Statistical Significance of Long-Range "Optimal Climate Normal" Temperature and Precipitation Forecasts

DANIEL S. WILKS

Department of Soil, Crop and Atmospheric Sciences Cornell University, Ithaca, New York

(Manuscript received 3 May 1995, in final form 18 September 1995)

ABSTRACT

A simple approach to long-range forecasting of monthly or seasonal quantities is as the average of observations over some number of the most recent years. Finding this "optimal climate normal" (OCN) involves examining the relationships between the observed variable and averages of its values over the previous one to 30 years and selecting the averaging period yielding the best results. This procedure involves a multiplicity of comparisons, which will lead to misleadingly positive results for developmental data. The statistical significance of these OCNs are assessed here using a resampling procedure, in which time series of U.S. Climate Division data are repeatedly shuffled to produce statistical distributions of forecast performance measures, under the null hypothesis that the OCNs exhibit no predictive skill. Substantial areas in the United States are found for which forecast performance appears to be significantly better than would occur by chance.

Another complication in the assessment of the statistical significance of the OCNs derives from the spatial correlation exhibited by the data. Because of this correlation, instances of Type I errors (false rejections of local null hypotheses) will tend to occur with spatial coherency and accordingly have the potential to be confused with regions for which there may be real predictability. The "field significance" of the collections of local tests is also assessed here by simultaneously and coherently shuffling the time series for the Climate Divisions. Areas exhibiting significant local tests are large enough to conclude that seasonal OCN temperature forecasts exhibit significant skill over parts of the United States for all seasons except SON, OND, and NDJ, and that seasonal OCN precipitation forecasts are significantly skillful only in the fall. Statistical significance is weaker for monthly than for seasonal OCN temperature forecasts, and the monthly OCN precipitation forecasts do not exhibit significant predictive skill.

1. Introduction

Very long range forecasting—with lead times on the order of one year—is a difficult problem, even when the predictand is an average over one month or season. The accuracy of such forecasts is currently quite limited, and accordingly, relatively unsophisticated methods can be competitive. A very simple and operationally straightforward method for constructing such forecasts is as the average over that number of the most recent years' values found to give the best "hindcast" performance on a set of historical data. Forecasts prepared according to this method have lead times of approximately 11 months when the predictand is a monthly average or nine months when the predictand is a three-month average.

Searching the climatic record at a location for that averaging period yielding the best hindcast performance is not a new idea. Some relatively recent discussions of the screening of different averaging periods as long-range predictors are provided in Dixon and Shulman (1984), Lamb and Changnon (1981), and Sabin and Shulman (1985). Generally, the characteristics of these revised climate normals are assessed in relation to the conventional 30-yr normal (e.g., Kunkel and Court 1990). A variety of earlier studies addressing this issue, some dating to the first half of the century, are also cited in the above papers.

This simple averaging procedure is currently used as one contributor to new operational long-lead forecasts produced by the U.S. Climate Prediction Center (Huang et al. 1996, hereafter HVB), who call the average over that number of years yielding the best performance the "optimal climate normal" (OCN). Operating on area-averaged (U.S. Climate Division) data for seasonal (i.e., running averages of three months) temperature, HVB selected averaging periods that yielded best hindcast performance for the years 1961–1993. Analogous procedures underlie the operational OCN forecasts of seasonal precipitation and of monthly temperature and precipitation.

While the procedure for computing the OCN forecasts is operationally straightforward, an objective evaluation of the resulting forecasts is conceptually more difficult. At individual locations, the procedure of

Corresponding author address: D. S. Wilks, Department of Soil, Crop and Atmospheric Sciences, Cornell University, Ithaca, NY 14853-1901.
E-mail: dsw5@cornell.edu

blindly screening the hindcast performances of all averaging periods between one and 30 years involves a multiplicity of comparisons. Choosing the single best averaging period from many possibilities should accordingly lead to some degree of "artificial skill" or misleadingly positive results for the developmental data. While a cross-validation approach would be a desirable safeguard, the procedure by which it might be constructed is not clear due to the overlapping of the averaging periods and the shortness of the available record relative to the 30-yr maximum averaging period. Recognizing this problem, HVB used a deflator due to Barnston and Van den Dool (1993) to estimate forecast performance on independent data.

An equally serious problem relates to the multiplicity of simultaneous evaluations of hindcast performance at many locations. Because of the strong spatial correlation typically exhibited by monthly and seasonal temperature and precipitation data, the apparently (but erroneously) good hindcast performances occasionally arising by chance will tend to occur with spatial coherency, and these areas accordingly have the potential to be confused with regions for which there may be real predictability. This aspect of the problem is especially of concern because the selection of averaging periods at the individual locations is, by the nature of the searching procedure, "tuned" to the sampling variations particular to the available data records.

