skills from SST and the mixed predictor experiments, respectively. The average value of local skill in percent for each experiment (S_G) is shown in the lower left-hand corner of each panel.

The results of the mixed predictor experiments are given in Table 4 and the best result shown in Fig. 3c. Winter scores are negligible. Apparently, the lack of correlation between the prior air temperature and SLP and subsequent winter T-values overwhelm the modest correlation associated with the prior SST data. This latter information was apparently diffused in the decomposition (A2) while the former predictor data were included in the model fitting procedure to the detriment of the forecast skill. Attempts to prefilter the predictors (i.e., select $p < 16$) did not appreciably affect the results.

The distribution of summer forecast skill (Fig. 3c) is similar to that obtained in the persistence and SST experiments (Figs. 3a, b). The main difference is the loss of some skill in the Mississippi Valley and the Northern Plains. All in all, the mixed predictor experiment procedure produces essentially the same results described above. Further, the persistence and SST predictors produced nearly the same spatial patterns of skill. One may conclude, therefore, that the forecast skill is due to a signal whose temporal properties are common to both the SST field of the Northern Hemisphere and the T-field over the United States. The nature of that signal will be discussed in section 5.

b. Monthly forecast skill

Forecasts for the individual months that make up summer and winter were made with the same set of predictors noted in section 4a. In this case, the SLP or air temperature data for the three months immediately preceding the month of interest were used in the model, e.g., November, December and January data were used to forecast February. The SST forecast experiments used both the preceding three months (as above) and the seasonally averaged data for the season prior to those months, e.g., summer, September, October and November SST data to predict December air temperature.

1) WINTER MONTHS

Persistence. The results (Table 1) show that December has the highest forecast scores while February has the lowest scores. In the former month, the skills are highly significant, but again low in magnitude over the eastern third of the country (Fig. 4). However, the failure of the forecast model in January and February means that, on average, the three months that make up "winter" are largely uncorrelated with each other. This fact will be revisited in section 5b.

While most of the results shown in Fig. 4 are what would be expected from earlier studies of persistence (e.g., Namias, 1978; Barnett, 1981a; van den Dool et al., 1986), there is one glaring difference. The current results suggest that the stations immediately along the West Coast, particularly San Diego and San Francisco,

TABLE 2. Skill scores* given as a percentage of correct tercile forecasts, obtained using Northern Hemisphere SLP to predict station temperatures in the month/season shown. A value of 33% is expected by chance.

Station	Summer	Jun	Jul	Aug	Winter	Dec	Jan	Feb
Jacksonville	36	40	58	38	30	42	48	51
Charleston	38	28	34	36	34	51	51	46
Mobile	28	36	38	44	28	48	57	51
Abilene	40	36	30	38	38	40	36	44
El Paso	34	36	34	20	34	30	14	12
Phoenix	26	34	34	28	30	44	42	36
San Diego	48	26	40	32	36	30	55	38
Asheville	34	44	48	30	28	38	59	44
Nashville	32	30	36	40	32	48	57	51
Little Rock	38	24	34	26	30	34	53	48
Albuquerque	34	44	34	24	36	26	26	44
Washington DC	30	30	42	36	24	36	59	30
Columbus	30	28	34	30	26	36	53	42
St. Louis	20	28	30	38	26	44	44	51
Denver	20	30	26	26	22	20	34	36
Sacramento	16	30	36	32	30	40	44	46
San Francisco	30	16	34	30	32	40	55	44
Blue Hill	34	46	30	40	34	42	59	24
Chicago	24	40	34	20	34	46	51	40
Detroit	44	40	26	36	34	47	57	36
Des Moines	40	38	34	36	24	51	28	24
North Platte	24	42	32	26	38	40	34	36
Salt Lake City	62	28	22	46	24	61	44	68
Winnemucca	26	32	28	38	26	48	57	40
Toronto	32	34	28	32	26	40	61	38
Rapid City	30	42	28	30	28	32	38	44
Sheridan	38	34	26	34	36	42	40	44
Boise	32	30	34	46	38	40	51	34
Portland	26	34	28	32	38	38	44	44
Duluth	48	51	30	40	38	40	38	30
Bismarck	32	44	22	34	30	36	36	40
Helena	40	36	30	36	28	34	40	38
Spokane	30	34	24	40	30	28	46	44

* Values greater than 41, 44 and 48 are significant at the 0.10, 0.05 and 0.01 levels, respectively.

are poorly correlated in time. On close inspection this result did not hold up and was found to be due to a shortcoming of the experimental design in the present study. The problem was alluded to in section 3a and will be discussed in more detail in section 7. The problem did not affect the general pattern of results shown in Fig. 4a.

SLP predictors. The monthly scores for the winter season were the highest found in this study (see Table 2, Fig. 4). Significant local scores (10%) existed at nearly two-thirds of the stations during February. During January one half of the stations had local skill scores whose significance was in excess of 1%. Taking into account the spatial correlation between the stations [Eq. (A17)] showed there were only 4 or 5 degrees of freedom in the 33 station set during the winter months. Nevertheless, the probability of obtaining the above spatial distribution of local skills by chance is less than 0.01%. In all three months, but particularly January, the numerical value of the scores was high enough to be practically useful. Note also that the spatial distribution of forecast skill now closely resembles a natural mode of variance in the temperature field (e.g., Diaz and Fulbright, 1981, Fig. 1).

The general pattern of skill across the country showed maxima in the eastern and western thirds of the country. The minimum in the central portion of the United States is similar to the result previously found by Barnett (1981a), Preisendorfer and Mobley (1984) and others for winter season prediction studies. The exception to this statement occurred in February when a number of significant local scores were apparent in the central portions of the country.

SST predictors. SST data from the 21 regions shown in Fig. 2 were used to predict air temperatures for December, January and February. In each case, the SST for each of the preceding three months and summer season were used as predictors (cf. discussion of winter season predictor set). The results are listed in Table 3, and the best case is illustrated in Fig. 4c.

Both December and February had global skill levels that were significant at the 5% level. During January five stations had a significant local skill score. Essentially, the SST were not of as much help in forecasting

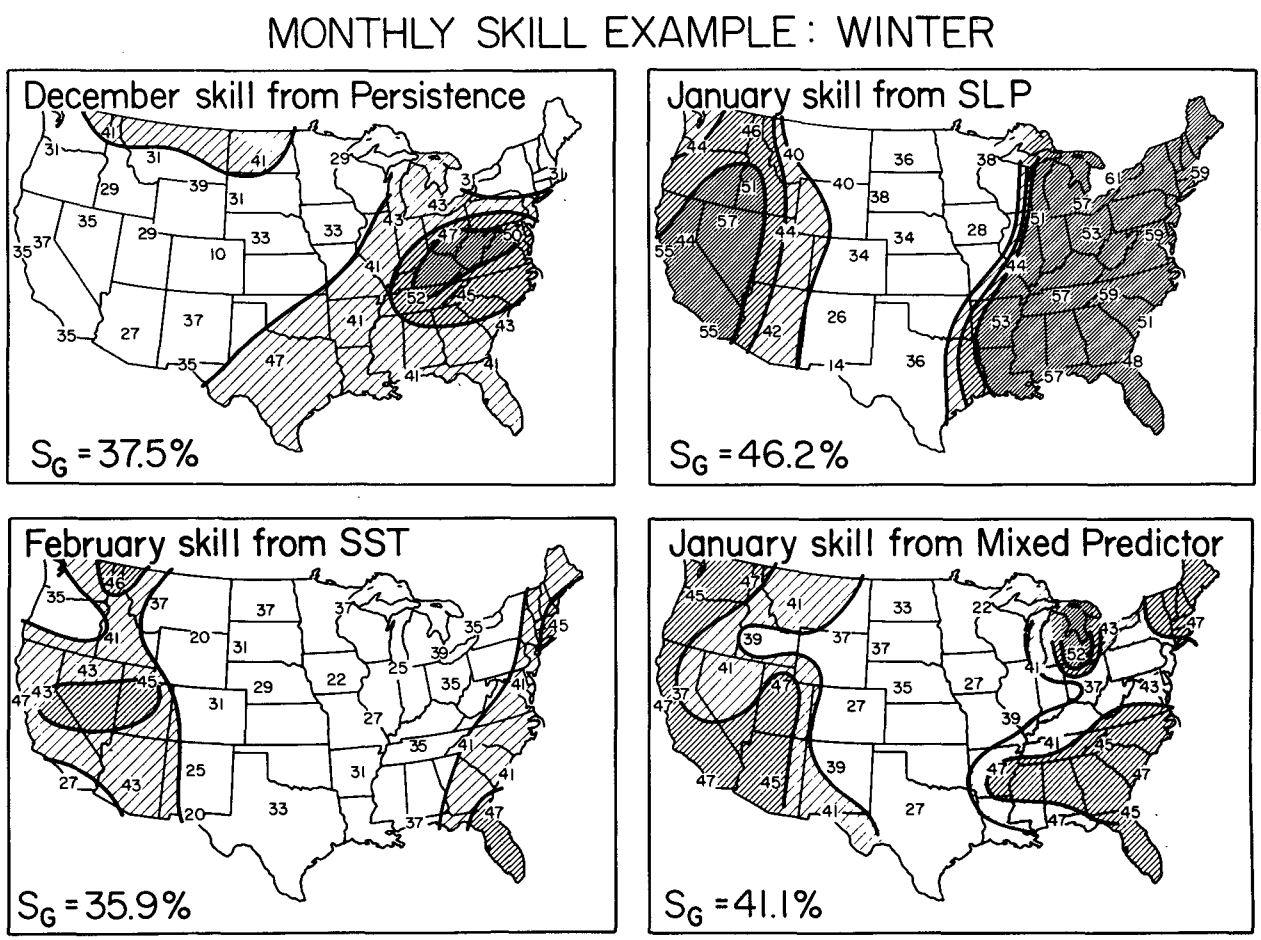

FIG. 4. Examples of surface air temperature forecast skill obtained for specific months during the winter season from different predictors. The conventions are identical to those described for Fig. 3.

air temperature anomalies during this month as during the other cold season months.

The spatial distribution of skill during February (Fig. 4c) is centered along both the eastern and western parts of the country in a pattern reminiscent of the distributions found in the SLP experiment (above) and by earlier workers. The distribution of skill in December was rather different with the majority of skill located in the eastern third of the country, particularly the upper Mississippi Valley. The high skill region located in the western United States during February largely disappears in December, although local scores are typically only a few percentage points below the 10% significance level. A distinct minimum in skill is evident in the central portion of the country in both months.

Mixed predictors. This experiment used the principal components of the individual predictor fields as basic input, e.g., as the *Y*-field. The procedure is as described in the section on seasonal forecast skill. The results are listed in Table 4 and the strongest result shown in Fig. 4d.

The spatial distributions of skill are highly significant in January and February (5%) and marginally significant in December (10%). The strongest result again occurs in January (Fig. 4d) and bears a strong qualitative resemblance to the results obtained from the SLP field alone. Yet, the mixed predictor experiment generally had numerically lower local skill scores than the SLP experiment, particularly when $p = 16$, i.e., no prefiltering of the input predictor set. It is concluded that the additional predictor information (SST, persistence) carried no information that was not already in the SLP field.

2) SUMMER MONTHS

Persistence. The months of June and July have the highest forecast skill scores from persistence. Of the two, the results for June are probably the strongest, particularly in the western states and Midwest (Fig. 5a). Both August and July show significant skill in the western third of the country, thus suggesting significant

TABLE 3. Skill scores* given as a percentage of correct tercile forecasts, obtained using Northern Hemisphere and equatorial SST to predict station temperatures in the month/season shown. A value of 33% is expected by chance.