This study approaches both the local and spatial OCN multiplicity problems through nonparametric hypothesis testing procedures. Seasonal and monthly temperature and precipitation data from United States climate divisions for the years 1931–1993, as in HVB and described in section 2, are used in order to allow valid comparisons with that study. The statistical significance of the OCNs for single-location dependent data specifications is assessed using a resampling procedure, described in section 3, in which individual data series are repeatedly shuffled and statistical distributions of various performance measures are produced, under the null hypothesis that the OCNs exhibit no predictive skill relative to the conventional 30-yr normals. The "field significance" of the collections of local tests described above is similarly assessed in section 4 by simultaneously and coherently shuffling the time series for all the climate divisions and summarizing the results of individual-location tests under this sampling scheme. Section 5 explores the possible physical basis of the OCN skill, and section 6 presents the results of hypothesis tests consistent with operational constraints that have emerged in connection with use of the OCN forecasts.

2. Data, forecasts, and verification measures

The U.S. Climate Division data, for 1931–1993 (National Climatic Data Center 1994) are used in the following in order to maintain maximum comparability with HVB. Monthly data for the 344 divisions in the 48 conterminous states were obtained from the National Climatic Data Center.

Analyses of both the raw monthly and composited seasonal data series are reported below. Monthly series for temperature and precipitation for each calendar month were used directly. For each "season," comprised of triplets of consecutive months (January–February–March, February–March–April, etc.), the respective monthly temperatures for each division were averaged, and the monthly precipitation values were totaled. The result in each case is a time series, T_i, $i = 1, \cdots, 63$; in which $i = 1$ corresponds to the year 1931, and $i = 63$ indicates the year 1993. This notation will be applied to both the temperature and the precipitation data.

The k-year "climate normal" predictor for the data value T_i is simply the average over the previous k values in that time series, that is

$$\bar{T}_{i,k} = \frac{1}{k} \sum_{j=1}^{k} T_{i-j}. \tag{1}$$

The averaging period is allowed to vary between $k = 1$ (the "persistence" forecast) and $k = 30$ (the annually updated "30-yr normal"). Accordingly, the first year for which hindcasts can be computed for the present dataset is 1961 ($i = 31$). For each data series, the averaging period for which (1) produces the best specifications for the years 1961–1993 is chosen as "optimal."

The correspondence between the OCNs and data for the years being hindcast was judged in HVB primarily using correlation measures of the form

$$\mathrm{COR}(k) = \frac{\sum_{i=31}^{n} \hat{T}_i^f \hat{T}_i^{ob}}{\left[\sum_{i=31}^{n} (\hat{T}_i^f)^2 \sum_{i=31}^{n} (\hat{T}_i^{ob})^2 \right]^{1/2}}, \tag{2}$$

where $n = 63$ years for the present dataset, the superscript "f" denotes forecast, and the superscript "ob" indicates the target observed datum to be specified. Larger values of $\mathrm{COR}(k)$ indicate better forecasts. The circumflex accents indicate anomalies constructed as

$$\hat{T}_i^f = \bar{T}_{i,k} - \langle T_i^f \rangle \tag{3a}$$

and

$$\hat{T}_i^{ob} = T_i - \langle T_i^{ob} \rangle. \tag{3b}$$

Different choices for the averages $\langle T_i \rangle$ yield somewhat different correlation scores in (2). Two such averages were used in HVB. The first of these is defined by

$$\langle T_i^f \rangle = \langle T_i^{ob} \rangle = \begin{cases} \frac{1}{30}\sum_{j=1}^{30} T_j, & 30 < i \leq 43 \\ \frac{1}{30}\sum_{j=11}^{40} T_j, & 43 < i \leq 53 \\ \frac{1}{30}\sum_{j=21}^{50} T_j, & 53 < i \leq 63\,, \end{cases} \quad (4)$$

which simulate the use of "aging WMO normals," or 30-yr averages updated decadally (e.g., 1951–1980) and available a few years following the end of the averaging period. Denote as $\mathrm{COR}_1(k)$ the agreement measure in (2), when (4) is used to define the anomalies.

The second measure used in HVB uses the annually updated 30-yr mean

$$\langle T_i^f \rangle = \langle T_i^{ob} \rangle = \bar{T}_{i,30}. \quad (5)$$

This is simply the average of the 30 values previous to the forecast value, as given in (1). Denote as $\mathrm{COR}_2(k)$ the agreement measure in (2), when (5) is used to define the anomalies.

The measures $\mathrm{COR}_1(k)$ and $\mathrm{COR}_2(k)$ are "anomaly correlations," in that the data values are approximately centered by subtraction of an externally derived average. These are somewhat different from the conventional Pearson product-moment (i.e., ordinary linear) correlation coefficient, which is obtained from (2) using

$$\langle T_i^f \rangle = \frac{1}{n-31}\sum_{i=31}^{n} \bar{T}_{i,k} \quad (6a)$$

and

$$\langle T_i^{ob} \rangle = \frac{1}{n-31}\sum_{i=31}^{n} T_i. \quad (6b)$$

Also considered here is the normalized mean-squared error,

$$\mathrm{MSE}^*(k) = \frac{\sum_{i=31}^{n} (\bar{T}_{i,k} - T_i)^2}{\sum_{i=31}^{n} (\bar{T}_{i,30} - T_i)^2}. \quad (7)$$

The numerator in (7) is proportional to the forecast mean-squared error. The normalizing factor in the denominator of (7), which is proportional to the mean-squared error for the annually updated 30-yr normal, is included to allow $\mathrm{MSE}^*(k)$ values for different locations to be comparable, regardless of their intrinsic degrees of interannual variability. Smaller values of $\mathrm{MSE}^*(k)$ indicate better forecasts.

3. Local statistical significance

a. Concepts

Determining the "optimal" value of the averaging period, k, involves examining the accuracy with which the data values for each station can be specified using (1), as reflected by measures such as (2) or (7), for each of the 30 values of k considered. That value of the averaging period yielding the best hindcast performance among these 30 values on the available data is chosen.

Although this procedure is intuitively appealing, the results are difficult to interpret because of the many trial values that are considered. Even if there is little real relationship between the predictor in (1) and the data values T_i, there may be ample opportunity for chance variations in the relatively short data series to yield a value of k for which a chosen accuracy measure indicates appreciable predictability. That is, it is the value of k yielding the (apparently) most accurate among many hindcasts being chosen, rather than the accuracy of forecasts based on an independently chosen averaging period being evaluated. The consequence of this multiplicity of comparisons is that some degree of "artificial skill," or misleadingly positive results for the developmental data, is expected. In the worst case, predictability may be inferred as an artifact of the exhaustive fitting procedure when none really exists.

A usual remedy for dependent-sampling problems of this kind is to conduct a cross-validation analysis (e.g., Efron 1982; Elsner and Schmertmann 1994; Wilks 1995). This procedure involves repeating the forecast-fitting exercise many times, each with a different subset of the developmental data (often a single point) withheld. The forecast algorithm is thus not "tuned" to the withheld points, and the accuracy with which these points can be forecast is then evaluated. In the present problem, however, the averaging periods overlap, and the length of the data series (63 years) is short with respect to the maximum averaging period (30 years), so that it is not clear how to design a cross-validation procedure having an adequate sample size.

Recognizing this local multiplicity problem, HVB deflated the dependent-sample correlation measures $\mathrm{COR}_1(k)$ and $\mathrm{COR}_2(k)$ using (Barnston and Van den Dool 1993)

$$\mathrm{COR}'(k) = \frac{N'\,\mathrm{COR}^2(k) - 1}{(N'-1)\,\mathrm{COR}(k)}, \quad \mathrm{COR}(k) > (N')^{-1/2}, \quad (8)$$

where $N' = 20$ (rather than $N = 33$, for specifications of the years 1961–1993) was used as an estimate of the "effective independent sample size." Only OCN forecasts exhibiting $\mathrm{COR}'(k) > 0.30$, corresponding to $\mathrm{COR}_1(k)$ or $\mathrm{COR}_2(k) > 0.408$, were then regarded as reflecting real predictability.

An alternative but complementary approach is to view the problem in the context of an hypothesis test: under the null hypothesis that (1) provides no better information regarding the data values T_i than a 30-yr climatological average, what is the probability that results equal to or stronger than those observed for a

TABLE 1. Critical values (10%, 5%, and 1% levels) for the performance statistics in (2) and (7) for underlying data following Gaussian distributions and $n = 63$, when (a) the best averaging period (k) among 30 is chosen and (b) the averaging period is chosen randomly. The critical levels in (a) are appropriate to the OCN forecasts, and the cutoff value of 0.408 adopted by HVB is very close to the 5% critical value for $COR_2(k)$. The differences between corresponding values in (a) and (b) illustrate the impact of searching among many choices for the "best" predictor, even under the null hypothesis of no real relationship.

(a)				(b)			
Measure	10%	5%	1%	Measure	10%	5%	1%
$COR_1(k)$	0.374	0.425	0.517	COR_1	0.258	0.318	0.428
$COR_2(k)$	0.364	0.404	0.484	COR_2	0.198	0.254	0.354
$MSE^*(k)$	0.928	0.901	0.828	MSE^*	0.980	0.960	0.914

given climate division could have arisen by chance? If that probability is sufficiently small (perhaps 0.05 or less), the null hypothesis is rejected and one can conclude that the OCN forecast exhibits statistically significant predictive accuracy.

The observed values of the performance statistics in (2) and (7), as the best of 30 trials, are complicated functions of the data T_i. Accordingly, the functional form of their sampling distributions under the null hypothesis (called the null distribution) is not known, and this precludes use of conventional parametric statistical tests. A sound and practical alternative is to evaluate the significance of the OCN predictors using a resampling test (e.g., Mielke et al. 1981; Wilks 1995). Resampling tests operate on the set of available data by simulating the data-generation process and repeatedly computing the test statistic [Eqs. (2) or (7)] under different permutations of the observations consistent with the null hypothesis. The collection of these values provides a reference distribution consistent with the null hypothesis, against which the actually observed value of the test statistic can be compared for unusualness.