Station	Summer	Jun	Jul	Aug	Winter	Dec	Jan	Feb
Jacksonville	45	37	41	41	50	31	47	47
Charleston	22	27	37	33	45	33	33	41
Mobile	37	35	29	35	43	35	37	37
Abilene	45	41	39	45	33	31	31	33
El Paso	31	31	41	41	27	45	35	20
Phoenix	56	47	68	56	18	39	31	43
San Diego	27	31	16	27	39	41	27	27
Asheville	39	31	41	35	41	35	39	41
Nashville	43	35	41	43	43	47	39	35
Little Rock	47	37	31	33	35	43	33	31
Albuquerque	41	20	37	45	33	33	29	25
Washington DC	29	31	20	27	37	41	35	41
Columbus	27	35	31	31	33	47	33	35
St. Louis	22	35	31	41	31	41	20	27
Denver	27	37	39	45	30	45	31	31
Sacramento	43	45	43	50	22	37	29	43
San Francisco	37	37	33	43	39	33	35	47
Blue Hill	27	35	35	31	31	39	39	45
Chicago	16	33	14	31	27	47	29	25
Detroit	27	37	25	29	22	43	37	39
Des Moines	33	25	31	37	39	43	27	22
North Platte	35	41	37	37	29	35	31	29
Salt Lake	66	69	68	58	20	45	65	45
Winnemucca	39	41	43	47	25	39	37	43
Toronto	31	39	20	27	25	43	35	35
Rapid City	29	35	33	31	43	31	31	31
Sheridan	33	45	37	31	39	27	31	20
Boise	35	35	29	43	25	35	33	41
Portland	45	41	45	41	29	39	22	35
Duluth	45	33	39	37	39	35	27	37
Bismarck	29	39	31	29	43	25	31	37
Helena	37	47	43	39	41	39	29	37
Spokane	48	41	39	37	25	47	29	46

* Values greater than 41, 44 and 48 are significant at the 0.10, 0.05 and 0.01 levels, respectively.

correlation between the three months that traditionally make up summer. July and, to some extent, August also show significant but low skill in the southeastern third of the country. No semblance of this regional skill exists for June. Given that there are only about six effective spatial degrees of freedom for the summer months, the associated global skills are still significant at or considerably above the 5% level.

SLP predictors. The SLP data has little ability to forecast monthly temperatures during the summer (Table 2). The best result (June, Fig. 5b) shows a small region of significant local skill in the northern plains and at few isolated stations; these latter are likely due to chance. However, the skill in this area/stations does not maintain itself in subsequent months. At any rate, the numerical scores are low enough to be of little practical interest.

SST predictors. In this experiment the winter SST and SSTs for the three preceding (spring) months were used to predict the T-field for each summer month. The results are given in Table 3 and the best case illustrated in Fig. 5c. The global skills are significant relative to chance for all three months, even after allowing for the fact that $q_e \approx 6$.

The spatial distributions of skill show a maximum in the western third of the country for all three months. However, as one proceeds from June to August the magnitude of the skills tend to increase. Also, a lobe of significant skill builds from the southeastern part of the country from June through July until it joins the high skill region in the western United States in August (Fig. 5c). These distributions of monthly skill are generally in accord with that obtained from the forecast of summer season air temperature from SST (section 4a).

Mixed predictors. The prediction experiments for the summer months were repeated using the principal components of the individual predictor field as described above. The results are listed in Table 4. The spatial distribution of skill for the most successful set of forecasts, June, is shown in Fig. 5d. Results in July dropped considerably but were still significant (10%) in a global sense. By August, however, only seven sta-

TABLE 4. Skill scores* given as a percentage of correct tercile forecasts, obtained using SST, SLP and a measure of persistence to predict station temperatures in the month/season shown. A value of 33% is expected by chance.

Station	Summer	Jun	Jul	Aug	Winter	Dec	Jan	Feb
Jacksonville	39	41	37	18	29	33	45	41
Charleston	37	45	39	33	33	33	47	45
Mobile	35	39	37	31	33	39	47	50
Abilene	41	39	37	27	20	39	27	41
El Paso	37	27	29	31	25	37	41	35
Phoenix	47	47	43	35	25	33	45	41
San Diego	35	35	47	35	37	27	47	39
Asheville	41	41	47	14	35	37	45	31
Nashville	33	47	52	22	31	50	41	43
Little Rock	54	27	35	10	29	47	47	41
Albuquerque	35	27	33	43	27	33	39	37
Washington DC	35	22	41	25	35	39	43	31
Columbus	41	33	37	27	27	47	37	27
St. Louis	35	20	35	37	22	56	39	43
Denver	35	47	35	39	22	31	27	37
Sacramento	41	47	41	47	25	33	37	41
San Francisco	41	41	41	35	27	39	47	43
Blue Hill	31	31	25	18	25	35	47	25
Chicago	25	25	37	20	20	43	41	33
Detroit	25	27	41	25	22	39	52	22
Des Moines	25	41	52	37	33	39	27	45
North Platte	35	45	29	43	31	25	35	31
Salt Lake	50	52	33	31	25	35	47	41
Winnemucca	54	39	33	35	14	43	41	37
Toronto	20	29	33	18	22	33	43	27
Rapid City	33	43	39	43	35	39	37	27
Sheridan	33	52	41	47	37	25	37	33
Boise	47	37	37	29	25	33	39	37
Portland	37	41	33	31	27	39	45	35
Duluth	43	39	29	45	35	35	30	20
Bismarck	37	43	33	60	35	45	33	35
Helena	37	43	35	31	33	25	41	35
Spokane	47	37	37	31	43	25	47	39

* Values greater than 41, 44 and 48 are significant at the 0.10, 0.05 and 0.01 levels, respectively.

tions had a significant local skill score and the global skill was significant at the 12% level.

Given the high degree of similarity between the persistence experiment (Fig. 5a) and those of the current experiment, it can be concluded that the set of mixed predictors introduces no new information above that already present in the recent history of the air temperature field itself.

5. Origins of seasonal forecast skill

This section uses the model diagnostic features described in sections 3c and A2 to determine the space–time evolution of the climate system that gave rise to the most successful seasonal forecasts. The results may give important clues to the physical processes that underlie extended range predictability. At the minimum, they show clearly what features of the climate system must be understood if we are to explain the causes of extended range predictability.

a. Predictability due to SST

The strongest seasonal prediction result came from attempts to forecast summer air temperature using SST as predictors. The skill came from two very different space–time scales of ocean variability and these are described below.

Slightly over one half (60%) of the predictive skill shown in Fig. 3b was associated with a more or less uniform change in surface temperature in the tropical and high latitude oceans. The equatorial oceans do not substantially participate in this signal. This is illustrated in Fig. 6 (upper), which shows the segment of the canonical predictor (g_1) map associated with May. This portion of the g-map accounted for 21% of the predictive skill. The other two months and winter season were equally as important as May. Thus, all three predictor months and the prior winter season were equally useful in accounting for the predictive skill. Further, the distribution of g-values for the other months and winter were virtually identical to the values shown in Fig. 6 (upper). Thus, the SST predictor information is

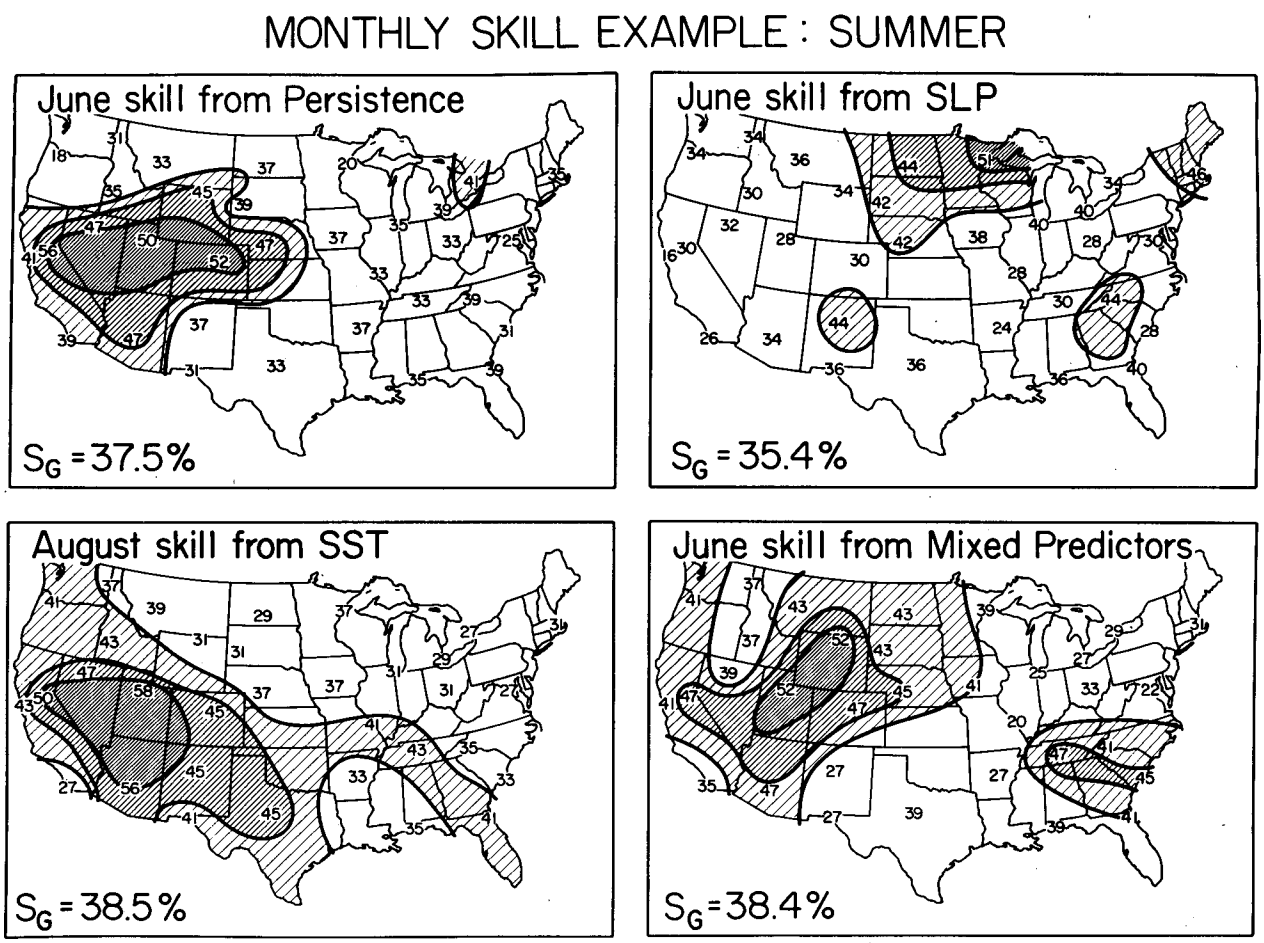

FIG. 5. As in Fig. 4 except for specific months during the summer season.

highly redundant in time; May data, say, could have done the job by itself.

The nature of the signal shown above becomes clear by investigating its temporal strength, $u_1(t)$ (Fig. 7a). In view of (A12) and the sign of the larger g-values (Fig. 6, upper), it is clear that $u_1(t)$ represents a warming of the Northern Hemisphere oceans from the early 1930s, when the data begin, to the late 1940s or early 1950s followed by a cooling until the mid-1960s or 1970s. Perhaps a slight warming then begins until the end of the dataset. This signal is similar to the behavior of the Northern Hemisphere surface temperature field as described by Jones et al. (1982) and others. The former authors' pentad-averaged estimates of this parameter are shown in Fig. 7a. The similarity with the $u_1(t)$ is reasonably good, although there is a small phase shift between the two signals (cf. Barnett, 1984) and the extreme cold of the late 1960s is not so obvious over the oceans. Remember, however, the predictor/predictand data used in this study was detrended prior to analysis. The Jones et al. data have not been detrended. The spatially global nature of the canonical predictor signal and its long time scale explain why all the important components of the g_1-map were essentially identical in sign and value and why they were all equally useful predictively. The lagged ocean temperatures go back only six months in time; that interval is small compared to the decadal time scale of the $u_1(t)$.