In the present problem, any significant predictive skill (with respect to the 30-yr normals) exhibited by the OCN forecasts presumably derives from the data series T_i exhibiting at least somewhat consistent trends or cycles in time. Either cycles or pseudocycles in the data with periods shorter than 30 years, or consistent trends with periods longer than the sample size, could lead to OCN forecasts exhibiting significant skill with respect to the 30-yr normals. In either case, the null hypothesis of no predictive skill implies that each 63-member data series exhibits no real time dependence, and to the extent that the data appear to contradict this condition it is as a result of chance orderings. Therefore, according to the null hypothesis, each observed data series is but one of the $63! \approx 1.9826 \times 10^{87}$ equally likely possible orderings of the 63 data values. The reference distribution for any of the performance measures can be constructed by shuffling the T_i values sufficiently many times to sample adequately the 63! possibilities, computing the value of the test statistic for each of the 30 values of k in each random reordering, and recording the largest [or smallest, for $MSE^*(k)$] of each set of 30 test statistics. For the shuffled data, the best "predictions" overall for the resampled trials should be achieved by the longest averaging period, in this case 30 years [compare Monte-Carlo results in Dixon and Shulman (1984) and Sabin and Shulman (1985)]. If the observed value of the test statistic (best over $k = 1, \cdots, 30$; in the unshuffled data series) is larger [or smaller, for $MSE^*(k)$] than all but a small fraction of the values in the reference distribution, then the null hypothesis can be rejected at a level consistent with that fraction.

b. *Temperature series*

For the temperature data, it is not actually necessary to repeat the resampling process for each of the 344 climate divisions and each of the 12 months or seasons. The distributions of monthly and seasonal temperatures are very nearly Gaussian, which is to be expected according to the central limit theorem, since these are averages of approximately 60 (for monthly values) or 180 (for seasonal values) individual daily maximum and minimum temperature observations. In addition, the performance measures in (2) and (7) are nondimensional, so that their behavior under the null distribution for Gaussian variates can be simulated once and for all using synthetically generated standard ($\mu = 0$, $\sigma^2 = 1$) Gaussian random numbers.

Table 1a shows the 10%, 5%, and 1% critical values for the null distributions of $COR_1(k)$, $COR_2(k)$, and $MSE^*(k)$ computed using 10 000 separate samples of $n = 63$ independent standard Gaussian variates. The values in Table 1a agree quite closely with those obtained through the resampling of actual seasonal temperature series of the same length. The critical values for $COR_1(k)$ and $COR_2(k)$ are similar but not identical, with consistently larger values for $COR_1(k)$ being required to reject the null hypothesis at a given level. Table 1a indicates that the cutoff value of 0.408 adopted by HVB [Eq. (8)] is very close to the 5% critical value for $COR_2(k)$. The critical values for the correlation measures increase for more stringent rejection levels, but those for $MSE^*(k)$ decrease because smaller values for this statistic indicate better forecasts.

The critical values for MSE*(k) indicate, for example, that in 10%, 5%, and 1% of the 10 000 random series, a value of k could be found that yielded MSE as small or smaller than the MSE for $k = 30$ multiplied by the factors 0.928, 0.901, and 0.828, respectively.

The values in Table 1a are appropriate for evaluation of the statistical significance of the OCN temperature forecasts, for $n = 63$. It is interesting to compare these to the values in Table 1b, which were produced in the same way, except for the crucial difference that the 30 possible values of k were not searched for the one yielding the best specifications. Rather, for each of the 10 000 trials, a single value of k was chosen randomly. As the best predictors from sets of one, rather than 30 possibilities, the values comprising the distributions summarized in Table 1b are necessarily smaller (larger, for MSE*) than the corresponding results in Table 1a. The differences between the corresponding entries in Tables 1a and 1b illustrate the potential danger involved in searching a wide range of possibilities when choosing an "optimal" averaging period: even if there is no real relationship in the data, a few of the comparisons are likely to yield apparently good results by chance. Focusing on these best values in the dependent dataset yields exaggerated and rosy results.

Fuller pictures of the null distributions for the statistics $COR_2(k)$ and MSE*(k) are presented in Fig. 1. Figure 1a shows the upper half of null distribution for $COR_2(k)$ for sample sizes $n = 40$, 50, 60, 80, 100, and 200. Figure 1b shows the lower half of the null distribution of MSE*(k), for these same sample sizes. The dashed horizontal lines in these panels indicate the 10%, 5%, and 1% critical levels. For example, Fig. 1b indicates that approximately 2% of the 10 000 synthetic independent Gaussian data series with $n = 40$ yielded values of MSE*(k) smaller than 0.6, that about 5% of these values were smaller than 0.7, and that 10% were smaller than approximately 0.775. That is, for data series with $n = 40$ [i.e., only $n - 31 = 9$ terms in the numerator and denominator of Eq. (7)], an observed value of MSE*(k) would need to be no larger than 0.7 for it to be declared statistically significant at the 5% level, but for $n = 50$, MSE*(k) as large as about 0.85 would be significant at that level. Figure 1 reflects the inflation arising from the searching of each of the data series for the value of k yielding the best results. Corresponding results for k being chosen either a priori or randomly (not shown) would be to the left of the corresponding curves in Fig. 1a and to the right of the curves shown in Fig. 1b (compare Table 1b).

Figure 2 shows the spatial distribution of locally significant tests at the 344 climate divisions for the (a) MAM, (b) JJA, (c) SON, and (d) DJF seasonal temperature data using the MSE*(k) performance criterion. Large circles, pluses, and heavier dots indicate significance at the 1%, 5%, and 10% levels, respectively. For JJA and DJF (Figs. 2b and 2d), large areas are significant at the 1% and 5% levels in the eastern one-third of the United States, and the same is true of the western portion of the country in MAM (Fig. 2a). These are fairly convincing indications that, in these seasons at least, the OCNs perform better than would be expected by chance alone. For SON (Fig. 2c), there is a group of divisions in the upper midwest that show local significance at the 5% level, and a few scattered stations also show significant results. Because of the multiplicity of (spatially) correlated tests considered here, that such a pattern would not have arisen by

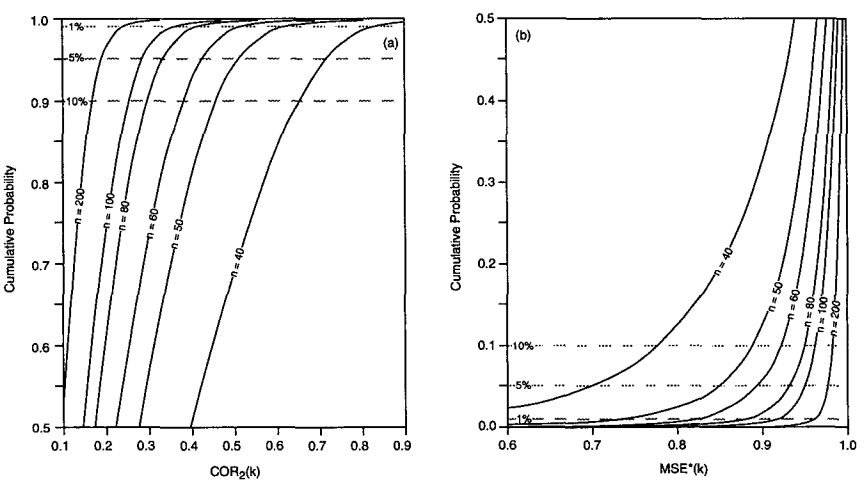

FIG. 1. Upper and lower halves, respectively, of the null distributions of (a) $COR_2(k)$ and (b) MSE*(k) for selected data series lengths. The upper tail of the $COR_2(k)$ distribution is shown because larger values of this statistic indicate better forecasts, and the lower tail of the distribution of MSE*(k) is shown because smaller values of this statistic are better. Dashed lines indicate 10%, 5%, and 1% significance levels.