Inspection of the canonical predictand map (h_1, not shown) and $u_1(t)$ shows that the forecast skill is associated with decreasing temperatures in the eastern half of the country and increasing temperatures in the western half during the period 1950–75. This is just the pattern of change obtained over nearly the same time span by van Loon and Williams (1976, their Fig. 2), thus confirming the current analysis. However, the current study now shows that these changes are associated with a global climate change. They are not simply local changes.

A caution is in order with regard to the preceding discussion. The predictand signal clearly violates the independence criteria associated with the cross validation methodology. However, it was of such low magnitude as to be buried in the noise when the de-

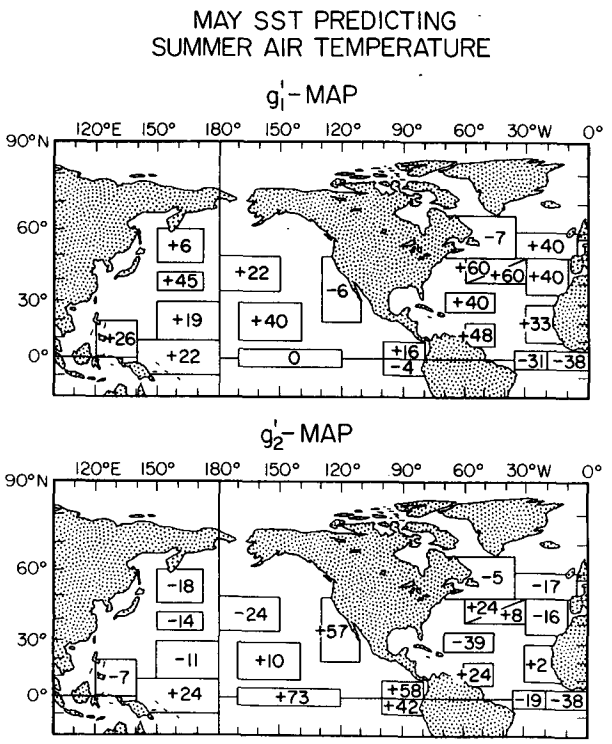

FIG. 6. Canonical predictor maps relating prior, regionally averaged SST to summer air temperature. The numbers represent the relative importance of each region in producing the summer forecast skill associated with SST (see Appendix). The upper panel shows the first canonical map while the lower panel shows the second canonical map.

correlation scales of the warm season temperature were computed. Given this fact, it seems likely that the results given above are qualitatively correct but the magnitude of the skill shown in Fig. 3 is somewhat inflated.

The remaining predictability (40%) during the summer season comes largely from the equatorial Pacific (cf. Fig. 6, lower), although the equatorial Atlantic and midlatitude North Pacific, both in antiphase with the equatorial Pacific, contribute to the skill. In contrast to the first canonical mode discussed above, the value of antecedent predictor data drops as one recedes further from the summer season.

The SST predictor pattern determined here is similar to the first EOF of the global SST field shown by Hsiung and Newell (1983; see also Weare et al., 1976). The associated canonical component vector (u_2) is shown in Fig. 7b and, not surprisingly, closely resembles the first principal component shown by Hsiung and Newell (cf. their Fig. 4). It is clearly dominated by ENSO events, a result already suggested by Fig. 6. It is worth noting here that the winter season forecast skill is also associated with g-maps and u-series that resemble those shown above. The summer forecast skill, like that of winter, is concentrated in the southeastern third of the country. The sense of the h_2, g_2 maps and the u_2 series is such that warm SST events in the equatorial Pacific go with cooler than normal summer temperatures over the southeast. The same relation holds for the winter (cf. Barnett, 1981a).

In summary, about half of the summer forecast skill is associated with multidecadal changes in the temperature of the Northern Hemisphere oceans and surface air temperature field. The reasons for these changes are unknown. The remaining skill is associated with the shorter period ENSO events but also involves changes in SST in other ocean regions besides the equatorial Pacific. The canonical predictor pattern associated with this forecast skill is nearly identical to the first EOF of the global SST field.

b. Nonpredictability of winter season air temperature

None of the models discussed above did a good job of predicting the winter season air temperature. Yet the models for the individual months of winter did extremely well. This apparent difference may be explained by two factors.

(i) The temperature fields for the individual months of winter are not well correlated with each other. Thus, forming an average of the three months is close to averaging three unrelated variables together. This fact is demonstrated in Fig. 8. Only 6 of the 33 stations showed significant correlations between all three months of winter, while 11 stations demonstrate no significant correlation between any of the months of

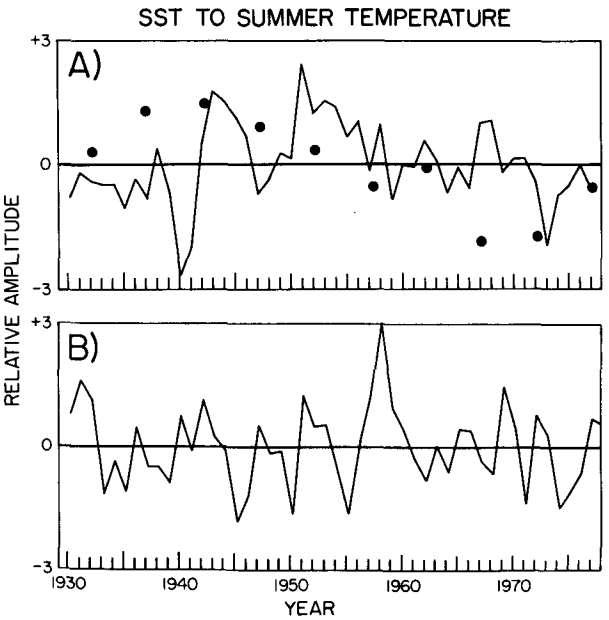

FIG. 7. (a) Canonical coefficient vector, u_1, associated with the g_1 map shown in Fig. 6 (upper). The heavy dots are pentad averages of Northern Hemisphere surface air temperature from Jones et al. (1982). (b) The second canonical coefficient vector, u_2, associated with the g_2 map shown in Fig. 6 (lower).

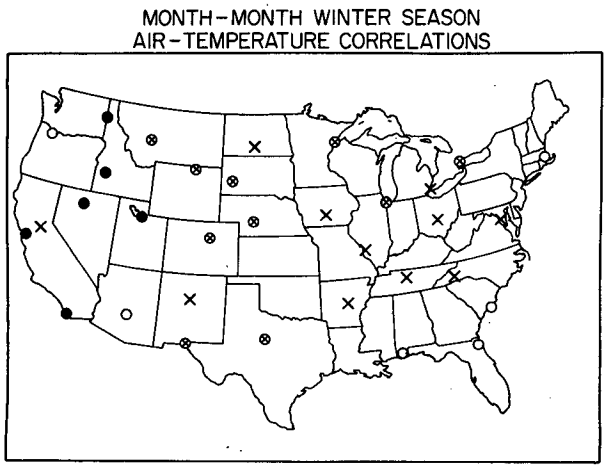

FIG. 8. Month-to-month air temperature correlation during the winter season. The solid circles represent stations for which all three winter months are significantly correlated with each other. The open circles represent stations where two out of three winter months are correlated. The ×'s represent stations where two out of the three winter months are not correlated. The circled ×'s represent stations where none of the three winter months are correlated. Over most of the United States the month-to-month air temperature changes during the winter season are unrelated.

winter. It appears that this way of looking at persistence within the winter season has not been done before, although the work of Dickson (1967) is close and produced much the same result as found here. Other persistence studies (e.g., van den Dool et al., 1986) perform substantial space–time averaging of their results and thus have missed conclusions drawn here. At any rate, it seems clear that the strategy of trying to forecast a seasonal aggregate of uncorrelated variates with a single model will not be very successful (and it wasn't). This brings into question the practical usefulness of defining a "winter" season for prediction purposes.

(ii) The results of section 4 showed that the individual months of winter were relatively well forecast. The discussion of section 6 will show that the key features of the atmospheric field that gave this forecast skill were different from month to month. Thus, no single model will capture well the air temperature variance that is associated with the traditional definition of winter. If one insists on predicting a traditional winter average, then apparently it will be necessary to forecast individually each of the winter months and then to average those forecasts into a "winter" forecast.

6. Origins of monthly forecast skill

a. Space–time evolution of predictor patterns

The monthly skills for the winter season from SLP predictors were the best found in this study. The g-maps associated with this skill suggest the space structure and temporal evolution of the SLP field responsible for the success. This is demonstrated in Fig. 9 where the g'_1 map, which accounts for over 80% of the January skill (Fig. 4b), is partitioned into the months that made up the predictor field [cf. Eq. (A1)]. Thus, the relative predictive importance (RPI) of October, November and December can be evaluated (see Appendix, section 2).

The results show that approximately 57% of the skill in the January forecasts comes from SLP variations in December alone. However, traces of the December pattern are clearly evident in the preceding two months, particularly over the ocean regions immediately adjacent to the U.S. mainland. The key pattern is a strong low (high) pressure anomaly over the central North Pacific Ocean, associated high (low) over the Rockies and southwest Canada, and another low (high) over the southeastern corner of the United States and extending to Bermuda. This is reminiscent of the Pacific–North American (PNA) pattern (Horel and Wallace, 1981; Wallace and Gutzler, 1981). Based on Wallace and Gutzler's work we can conclude that the patterns seen in the SLP data are essentially barotropic and thus representative of the overlying troposphere. Note that the pattern is well represented as a standing wave since its centers of action show little movement as it develops.

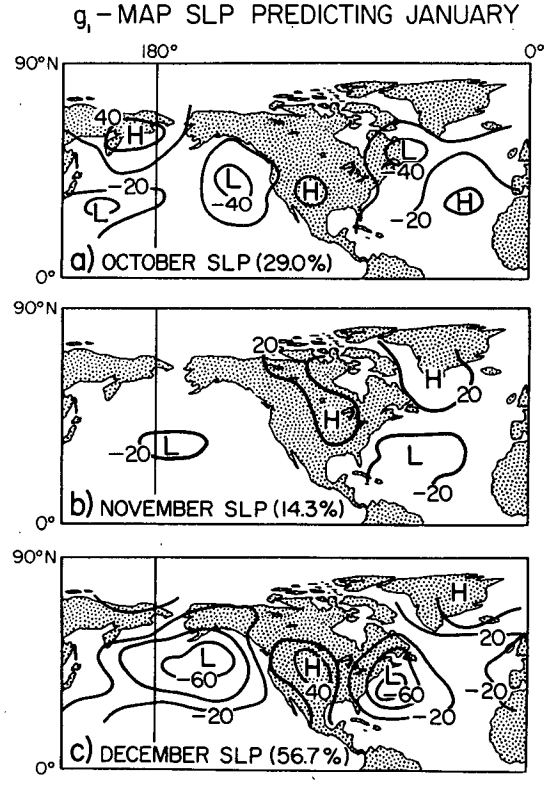

FIG. 9. First canonical predictor maps associated with surface air temperature forecast skill during the month of January. The percentages in parentheses in each panel represent the relative importance of each month to the overall forecast skill associated with this mode. The contours on the maps represent relative (normalized) importance of the SLP signal to the subsequent air temperature forecast.