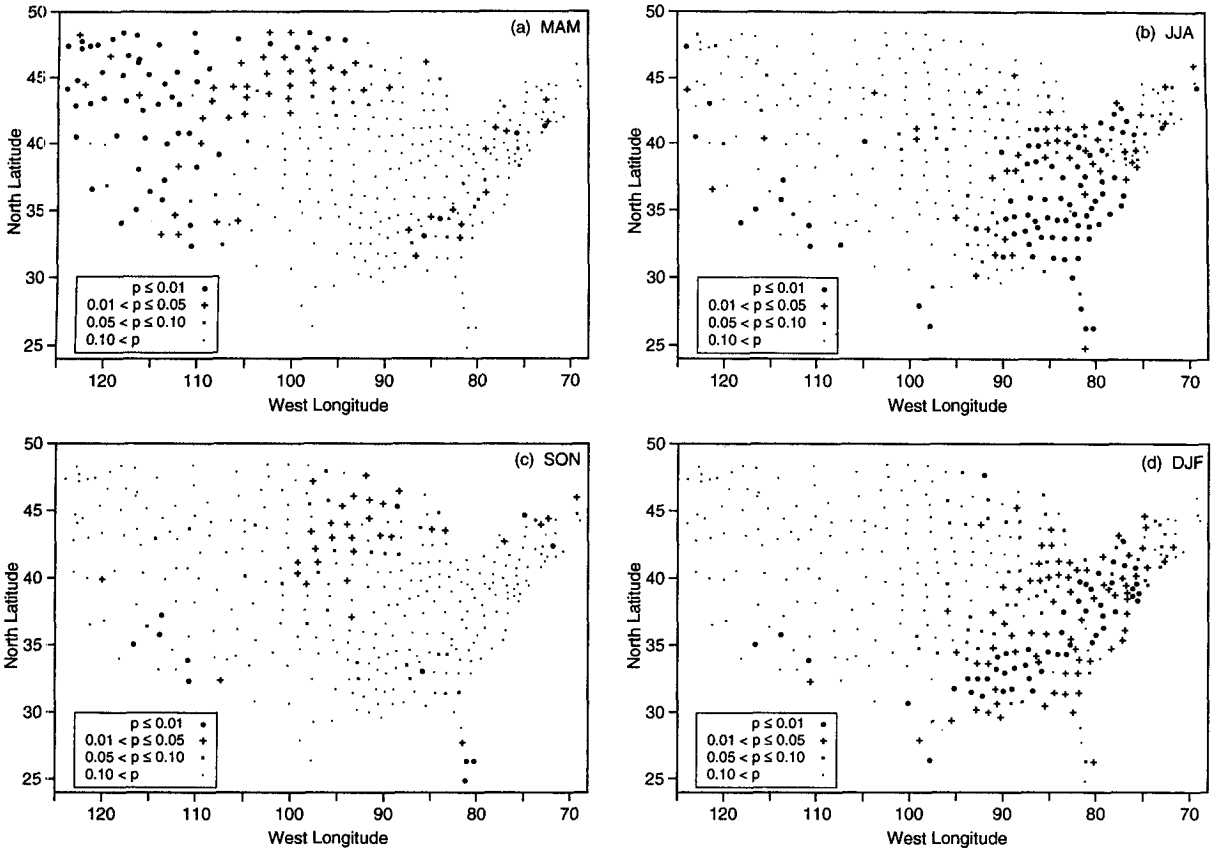

FIG. 2. Spatial distributions of the statistical significance of local OCNs for 1961–1993, according to the critical values for MSE*(k) in Table 1a, for seasonal U.S. Climate Division temperature data. (a) March–April–May, (b) June–July–August, (c) September–October–November, (d) December–January–February.

chance is less clear. This question will be investigated in section 4.

The spatial distributions of locally significant tests shown in Fig. 2 are very similar to, although not exactly the same as, those which result from use of the $COR_1(k)$ and $COR_2(k)$ criteria (not shown). These maps also resemble qualitatively the spatial distributions of the $COR_1(k)$ statistic presented in HVB.

c. *Precipitation series*

The Gaussian critical levels presented in Table 1a and Fig. 1 are appropriate to the divisional precipitation data only in those cases where the statistical distributions of the data values are not excessively skewed. Critical levels for resampling tests based on individual divisional precipitation data series, as described in section 3a, agree well with those in Table 1a for those divisions where gamma distributions fit to the data values exhibit shape parameters larger than about five. For some of the divisions yielding shape parameters smaller than this value, however, the appropriate critical values are substantially larger, and the differences are larger for the more stringent test levels. For these divisions, use of the values in Table 1a would lead to unjustified rejection of local null hypotheses, leading to false confidence in the results. Substantial areas of the United States exhibit seasonal precipitation totals that are sufficiently skewed to be fit by gamma distributions with shape parameters smaller than five (Ropelewski and Jalickee 1983), and this is true for monthly precipitation over even larger areas (Wilks and Eggleston 1992). Therefore, all significance tests for precipitation series presented here are based on individual divisional resampling procedures, rather than critical levels derived from Table 1a or Fig. 1.

Figure 3 shows the spatial distributions of locally significant tests for divisional precipitation totals over (a) MAM, (b) JJA, (c) SON, and (d) DJF, again for the MSE*(k) criterion. In contrast to the corresponding results for seasonal temperatures (Fig. 2) there are relatively few divisions for which seasonal OCN forecasts of precipitation are sufficiently accurate for their performance to be regarded as not having arisen by chance. The strongest result is for SON (Fig. 3c), in which divisions around the Great Lakes through a por-