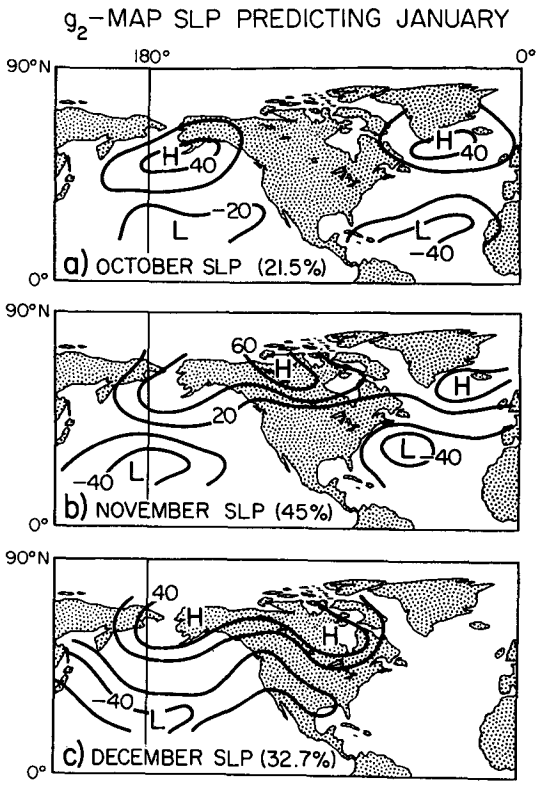

FIG. 10. As in Fig. 9 except for the second canonical mode.

The g'_2 function, which accounts for all the remaining January skill (20%), is shown in Fig. 10. It represents a very different, yet important, distribution of SLP, as we shall see below. Note that all three predictor months contribute to the skill. This means the time scale associated with this pattern is much longer than the time scale associated with g_1. The higher order canonical modes were insignificant compared to modes 1 and 2.

The first principal predictand map (h_1) for January associated with the g_1-pattern is shown in Fig. 11. As expected, it compares well with the map of January forecast skill. The distribution of SLP that leads to the h_1 pattern strongly suggests that it is just the advection of warm/cold air masses that is largely responsible for the observed distribution of air temperature anomalies.

The first two principal predictor patterns associated with the February forecast skill are shown in Figs. 12 and 13. Comparing them with Figs. 9 and 10 shows that the dominant predictor patterns for January and February have simply reversed roles, e.g., g_1 for January forecasts has become g_2 for February forecasts. Exceptions to this statement occur over the Atlantic Ocean, but the differences are not large. The g_1 function for the February forecast accounts for over 70% of the global forecast skill for that month. January is clearly the dominant month within this g-map, accounting for 64% of the total skill associated with this first canonical mode. Thus the forecast skill for January and

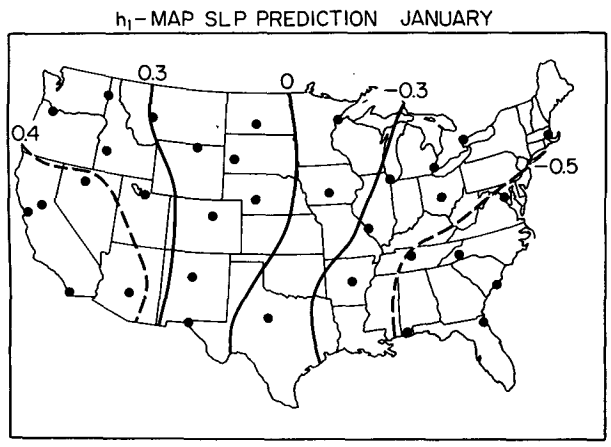

FIG. 11. The canonical predictand (h_1) map associated with January forecast skill. The contours show the most predictable pattern of surface air temperature during January. Units are in standard deviations.

February (and December, not shown) are due to rather different distributions of SLP. This helps explain the prior results in section 5b, which showed low month-to-month correlation during the winter season.

It should also be noted that the February pattern, whose full spatial extent is not completely resolved,

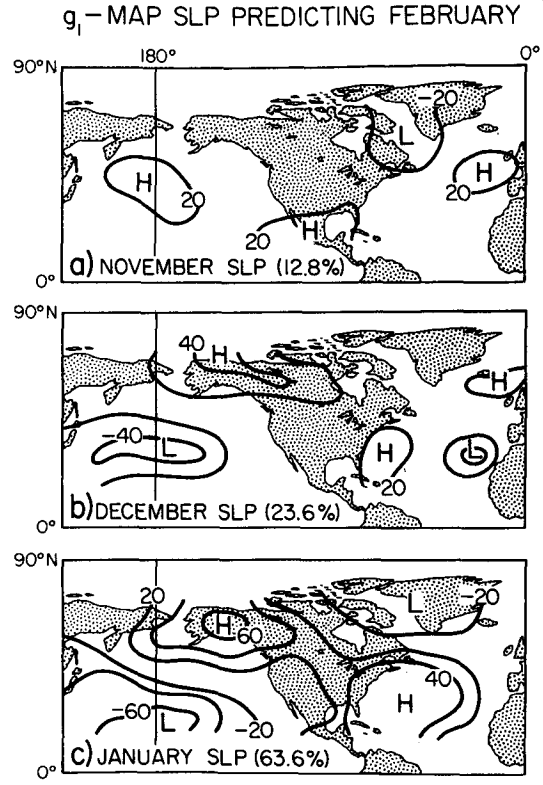

FIG. 12. As in Fig. 9 except for February.

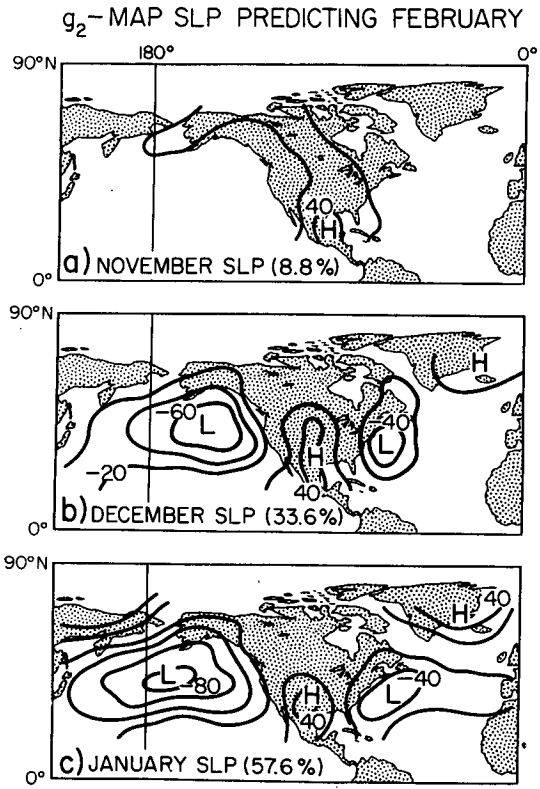

FIG. 13. As in Fig. 10 except for February.

bears a resemblance to the North Pacific oscillation of Walker and Bliss (1932) or more recently to the West Pacific and West Atlantic teleconnection patterns discussed by Wallace and Gutzler (1981) or the tropical/Northern Hemisphere pattern mentioned by Barnston and Livezey (1987). Like the g_1 pattern for January, it is also likely to be a barotropic feature.

A basic property of the large-scale structures described above is their characteristic time scale. The rapid intensification of the g_1 patterns in the month immediately preceding the January and February forecast periods suggests the time scale question cannot be well investigated with the monthly averaged SLP data used here. However, monthly data can provide a crude estimate of the characteristic lifetime of the large-scale structures. These characteristic times were estimated as follows:

(a) A specification experiment was run where the SLP fields for the months of November through March were used to "predict" January air temperature.

(b) The g_1 map segment for January was isolated and will be referred to as $g_1(0)$.

(c) The pattern correlation (PC) between $g_1(0)$ and the g-patterns for the preceding and subsequent two months was computed, e.g., $\langle g_1(0)g(-1)\rangle_x$ = PC between January and December. The results are shown in Table 5a.

(d) A simple significance test on the PC determined the decorrelation time of the principal predictor patterns relative to January.

(e) The same procedure was repeated for a February specification using data from December through April (ref. Table 5b).

The results (Table 5) show that both large-scale structures discussed above grow to a maximum in a space of two to three months. However, their decay occurs within one month (e.g., February–March PC = −0.19), leading to the notion of a total collapse of the structure in a time that is short relative to their growth period. This strong asymmetry suggests a highly nonlinear mechanism is responsible for their appearance in the first place.

The above results would be misleading if the structures showed propagation. The January specification structures (not shown) appear quasi-permanent from November to January, with the centers of action beginning to shift in February and perhaps accounting for the fall off in PC. The February structures show high permanency over the Pacific and North America for December–February. However, these characteristic patterns are simply not evident in the March map and hence the small PC value.

An interesting feature of the monthly patterns is the strong difference in skill they give in the specification. December and particularly January account for the majority of specification skill in January air temperatures. However, for February specification the months of December through February are all equally impor-

TABLE 5. The G-map Pattern Correlation (PC) between month of specification and both antecedent and subsequent months. Relative Predictive Importance (RPI) shows the percent of total skill contributed by the various months used in the specification.

Months	PC	RPI (%)
	a. January specification	
Jan–Nov	0.56	6.9
Jan–Dec	0.86*	24.5
Jan–Jan	1.00*	47.0
Jan–Feb	0.55	14.0
Jan–Mar	−0.70	7.6
		100.0 sum
	b. February specification	
Feb–Dec	0.73*	24.7
Feb–Jan	0.74*	25.8
Feb–Feb	1.00*	25.8
Feb–Mar	−0.19	9.6
Feb–Apr	0.18	14.1
		100.0 sum

* 95% significant assuming four degrees of freedom in the spatial fields.

tant. These results again attest to the spatially steady nature of the structures.

b. *Global connections*

It is logical to wonder if the primary g-patterns discussed above are relatively "local," i.e., confined to the region immediately surrounding North America. The answer to this question was obtained in several ways. A prediction experiment using the near-global SLP field as the predictor data was carried out for January air temperatures. The resulting forecast skills (not shown) were considerably lower than those found previously and not highly significant in a global sense. Significant local skill was found in the southeastern section of the country only. The associated g-maps (Fig. 14) suggested the key predictor features noted above have only a modest relation to global-scale variation in SLP, i.e., the Southern Oscillation (SO) in the Southern Hemisphere. The teleconnection to higher northern latitudes (e.g., Fig. 9c) is similar to that described by Bjerknes (1966). However, the correlation between $u_1(t)$ for this global analysis and an SO index was only −0.45. We shall see later that the tropical connection to the January g_1 map is not strong (ref. section 7a).

The February forecast experiment was also repeated as above on the near-global SLP set. Again, the global skill was of modest significance. The g-patterns shown for the limited SLP prediction experiment were found to be stable and again apparently linked to variations in the tropics and, particularly, the Southern Hemisphere. Perhaps most striking, the shape of g_1 patterns closely resemble the large-scale, moving global SLP patterns derived by Barnett (1985) by a very different method. This is illustrated in Fig. 15, which shows the spatial detail of the February pattern and its clear link to the tropics. The time sequence of the g_1 patterns

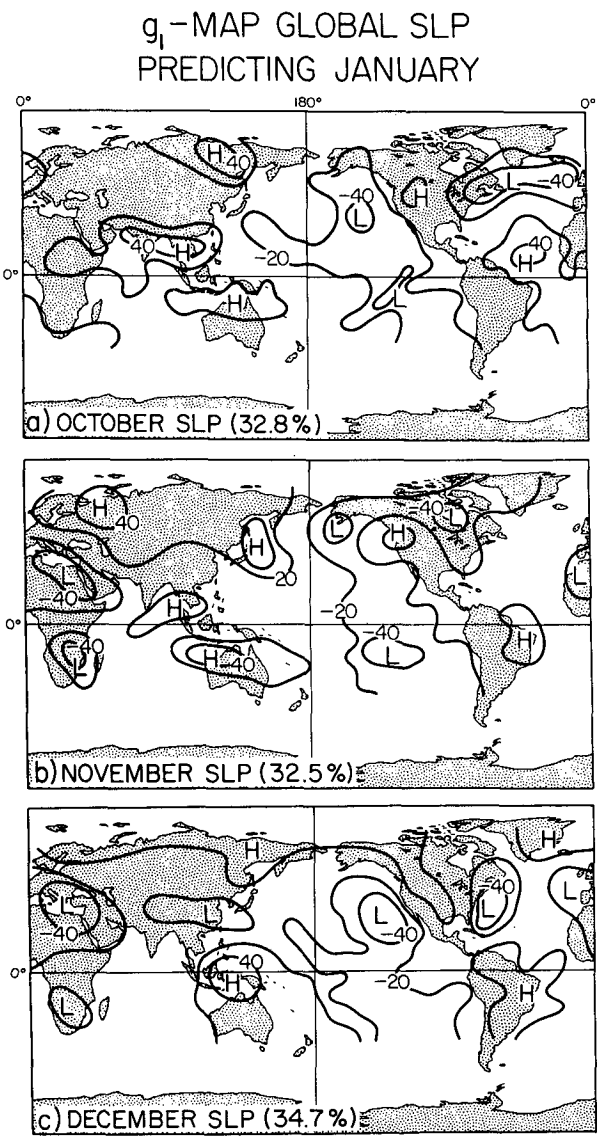

FIG. 14. As in Fig. 9 except the near-global SLP fields were used as predictors of January air temperature.