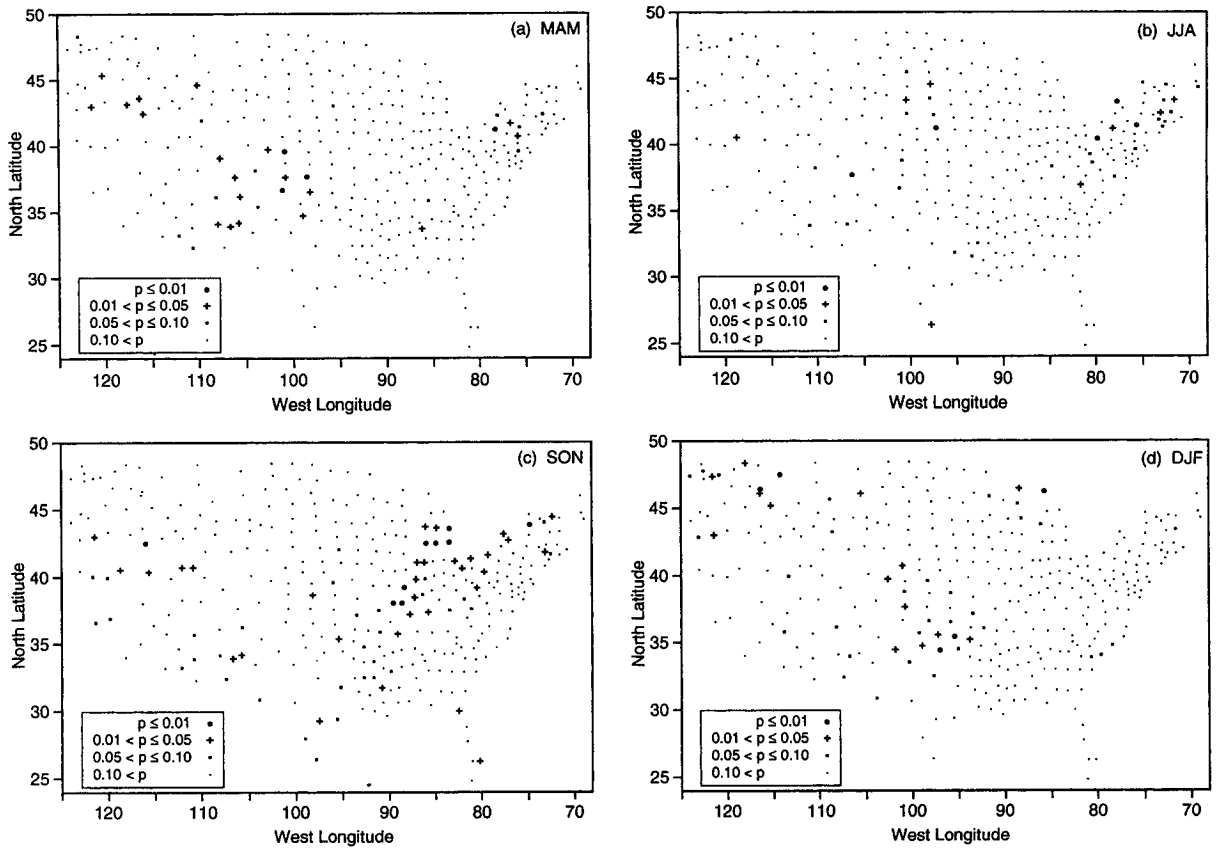

FIG. 3. As Fig. 2, for seasonal precipitation. The critical levels have been taken from individual division by division resampling tests because the precipitation data at drier locations are strongly non-Gaussian.

tion of the Midwest appear to show a coherent pattern of statistical significance at the 5% and 1% levels. The extent to which these patterns of local significance may be regarded as significant in a larger sense is examined in the next section.

4. Field significance

Section 3 described evaluation of the statistical significance of local OCN forecasts for the 344 United States climate divisions. In order to fully evaluate the promise of the OCN forecasts, it is necessary to evaluate the statistical significance of the overall pattern of forecast performance, that is, the "field significance" (Livezey and Chen 1983) of patterns such as those shown in Figs. 2 and 3.

This assessment of the overall significance of the OCNs jointly at all 344 climate divisions is complicated by two factors. The first derives simply from the large numbers of individual tests being evaluated simultaneously. Even if the forecasts were not better than the 30-yr average at any of the 344 climate divisions, some fraction of these many individual tests would yield nominally significant results purely by chance. If the null hypothesis is true, any given test will yield a falsely significant result (type I error) at the 5% level, say, with probability 0.05. For independent tests, it is straightforward to deal with this problem using the binomial distribution (Livezey and Chen 1983; von Storch 1982; Wilks 1995). For example, in a hypothetical collection of 20 independent tests for which the null hypothesis is really true, one expects (in the statistical sense) one false significant result at the 5% level. However, the probability is greater than 0.25 that there will be at least two such results among 20 independent tests.

The second complication is more subtle and derives from the spatial correlation of the divisional temperature and precipitation data. Because of these correlations, false rejections of local null hypotheses will tend to occur with spatial coherency and will accordingly have the potential to be confused with regions for which there may be real predictability. The multiplicity problem for correlated tests is more difficult to deal with and is in general approached through construction of an appropriate resampling test (Livezey and Chen 1983).