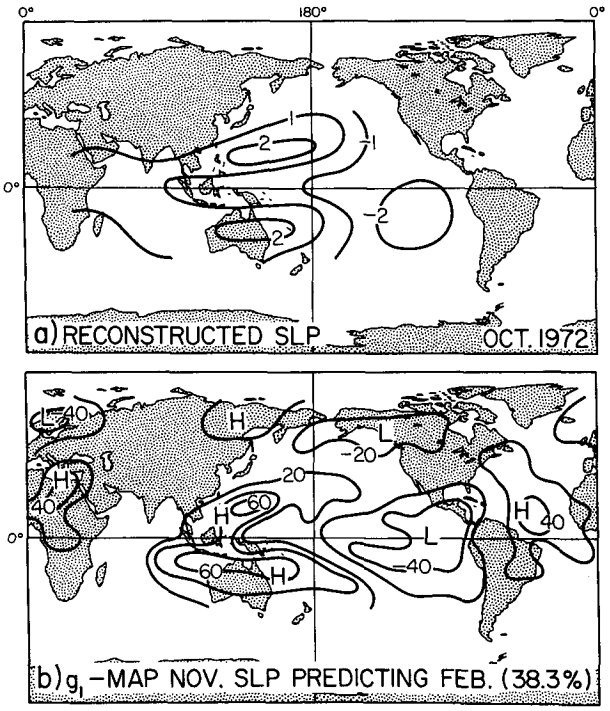

FIG. 15. (a) Reconstructed SLP pattern (relative units) associated with large-scale moving sea level pressure patterns described by Barnett (1985) and (b) November normalized g_1 map values obtained from using near-global SLP to predict February air temperature. The percentage in parentheses (lower panel) indicates the relative importance of this month to the distribution of February forecast skill. The illustration suggests the importance of the near-global moving SLP pattern to forecasts of February air temperature over the United States.

(not shown) in this study also suggests the type of motion and signal-bifurcation about the equator found in Barnett (1985). Thus, the February results suggest the predictive skill is in part due to a global-scale SLP variation that is linked to more local changes over the North Pacific and North America.

The suggestions posed above may be partially tested by quantitatively answering the question: How much of the forecast skill is due to tropical forcing and how much is due to local forcing? The answer to this question was obtained as follows. The SOI index was regressed against the regional SLP field (20°–70°N) and the variance associated with the SOI removed. The residual SLP field was then used in an experiment identical to the one that produced the results shown in Fig. 4. The SLP predictor field with the SO signal removed produced an average value of local skill for January forecasts of $S_G = 42.7$ (versus 46.2% previously) and a spatial distribution of skill almost identical to that found previously. The only noticeable difference between experiments was a modest reduction in forecast skills in the eastern portion of the country for the SOI-removed case. The largest reduction was 18% at Toronto (from 61% to 43%) but the more typical value was 7%–10%. However, the local skills in the eastern third of the country remained significant.

The forecast results for February showed a much stronger loss of skill in the southeast and virtually all skill in the central portion of the country. Both areas now demonstrated no significant values of local skill. However, the global skill was still statistically significant, but only marginally. The interhemispheric connections are apparently much stronger for the February predictor pattern than for January.

We conclude that global-scale SLP phenomena associated with the SO do contribute to the forecast skill over North America, particularly in the southeast (cf. Barnett, 1981a). However, the use of these global predictors gives values of S_G over the United States that are generally lower than those obtained from local SLP predictors. Thus, local changes in SLP, changes that are uncorrelated with an SOI, are more effective predictors and capable, by themselves, of producing significant forecast models. This statement is most true for January forecasts. It is also true for February forecasts, but only in an integral sense. In this month, removal of the SO signal also removes most of the predictive skill over the eastern third of the nation.

Thus, much of the forecast skill is due to large-scale coherent structures in the SLP that generate rapidly relative to a month and then persist for one to two months, giving them typical lifetimes of two to three months. The main energy in these structures is confined to the regions over North America and its contiguous oceans. These patterns bear similarity to known teleconnection patterns. It should also be clear to the reader that the monthly data used in this paper is not really satisfactory to resolve the life history of these features.

7. Discussion

This section presents a brief discussion of several aspects of this study. Its purpose is to pull together different results and also to present some critical insights to the work and analysis method.

a. Origins of predictive skill

The results of sections 5 and 6 give a relatively clear picture of the space–time evolution of the climate fields that leads to predictive skill. However, the results must be tempered by the fact that two of the main features delineated, the near-global warming trends on decadal time scales and the extremes of the ENSO events, are not unknown features. What is new is the relative importance of large-scale coherent structures in the SLP field to the wintertime predictive skill. Further, it appears that virtually all the forecast skill found for surface air temperature over the United States is due to these three phenomena.

A fundamental finding of this study has to do with midlatitude SLP features that give rise to the forecast skill. The first canonical predictor (g_1) pattern that gave over 80% of the January global forecast skill (Fig. 9c) was seen to originate mainly in December. This feature resembled the PNA pattern. Additional studies given here and to be reported elsewhere showed that this pattern reaches maximum intensity in January and begins to damp strongly in February. But the g_2 function for January, which accounted for 20% of the skill, was nearly identical to the g_1 pattern for February, particularly over the Pacific Ocean and North America. During February, this pattern, which resembled a combination of known teleconnection patterns, became dominant and accounted for 70% of the February forecast skill. The g_2 function for February forecasts was just the g_1 pattern for January.

The above results raise two possibilities: (a) One can imagine that the main January pattern (Fig. 9c) is dynamically transformed into the main February distribution of SLP (Fig. 12c); (b) alternatively, the two patterns may have nothing to do with each other and the canonical mode switching is due only to the facts that they tend to occur at different times of the year or in different years and, when they do occur, persist for two to three months. The validity of these two possibilities was checked by simply cross correlating the S_G time series for January and February. The correlation was −0.07, an insignificant value. Thus, a successful January forecast is unrelated to the forecast skill expected for February and vice versa. If (a) above was true, the correlation should have been high, thus guaranteeing that a good January forecast would be followed by a good February forecast. This was not the case. Thus, we accept (b). The above results mean there are two separate, uncorrelated large-scale structures in the SLP field that give the forecast skill. Further, these distributions of SLP apparently represent preferred modes

of atmospheric variability. Their occurrence is tightly tied to the annual cycle, but their most energetic life times are roughly equivalent, two to three months, although one of the structures does persist much longer than the other. Given only these facts, we cannot determine the physical processes responsible for the main SLP predictor signals. We can say, based on earlier results (section 6b), that while the general state of the climate system, e.g., ENSO or non-ENSO, may set the stage for these regional events, the ENSO phenomena is neither a necessary nor a sufficient condition for a successful forecast nor for the actual occurrence of the large-scale structures.

The results of the mixed predictor experiments suggest that the precursor signals that lead to forecast skill tend to be found in one or more of the climate predictor fields used in this study. If there were uniquely different skill-producing signals in the different fields, then the mixed predictor experiments should have given higher scores than any one of the individual field experiments. This was not the case.

Finally, the result that certain fields should be more useful as predictors than others depending on the nature of the predictor signal seems logical. For instance, the small, low-frequency signal associated with the decadal scale changes would be expected most clearly in the ocean due to their large thermal inertia, long time scale and the resulting signal/noise ratio that would be larger than that expected for, say, SLP. By the same token, rapidly changing predictor patterns such as determined for the winter months should be manifested best in the atmosphere, as opposed to, say, the more sluggish ocean variables.

b. *Forecast skills*

The scoring on the forecasts represents approximately the procedure one would follow in evaluating the relative skills of a forecast model over approximately 50 years of pseudo-independent testing. It is gratifying that large spatial regions of significant skill exist. Yet the values are generally low, even considering the liberal method of terciling (cf. section 3b). Remembering that the analysis methods used here are specifically designed to get as much skill as possible from a linear model suggests that the *average* ability to forecast short-term climate change is small. Major exceptions to this conclusion occur during the winter months, particularly January and February. In any event, the forecast skills estimated here are generally much less than those expected from studies of potential predictability. Indeed, if one is ever to approach the optimistic expectations of these studies, then it will be necessary to (a) find predictor data that is uncorrelated with SST, SLP and air temperature or (b) devise a highly nonlinear prediction scheme that is poorly approximated by any linear model or (c) introduce, at least conceptually, physical-synoptic considerations which are, of necessity, highly qualitative and, in addition, make no use of the predictor/predictand relationships presented here.

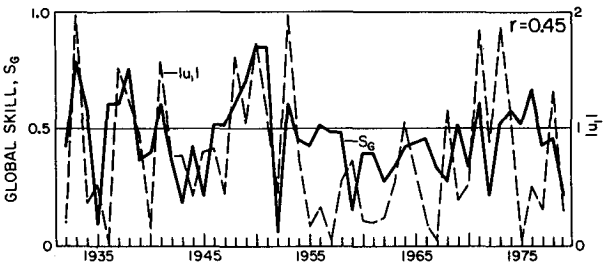

FIG. 16. The yearly values of average local skill (S_G) associated with the prediction of January air temperatures from regional SLP data (heavy line). The dashed line represents the time history of the absolute value of the first canonical coefficient vector for the January forecast model. The correlation between the two series (r) was 0.45, suggesting that $|u_1|$ can be used to predict the "goodness" of the forecast itself.

The idea of averaging skill over a number of years, the basis for the above comments, may not be a particularly meaningful measure of forecast ability in the first place. This point of view is supported by the temporal distribution of global skill obtained from forecasts of January air temperature and the $|u_1(t)|$ that accompanied each forecast (Fig. 16). The large variability in $|u_1|$ and large relative size of the associated μ_1 guarantee large variations in S_G. In other words, the mean value of S_G, by which we traditionally measure a model's success, gives only a partial view of a model's ability.

In view of the correlation between $|u_1|$ and S_G (0.45) it is clear that the value of $|u_1|$ gives an indication of how good a forecast to expect for each of the winter months. Since $|u_1|$ is estimated as the forecast is made [ref. (A3) and (A8)], it can be used to develop a forecast quality index. Examples of how this might be done are given for one station that had high forecast skill (Toronto) and one that had low skill (El Paso). Note the local skill scores are used in this example since S_G can be misleading itself, i.e., it is obtained by averaging over stations that had significant local scores *and* those that did not.

The paired values $|u_1|$ and S_L are presented in the form of a contingency table (Table 6) where the terciles of $|u_1|$, called Q_L, Q_N and Q_A, are used to order the forecasts whose skill is expressed as percent correct, percent 1-class errors (e.g., "A" forecast and "N" observed) and so on. By way of an example, if $|u_1|$ is in the upper tercile (Q_A) then the probability of a correct forecast at Toronto is 75%, while if $|u_1|$ is in the middle tercile (Q_N) then the probability of a correct forecast drops to 50% (versus 33% expected by chance.) Other stations in the eastern third of the country had similar results. On the other hand, at El Paso the similar probabilities are 18.8% and 12.5%. The scatter in the $|u_1|$ versus S_L distribution at El Paso is indicative of a low forecast ability situation, while the results for Toronto

TABLE 6. Local skill vs terciled forecast quality index for a high skill station (Toronto) and low skill station (El Paso). This table shows the probability of obtaining a correct forecast contingent on the absolute value of the canonical component vector when the latter is partitioned into terciles, e.g., Q_A refers to the upper 1/3 largest values of $|U_1|$. The results suggest the idea of a time-averaged forecast skill may be misleading.

	Toronto		
% correct	56.2	50.0	75.0
% 1-class errors	43.8	18.8	12.5
% 2-class errors	0.0	31.2	12.5
	Q_L	Q_N	Q_A
	El Paso		
% correct	25.0	12.5	18.8
% 1-class errors	50.0	37.5	25.0
% 2-class errors	25.0	50.0	56.2
	Q_L	Q_N	Q_A

indicate the opposite situation. The results show that even in the best of situations one cannot always expect to have a good forecast. Fortunately, the development of a forecast quality index provides a quantitative way to put a confidence factor on a forecast.

In summary, *average* forecast skills are generally low. In view of the methods and data fields used here, we should not anticipate significantly larger values using other linear methods. However, the skills for the winter months are high enough to be practically useful. The use of a forecast quality index seems a potentially powerful way to estimate in advance if a forecast is apt to be good or bad. This index also suggests that time-averaged skills, which are commonly used in the literature, give a misleading view of one's ability to forecast specific climate events.

c. *Sensitivity of results*

There were a number of decisions that had to be made in the course of carrying through the analysis described in section 3 and the Appendix. This subsection offers brief comments on the sensitivity of the results to these decisions.

Predictor truncation limit, *p*. EOF filtering rules were used to select p in (A2). If these were too stringent, then valuable predictor information could be lost to the analysis. We found that increasing p above the objectively determined limits quickly led to models with no forecast skill. In these cases, the CCA used the larger available degrees of freedom to fit the predictand data more accurately (increased hindcast skill) but at the expense of the forecast skill. We still cannot rule out the possibility that some very high eigenmode of Y that we have neglected could increase forecast skill, but we know of no way to evaluate this possibility in a satisfactory manner.

Predictand truncation limit, *q*. The filtering rules were used to estimate q in (A2). In general, this approach was satisfactory to resolve the large-scale skill patterns. However, we noted at least one problem on a very local level with this procedure. As noted in section 4b1, one would have expected persistence to be more skillful than it was in this study for the two stations on the California coast. It was found that these stations dominated eigenmodes 5–7 in the EOF approximation of T (cf. A2). However, the objectively determined expansion cutoff, q, was 3. Therefore the model had no chance to forecast the local variability associated with these stations. Thus, the station forecast skill found in this study must individually be thought of as lower limits for there may be isolated, local processes that could increase the skill at specific stations, e.g., persistence near large bodies of water (Namias, 1978; Van den Dool et al., 1986). Variation of q within reasonable limits did not affect the large-scale patterns discussed in sections 5–7.

Canonical mode truncation limit, *q″*. Three different approaches to estimating q'' were used and the resulting models all gave about the same result [cf. discussion following Eq. (A14)]. In general, all approaches gave q'' values ranging from 1 to 2. Larger values of q'' in (A14) gave models with rapidly declining S_G values. Further, the value of μ_1 was generally considerably larger and statistically distinct from the second and higher-order canonical correlations in most of the cases studied. In short, we do not feel that the results presented here were sensitive to our method of selecting q''.

Cross validation. Calculations showed the winter temperature data to be uncorrelated at time lags of one year while summer data was uncorrelated at one–two year lags. Under these conditions, the cross-validation techniques used here are strictly valid. As an additional check, we repeated both the January T prediction from SLP and the summer T prediction from SST using an expanded data omission window in the cross validation. Thus, for January, both the year before and year after t_v were omitted completely from the training and test sets. In the case of the summer forecasts, the two years immediately before and after t_v were similarly omitted. In both cases, the resulting skill scores, etc., were essentially the same as reported above.

A further check on the robustness of the results was obtained by using the 1930–80 data to build a prediction model relating SLP to January T. This model was then tested on the independent, but considerably more problematical, dataset for the period 1900–29. Again, the results reported above were reproduced, even in spatial detail, with surprisingly good fidelity; S_G = 42.1% on the early period versus S_G = 46.2% for recent times. In summary, the cross-validation approach seems to offer a satisfactory measure of model forecast ability and significance in the a posteriori setting of CCA.

Spatial degrees of freedom, q_e. The estimation of the spatial degrees of freedom is based on a normalized

integral of the *covariance* between all pairs of stations (A17). R. Livezy (personal communication) suggested that the *correlation* matrix between stations might be more appropriate in (A17). We have used this suggestion and find typical values of q_e by his approach are one to two times those obtained by use of the normalized covariance in (A17). Thus, the estimates of q_e stated in the text are, if anything, low. Our statements regarding spatial degrees of freedom and significance thus tend to be conservative.

Model stability. In the process of significance testing, a total of 48 ostensibly different models must be constructed. Reviewing the key model parameters ($u_i(t)$, g_i- and h_i-maps, μ_i, etc.), we found rms variations in the model coefficients of less than 10% over the ensemble of models for the leading two canonical modes; i.e., the models are stable. This means the interpretations of section 5–7 and results of section 4 are quite stable.

Multiple tests. Forecast experiments were conducted for six individual months and two seasons. Three different predictor fields were used. Given these 24 experiments, one might expect one of them to show an S_G value significant at the 95% level simply by chance. A "super global" significance test aimed at this problem was conducted. Provided one of the $\hat{S}_G$ values was significant at the 0.998 ($=0.95^{1/24}$), the results of this study can be claimed significant at the 95% level. Two values of $\hat{S}_G$ exceeded this significance level while six exceeded 0.991. If one allows that the values of q_e are conservative by a factor of 2 [see above] then 13 out of 24 values of $\hat{S}_G$ are significant at the 0.998 level or above. Thus, it appears the strongest results are not due to chance associated with multiple tests. More importantly, the predictor/predictand patterns for the most robust experiments make physical sense in the context of recent studies of climate dynamics. Thus, while some of the marginally significant results may be due to chance, it seems clear that the strongest ones represent real ocean/atmosphere relationships.

In summary, the model building procedure and significance testing appear robust and not sensitive to the objective methods used to determine key model parameters. The exception to this statement is the possible omission of local features in the T-field that could increase skill at a specific station.

8. Conclusions

The predictability of surface air temperature over the United States has been investigated with advanced statistical techniques designed to maximize predictive skill. The ability of a form of persistence, SST and SLP fields to forecast temperature was investigated. The cross-validation methods used to check the models should give a good approximation to the actual forecast skills expected on independent data. More important, the procedures show which space–time features of the climate system are responsible for the predictive skill.

The principal results of the study were as follows:

1) Analysis of the prediction models' performance showed that virtually all of the forecast skill found in this study came from three climatological features: (i) a decadal scale change in Northern Hemisphere temperature, (ii) ENSO-related phenomenon, and (iii) the "fast" development of two different large-scale coherent structures in the atmospheric field overlying North America and its contiguous oceans. These structures are quasi-stationary and have lifetimes of two to three months. However, their growth/decay cycles are highly asymmetric in time with the decay phase being fast relative to one month. The horizontal advection patterns associated with these structures explain their ability to predict surface temperatures.

2) The physical mechanisms responsible for the first two predictive signals are currently unknown, although the ENSO phenomenon is under intense study. The physics associated with the third signal, the large-scale structures, are also largely unknown. One can conjecture that these features are due mainly to internal atmospheric dynamics. This conclusion is based on their rapid, localized growth and decay, which seems to preclude an external forcing (e.g., SST), and to the fact that they are either weakly unrelated or only related to the other two predictive signals noted above. However, there are counter arguments that make this suggestion highly speculative. In any event, it is the signals associated with ENSO and the large-scale structures that must be understood and modeled if advances in short-term climate prediction are to be made.

3) The *average* forecast skills, while highly significant statistically, were low in numerical value and represent only a small amount of variance (5%–20%) in the air temperature field. The exceptions to this statement occurred for individual winter months of January and February when useful forecast skills were found. In all cases studied, the forecast skills are considerably below the suggested values obtained in potential predictability studies. The forecast skills were stable to model perturbation and reproduced well on a 30-year segment of data that was totally independent of any other aspect of this study.

4) The predictor signals are more or less common to all of the data fields used in this study, although they express themselves more strongly in one field or another depending on phase of the annual cycle. This means that additional forecast skill will not be obtained by *linear* techniques unless they include new information that is mutually uncorrelated with Northern Hemisphere SST, near-global SLP and the air temperature field itself.

5) The results suggest that averaging forecast scores over many years, like averaging over wide geographic regions, does not provide a meaningful measure of predictability. The expected model performance itself can be predicted at the time a forecast is made. Under

favorable circumstances we found successful tercile forecasts can be expected 70%–80% of the time. Future prediction models should offer such "forecast reliability indices" if their results are to be really useful, for it appears that climate forecasts cannot be made with the same reliability from year to year.

6) It was shown that the traditional definition of winter is not useful for forecasting purposes. This is so because the months that make up winter are largely uncorrelated with each other over large portions of the United States. Further, the nature of the atmospheric patterns that do a good job of predicting each winter month are themselves different from each other. Under these conditions no single statistical model can be expected to predict the aggregate of months normally termed winter.

Acknowledgments. This work was supported by the Climate Dynamics Program of the NSF under grants NSF ATM82-13279 and NSF ATM85-13713 and the National Climate Program Office (NOAA) under Grant NA81-AA-D-00054. One of us (R.P.) was supported by the Pacific Marine Environmental Laboratory of NOAA. Some of the work was carried out while one of us (T.P.B.) was a visiting scientist at the Max Planck Institute, University of Hamburg. The authors are indebted to Tony Tubbs for carrying out the extensive computations required and to Virginia Roberts and Philomène d'Ursin for preparation of the manuscript. John Horel, Jerry Namias and John Roads offered useful discussion during the work and comments on a first draft. Anonymous reviewers are largely responsible for the separation between the qualitative discussion of the methodology (section 3) and its more rigorous presentation (Appendix).

APPENDIX

Modeling Theory

This Appendix provides a brief review of the prediction methodology used in the main body of the paper and described qualitatively in section 3.

1. Data compression

A key feature of our analysis involves concatenating spatial fields of predictor data from different times so that one can define both the space and time evolution of the climate system that gives rise to predictive skill. Consider as predictors the SLP field denoted by SLP (t, τ, η), where $t = 1, 2, \cdots, n$ is a year counter; $\tau = 1, 2, \cdots 12$ is a month counter, e.g., $\tau = 1$ is for January; and $\eta = 1, 2, \cdots, m$ is a grid-point counter in two-dimensional space. Let us now suppose we wish to predict some temperature field $T(t, \tau, \eta)$ in December ($\tau = 12$) using SLP data from the prior three months (September, October, November, $\tau = 9, 10, 11$) of year t. The composite predictor set (Y') would be defined by

$$Y'(x,t) = \begin{cases} \mathrm{SLP}(t,9,\eta) & x = 1,2,\cdots,m; \quad \eta = 1,\cdots,m \\ \mathrm{SLP}(t,10,\eta) & x = m+1,\cdots,2m; \quad \eta = 1,\cdots,m \\ \mathrm{SLP}(t,11,\eta) & x = 2m+1,\cdots,3m; \quad \eta = 1,\cdots,m \end{cases} \tag{A1}$$

where $t = 1, 2, \cdots, n$ and $x = 1, 2, \cdots, 3m = p$. The predictand field $T'(x', t)$ is codified in the same way over its own (possibly distinct) domain of points $x' = 1, \cdots, q$. We build into the t index of $T'(x', t)$ a forecast time lead Δt so that $T'(x', t)$ comes Δt units of time later than $Y'(x, t)$, for each $t = 1, \cdots, n$. In our work, the truncated p, q are such that $q < p$. This fact allows simplifications of the algebraic theory below (similar simplifications can be made if $p \leqslant q$).

2. Prediction equations

The predictor data Y' and predictand data T' are next detrended over the span of data being used in the model building process and centered in time, e.g., $\langle T'(x', t)\rangle_t \equiv 0$ where $\langle\ \rangle_t$ denotes an average over $t = 1, \cdots, n$. We call these new datasets Y and T, respectively. Decomposing them into their truncated principal components gives

$$Y(x,t) = \sum_{j=1}^{p} \kappa_j^{1/2} \alpha_j(t) e_j(x) \quad x = 1,2,\cdots,p$$

$$T(x',t) = \sum_{j=1}^{q} \lambda_j^{1/2} \beta_j(t) f_j(x') \quad x' = 1,2,\cdots,q. \tag{A2}$$

The separate sets of eigenvectors $\mathbf{e}_j$ and $\mathbf{f}_j$ are found in the usual manner and are orthonormal. Principal component truncation rules have been employed to find p and q (Preisendorfer et al., 1981). In the following we use (p, q) to represent these truncation limits as appropriate. The variance and physical units are carried by the eigenvalues κ_j and λ_j. The principal components $\alpha_j(t)$ and $\beta_j(t)$ are evaluated as follows. For $\alpha_j(t)$

$$a_j(t) = \sum_{x=1}^{p} Y(x,t) e_j(x), \quad j = 1,\cdots,p \tag{A3}$$

and then normalize:

$$\alpha_j(t) = \kappa_j^{-1/2} a_j(t), \quad t = 1,\cdots,n.$$

Similar calculations are done for $\beta_j(t)$.

Optimal representation of $\mathbf{T}$ in terms of $\mathbf{Y}$ is obtained by first forming the set of all linear combinations of the α_j and β_j in the Euclidean vector space E_n:

$$\mathbf{u} = \sum_{j=1}^{p} \alpha_j r_j \quad \text{and} \quad \mathbf{v} = \sum_{k=1}^{q} \beta_k s_k \tag{A4}$$

where $\mathbf{r}$ and $\mathbf{s}$ are by construction arbitrary unit vectors in E_p and E_q, respectively. For each choice of $\mathbf{r}$ and $\mathbf{s}$ define the correlation

$$\langle u(t)v(t)\rangle_t = \mathbf{r}^T\mathbf{C}\mathbf{s} \tag{A5}$$

where T denotes transpose and $\mathbf{C}$ is the p by q matrix whose elements are

$$c_{jk} = \langle \alpha_j(t)\beta_k(t)\rangle_t. \tag{A6}$$

It can be shown that the correlation of $\mathbf{u}$ and $\mathbf{v}$ in (A5) is maximized if $\mathbf{r}$ and $\mathbf{s}$ are, respectively, the eigenvectors of the systems:

$$\left.\begin{aligned} [\mathbf{C}\mathbf{C}^T]\mathbf{r}_j &= \mu_j^2\mathbf{r}_j \quad j = 1,2,\cdots,p \\ [\mathbf{C}^T\mathbf{C}]\mathbf{s}_k &= \mu_k^2\mathbf{s}_k \quad k = 1,2,\cdots,q \end{aligned}\right\} \tag{A7}$$

where $\mathbf{r}_j = [r_{1j}, r_{2j}, \cdots, r_{pj}]^T$ and similarly for $\mathbf{s}_k$. Matrix theory shows that the coefficient product matrices in (A7) have the same nonzero eigenvalues (μ^2) and rank $l = \min[p, q]$ and that the $\mathbf{r}_j$ and the $\mathbf{s}_k$ form orthonormal sets of vectors in E_p and E_q, respectively. Remembering (A4), we obtain in this way the desired *canonical component vectors:*

$$\mathbf{u}_j = \sum_{i=1}^{p} \alpha_i r_{ij} \quad \text{and} \quad \mathbf{v}_k = \sum_{i=1}^{q} \beta_i s_{ik}. \tag{A8a}$$

The $\mathbf{u}_j$ and the $\mathbf{v}_k$ each form orthonormal sets of vectors in E_n.

It follows at once from (A8a) and the orthonormality of the α_j and β_k that

$$\langle \alpha_i(t)u_j(t)\rangle_t = r_{ij} \quad \text{and} \quad \langle \beta_j(t)v_k(t)\rangle_t = s_{jk}. \tag{A8b}$$

Moreover, in deriving (A7), we find, for the case $q < p$, that

$$\left.\begin{aligned} \mathbf{C}\mathbf{s}_k &= \mu_k\mathbf{r}_k \\ \mathbf{C}^T r_k &= \mu_k\mathbf{s}_k \end{aligned}\right\} \quad k = 1,\cdots,q<p. \tag{A9}$$

The μ_j are the non-negative square roots of the eigenvalues μ_j^2, $j = 1,\cdots,p$, coming from solutions of (A7). These may be arranged in descending order as follows:

$$\mu_1 = \cdots = \mu_s = 1 > \mu_{s+1} > \cdots > \mu_q > \mu_{q+1} = \cdots = \mu_p = 0 \tag{A10}$$

where $s = \max[0, p + q - (n - 1)]$. In our work the truncated p and q are such that $p + q < (n - 1)$, so $s = 0$; hence the nondegenerate eigenvalues are $\mu_1^2, \cdots, \mu_q^2$. The remainder, namely, $\mu_j^2, j = q + 1, \cdots, p$, are zero.

From (A5), (A9) and (A10), it follows that

$$\langle u_j(t)v_k(t)\rangle_t = \begin{cases} \mu_k\delta_{jk} & j,k = 1,2,\cdots,q \\ 0 & j = q+1,\cdots p;\quad k = 1,2,\cdots,q \end{cases} \tag{A11}$$

for the case $q < p$, which is the situation under study. Thus, the μ_j^2 in (A7) are seen to be the squares of the correlations between $\mathbf{u}_j$ and $\mathbf{v}_j$, and the μ_j are called canonical correlation coefficients.

The above results allow us to represent the Y and T datasets as linear combinations of their canonical component vectors:

$$\left.\begin{aligned} Y(x,t) &= \sum_{j=1}^{p} u_j(t)g_j(x) \\ T(x',t) &= \sum_{k=1}^{q} v_k(t)h_k(x') \end{aligned}\right\} \tag{A12a}$$

where we define

$$\left.\begin{aligned} g_j(x) &\equiv \langle Y(x,t)u_j(t)\rangle_t \\ h_k(x') &= \langle T(x',t)v_k(t)\rangle_t \end{aligned}\right\}. \tag{A12b}$$

The *canonical maps* $\mathbf{g}_j$ and $\mathbf{h}_k$ are vectors whose components show the correlation at a specific location (x or x') between Y or T and their respective canonical component time series (j or k). The $\mathbf{g}_j$ and $\mathbf{h}_k$ maps as they are defined are not unit vectors nor are they mutually orthogonal. It is convenient, for ease of interpretation, to deal with normalized versions of the g-maps. If these normalized maps are denoted by $\mathbf{g}'_j$, then

$$g'_j(x) = g_j(x)/\{\sum_x [g_j(x)]^2\}^{1/2}$$

so that

$$\sum_x [g'_j(x)]^2 \equiv 1.$$

Remembering that a g' map represents a set of data fields "stacked" in time [cf. (A1)], the relative predictive importance of each data field can be easily determined. For example, using (A1), the relative predictive importance (RPI) of September data is given by

$$\sum_{x=1}^{m} [g'_i(x)]^2$$

while the relative importance of November, say, is given by

$$\sum_{x=2m+1}^{3m} [g'_i(x)]^2.$$

Given the normalization of g', these fractional sums, when multiplied by 100, can be thought of as percent of predictability due to predictor data in a specific month.

The cross correlations between, say, T and the $\mathbf{u}_j$ are, by (A11) and (A12),

$$\langle T(x',t)u_j(t)\rangle_t = \begin{cases} h_j(x')\mu_j & j = 1,\cdots,q \\ 0 & j = q+1,\cdots,p \end{cases} \tag{A13}$$

for $q < p$ and illustrate the weighting supplied by the canonical correlation coefficient μ_j.

We wish to represent the n-dimensional predictand vector $\mathbf{T}(x',\cdot) = [T(x', 1),\cdots, T(x', n)]^T$ by a linear combination of the canonical component vectors $\mathbf{u}_j$ of the predictor dataset. Geometrically, this is accomplished by recalling (A11) and then projecting the T

vectors onto the q-dimensional vector space spanned by the first q of the $\mathbf{u}_j$; $j = 1, \cdots, q < p$. This is done by first forming the $n \times n$ projection matrix

$$\mathbf{P}_u = \sum_{j=1}^{q} \mathbf{u}_j \mathbf{u}_j^{\mathrm{T}}$$

which has the properties

$$\mathbf{P}_u^{\mathrm{T}} = \mathbf{P}_u \quad \text{and} \quad \mathbf{P}_u \mathbf{P}_u = \mathbf{P}_u.$$

Then the least-squares estimate of $\mathbf{T}$ by $\mathbf{Y}$ is

$$\hat{\mathbf{T}}(x', \cdot) \equiv \mathbf{P}_u \mathbf{T}(x', \cdot)$$

i.e., by (A13)

$$\hat{T}(x', t) = \sum_{j=1}^{q''} \mu_j u_j(t) h_j(x')$$

$$x' = 1, \cdots, q'' \leq q$$

$$t = 1, \cdots, n. \tag{A14}$$

As a final note, the appropriate number of $q'' \leq q$ canonical modes to retain in (A14) for the prediction $\hat{T}$ can be estimated in several ways. (a) The most obvious alternative is simply to observe the q'' value for which the global forecast skill, if significant, was a maximum. This approach has an a posteriori character, but it also does measure the *forecast* abilities of the model whereas the next two methods measure the significance of the hindcast model. (b) One could also choose to terminate the summation (A14) in a manner consistent with the method used to select p and q in (A2) (cf. Preisendorfer et al., 1981). Specifically, we select an $n \times pY$ set and an $n \times qT$ set from a population of independent, normally distributed random numbers and estimate the μ_k precisely as noted above for the "real" datasets. This is done for many realizations of (Y, T), allowing us to build a probability distribution function for each μ_k expected from a "no skill" situation, i.e., Y and T uncorrelated. The μ_k for the actual data are compared, for each k, with the distributions for the random case, and the summation terminated at q'' when the μ_q'' could with probability 0.05 have come from the random population. (c) Alternatively, we could use the same Monte Carlo procedures and the "nesting strategy" proposed by Barnett and Hasselmann (1979) on the μ_k to determine the model order. All three criteria produced essentially the same cutoff q'' values, typically 1 to 2. We chose the first method, which maximized global forecast skill, for our present study and so used that to guide our choice of q''.

3. Estimating forecast skill: The strategy

The methods of "cross validation" were used to estimate the model's true forecast skill. The approach goes as follows. Suppose we have $n + 1$ time samples of the raw Y and T fields. We refer to each of these data fields at a given time as a *data map*. For some time index ν, $1 \leq \nu \leq n + 1$, we remove the νth predictor and predictand data maps, denoting them by $Y''(x, \nu)$ and $T''(x', \nu)$. We next relabel, in a sequential fashion, the remaining n time indices of the remaining data maps. In this way we obtain at x, x', two raw training sets $\{Y'(x, t|\nu): t = 1, \cdots, n\}$ and $\{T'(x', t|\nu): t = 1, \cdots, n\}$ indexed for each ν precisely as $(\mathbf{Y}', \mathbf{T}')$ in section A1 above with the removed predictor and predictand pair indexed as $Y'(x, n + 1|\nu)$ and $T'(x', n + 1|\nu)$.

The procedures of section A2 are next followed explicitly to construct (train) the forecast model (A14). This includes finding, e.g., the present magnitudes of $\mu_j(\nu)$ and $h_i(x'|\nu)$ for the νth cross-validation case. This model is then tested on the predictor/predictand map pair that were withheld from the analysis. The procedure is as follows:

(a) Remove from $Y'(x, n + 1|\nu)$ and $T'(x', n + 1|\nu)$ the trend and means found for the νth (Y', T') set. Denote the resulting data by $Y(x, n + 1|\nu)$ and $T(x', n + 1|\nu)$.

(b) Use (A3) to estimate $\alpha_j(n + 1|\nu)$ where the required $\kappa_j(\nu)$ and $e_j(x|\nu)$ are those obtained from the current νth training set.

(c) Find the $\mathbf{r}_j$ by means of (A7) and use (A8a) to compute $u_j(n + 1|\nu)$ from $\alpha_j(n + 1|\nu)$.

(d) From (A14) and $u_j(n + 1|\nu)$ obtain the estimate $\hat{T}(x', n + 1|\nu)$ which then can be compared directly with the observed $T(x', n + 1|\nu)$.

The above process is repeated $n + 1$ times, sequentially omitting a single data map pair (Y'', T'') each time. The result is a series of $n + 1$ forecast data maps $\hat{\mathbf{T}}_\nu$, each of which can be compared directly with the observed map $\mathbf{T}_\nu$.

4. Estimating forecast skill: The method

There are numerous scoring methods that are available to test the predictive skill of (A14). The cross-validation approach gives us the opportunity to estimate true model forecast, not hindcast, performance. The skill levels were anticipated to be low. Therefore, we chose a liberal approach to scoring that is a variation of the standard tercile approach. First separate at each x' the observed values of $\mathbf{T}_\nu$ into three equally divided classes (terciles), e.g., above, normal and below categories. Next tercile the $\hat{\mathbf{T}}_\nu$ according to *its own* distribution of values at each x'.

The pairs of $(\mathbf{T}_\nu, \hat{\mathbf{T}}_\nu)$ for a specific predictand location, x', are checked to see if their respective terciles agree, e.g., "A" forecast and "A" observed. If they do, the forecast is correct. The total number correct (N_c) is used to get the percent correct and hence the *local skill* S_L (at x'):

$$S_L = 100\left(\frac{N_c}{n+1}\right). \tag{A15}$$

The significance of this number may be determined by comparison with the Stochaster's (the random fore-

caster's) binomial distribution for its number of correct forecasts. The average local skill, S_G, provides easy reference to the skill maps and is obtained by simply averaging the 33 values of S_L obtained in the forecast experiments.

In any one particular experiment there are q such measures of local skill, one for each predictand station $x' = 1, \cdots, q$. If all stations were independent of each other the binomial distribution could be used again to estimate the significance of the *global skill*, $\hat{S}_G$, where

$$\hat{S}_G = 100\left(\frac{N_s}{q}\right) \tag{A16}$$

and N_s is the number of stations that had significant S_L. When stations are correlated with each other, one must use other procedures, as pointed out by Livezey and Chen (1983). Essentially, q must be replaced by the *effective number* q_e of independent stations, and N_s must be similarly reduced, i.e., $N_s' = N_s(q_e/q)$. One may then use (q_e, N_s') in the binomial distribution to estimate the significance of $\hat{S}_G$. Here q_e would be the number of independent binomial trials and N_s' the number of successes out of those trials.

The value of q_e may be estimated from

$$q_e = q\left[1 + \left(\frac{\sum_{i=1}^{q}\sum_{j=1}^{q}{}' |c_{ij}|}{\sum_{j=1}^{q}\sigma_j^2}\right)\right]^{-1} \tag{A17}$$

where the prime denotes $i \neq j$ and where

$$c_{ij} = \langle T(x_i, t)T(x_j, t)\rangle_t$$

$$\sigma_i^2 = \langle T(x_i, t)^2\rangle_t.$$

REFERENCES

Anderson, C. W., 1984: *An Introduction to Multivariate Statistical Analysis,* 2nd ed., Wiley and Sons, 675 pp.

Barnett, T. P., 1977: An attempt to verify some theories of El Niño. *J. Phys. Oceanogr.*, **7**, 633–647.

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

——, 1981b: Statistical relations between ocean/atmosphere fluctuations in the tropical Pacific. *J. Phys. Oceanogr.*, **11**, 1043–1057.

——, 1984: Long term trends in surface temperature over the ocean. *Mon. Wea. Rev.*, **112**, 303–312.

——, 1985: Variations in near-global sea level pressure. *J. Atmos. Sci.*, **42**, 478–501.

——, 1986: Detection of changes in the global troposphere temperature field induced by greenhouse gases. *J. Geophys. Res.*, **91**(D6), 6659–6667.

——, and K. Hasselmann, 1979: Techniques of linear prediction with application to oceanic and atmospheric fields in the tropical pacific. *Rev. Geophys. Space Phys.*, **17**, 949–968.

Barnston, A. G., and R. E. Livezey, 1987: Classification, seasonality, and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea. Rev.*, **115** (in press).

Bjerknes, J., 1966: A possible response of the atmospheric Hadley circulation to equatorial anomalies of ocean temperature. *Tellus,* **18**, 820–829.

Blackmon, M. L., J. Geisler and E. Pitcher, 1983: A general circulation model study of January climate anomaly patterns associated with interannual variations of equatorial Pacific sea surface temperatures. *J. Atmos. Sci.*, **40**, 1410–1425.

Cayan, D., and A. Douglas, 1984: Urban influences on surface temperatures in the southwestern United States during recent decades. *J. Climate Appl. Meteor.*, **23**, 1520–1530.

Chervin, R., 1986: Interannual variability and seasonal climate predictability. *J. Atmos. Sci.*, **43**, 233–251.

Davis, R., 1977: Techniques for statistical analysis and prediction of geophysical fluid systems. *Geophys. Astrophys. Fluid Dyn.*, **8**, 245–277.

Diaz, H., and R. Quayle, 1978: The 1976–77 winter in the contiguous United States in comparison with past record. *Mon. Wea. Rev.*, **106**, 1393–1491.

Diaz, H. F., 1981: Eigenvector analysis of seasonal temperature, precipitation and synoptic-scale system frequency over the contiguous United States. Part II: Spring. *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.

Dickson, R. R., 1967: The climatological relationship between temperature of successive months in the United States. *J. Appl. Meteor.*, **6**, 31–38.

Dixon, K. W., and R. P. Harnack, 1986: The effect of intraseasonal circulation variability on winter temperature forecast skill. *Mon. Wea. Rev.*, **114**, 208–214.

Douglas, A., D. Cayan and J. Namias, 1982: Large-scale changes in North Pacific and North American weather patterns in recent decades. *Mon. Wea. Rev.*, **110**, 1852–1862.

Efron, B., 1983: Estimating the error rate of a prediction rule: Improvement on cross validation. *J. Amer. Stat. Assoc.*, **78**, 316–331.

Glahn, H., 1963: Canonical correlation and its relationship to discriminate analysis and multiple regression. *J. Atmos. Sci.*, **25**, 23–31.

Harnack, R. P., 1979: A further assessment of winter temperature predictions using objective methods. *Mon. Wea. Rev.*, **107**, 250–267.

——, 1982: Objective winter temperature forecasts: An update. An extension to the western United States. *Mon. Wea. Rev.*, **110**, 287–295.

——, and H. E. Lansberg, 1978: Winter season temperature outlooks by objective methods. *J. Geophys. Res.*, **83**(C7), 3601–3616.

——, and J. R. Lanzante, 1984: Specification of seasonal mean 700 mb height over North America by North Pacific and North Atlantic sea surface temperature. *Mon. Wea. Rev.*, **112**, 1626–1633.

Hasselmann, K., and T. P. Barnett, 1981: Techniques of linear prediction for systems with periodic statistics. *J. Atmos. Sci.*, **38**, 2275–2283.

Horel, J., and J. Wallace, 1981: Planetary scale atmospheric phenomena associated with the Southern Oscillation. *Mon. Wea. Rev.*, **109**, 813–829.

Hotelling, H., 1936: Relations between two sets of variates. *Biometrika,* **28**, 321–377.

Hsiung, J., and R. E. Newell, 1983: The principal nonseasonal modes of variation of global sea surface temperature. *J. Phys. Oceanogr.*, **13**, 1957–1967.

Inoue, M., and J. J. O'Brien, 1984: A forecasting model for the onset of a major El Niño. *Mon. Wea. Rev.*, **112**, 2326–2337.

Jones, P. D., T. Wigley and P. Kelly, 1982: Variations in surface air temperatures. Part I: Northern Hemisphere, 1881–1980. *Mon. Wea. Rev.*, **110**, 59–70.

Lau, N. C., 1985: Modeling the seasonal dependence of the atmospheric response to observed El Niños in 1962–76. *Mon. Wea. Rev.*, **13**, 1970–1996.

Lawley, D., 1959: Tests of significance in canonical analysis. *Biometrika,* **46**, 59–66.

Livezey, R., and W. Chen, 1983: Statistical field significance and its

determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **111,** 46–59.

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

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

Namias, J., 1975: Short period climatic variations. *Collected Works,* Vols. I–III, University of California Press. [Available from J. Namias, Scripps Institution of Oceanography, La Jolla, CA 92093.]

——, 1978: Persistence of U.S. seasonal temperatures up to one year. *Mon. Wea. Rev.*, **106,** 1557–1567.

Nicholls, N., 1987: The use of canonical correlation to study teleconnections. *Mon. Wea. Rev.*, **115** (in press).

Preisendorfer, R., and C. Mobley, 1984: Climate forecast verifications, U.S. mainland 1974–1983. *Mon. Wea. Rev.*, **112,** 809–825.

——, F. Zwiers and T. P. Barnett, 1981: Foundations of principal component selection rules. SIO Ref. Ser. 81-4. [Available from Scripps Institution of Oceanography, La Jolla, CA 92093.]

Rowntree, P. R., 1972: The influence of tropical east Pacific Ocean temperatures on the atmosphere. *Quart. J. Roy. Meteor. Soc.*, **98,** 290–321.

Shukla, J., and J. N. Wallace, 1983: Numerical simulation of the atmospheric response to equatorial sea surface temperature anomalies. *J. Atmos. Sci.*, **40,** 1613–1630.

Stone, M., 1974: Cross validatory choice and assessments of physical predictions. *J. Roy. Stat. Soc.*, **B36,** 111–147.

——, 1977: Estimatonics for and against cross validation. *Biometrika,* **64,** 29–38.

Trenberth, K. E., 1984a: Some effects of finite sample size and persistence on meteorological statistics. Part I: Autocorrelations. *Mon. Wea. Rev.*, **112,** 2359–2368.

——, 1984b: Some effects of finite sample size and persistence on meteorological statistics. Part II: Potential predictability. *Mon. Wea. Rev.*, **112,** 2369–2378.

——, and D. A. Paolino, 1980: The Northern Hemisphere sea-level pressure data set: Trends, errors and discontinuities. *Mon. Wea. Rev.*, **108,** 855–872.

Tukey, J., 1958: Bias and confidence in not-quite large samples. *Ann. Nath. Stat.*, **29,** p. 614.

van den Dool, H., W. Klein and J. Walsh, 1986: Geographic distribution and seasonality of persistence and monthly mean air temperatures over the United States. *Mon. Wea. Rev.*, **114,** 546–560.

van Loon, H., and J. Williams, 1976: The connection between trends of mean temperature and circulation at the surface. Part II: Summer. *Mon. Wea. Rev.*, **104,** 1003–1011.

Walker, G., and E. Bliss, 1932: World weather V. *Mem. Roy. Meteor. Soc.*, **IV**(36), 53–84.

Wallace, J. M., and D. Gutzler, 1981: Teleconnections in geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.*, **109,** 784–812.

Weare, B. C., A. R. Navato and R. E. Newell, 1976: Empirical orthogonal analysis of Pacific sea surface temperatures. *J. Phys. Oceanogr.*, **6,** 671–678.