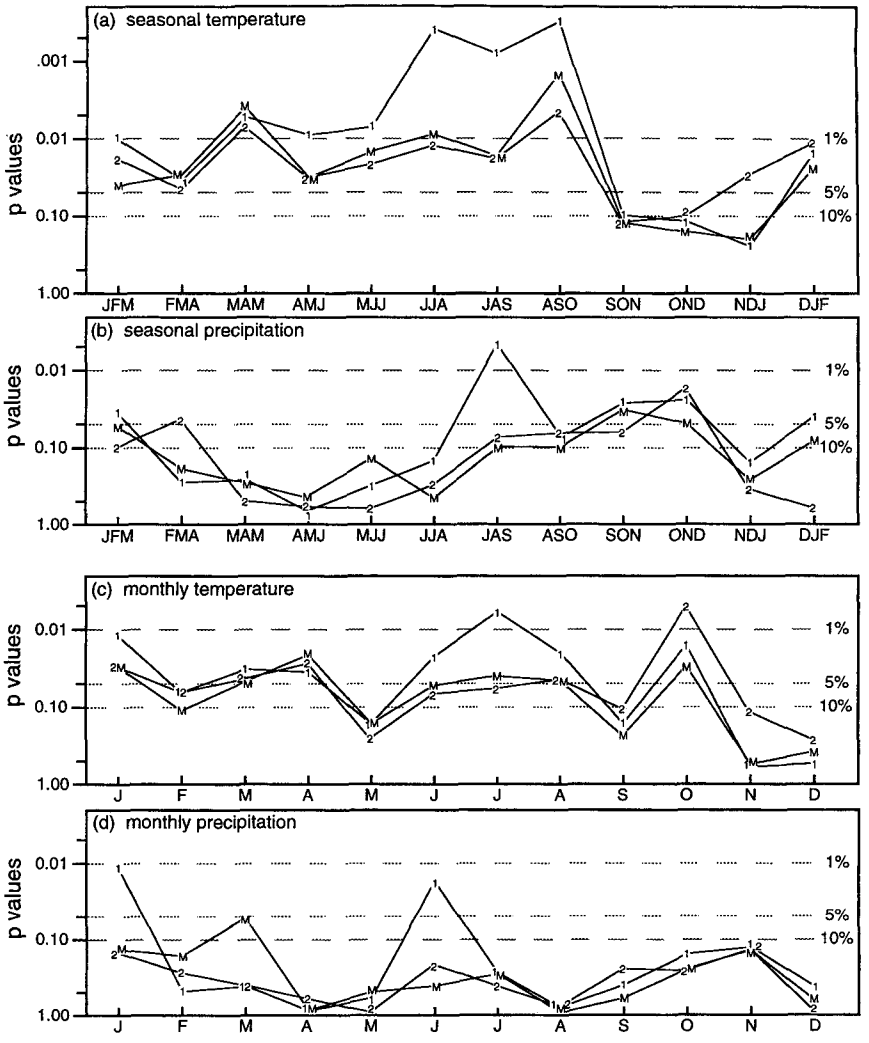

FIG. 4. *P*-values summarizing field significance tests for OCN forecasts of (a) seasonal temperature, (b) seasonal precipitation, (c) monthly temperature, and (d) monthly precipitation. Local tests for temperatures were conducted at the 5% level, and local precipitation tests were at the 10% level. The symbols "1," "2," and "*M*" indicate tests using the $COR_1(k)$, $COR_2(k)$, and MSE*(k) statistics, respectively.

Here the field significance of the collections of sets of local tests described above is assessed by simultaneously and coherently shuffling the individual time series for all the climate divisions. That is, for this resampling test the data consist of time series of 344-element vectors $\mathbf{T}_i$, $i = 1, \cdots, 63$. Sampling from the 63! permutations of these vectors simulates the condition implied by the null hypothesis that the particular time ordering of the data is arbitrary, while preserving the spatial relationships of the data in each year. Here 10 000 permutations of the vectors $\mathbf{T}_i$ were drawn, and local tests were conducted on these samples for each of the 344 divisions as described in section 3. The null distribution then consists of the areas (not numbers of tests, because sizes of the climate divisions vary widely) over which nominally significant tests occurred over the 10 000 trials. The statistical significance of the actual area over which locally significant tests occur is then assessed by comparison to the corresponding distribution of areas of (falsely) significant local tests under the null hypothesis.

Figure 4 summarizes the results of these field significance tests for forecasts of (a) seasonal temperature, (b) seasonal precipitation, (c) monthly temperature, and (d) monthly precipitation. The local tests upon which these results are based were conducted at the 5% level for temperatures (Figs. 4a and 4c) but at the 10% level for precipitation (Figs. 4b and 4d) since relatively few local precipitation tests were significant at the 5% level. The vertical scale for the field signifi-

cance p values are inverted and logarithmic, and horizontal dashed lines indicate field significance at the 1%, 5%, and 10% levels. The symbols "1," "2," and "M" indicate tests using the $COR_1(k)$, $COR_2(k)$, and MSE*(k) statistics, respectively. The results for tests involving these three measures are in general agreement, although $COR_1(k)$ shows a tendency to yield apparently more significant results in some cases.

Figure 4a indicates that the OCN forecasts for seasonal temperature are significantly better than using the 30-yr average, except for the seasons September–October–November, October–November–December, and November–December–January. The areas in Figs. 2a, 2b, and 2d over which local tests were significant at the 5% level (symbols ● and +) are large enough that they are very unlikely to have occurred by chance. On the other hand, the result for September–October–November (Fig. 2c) shows an area for tests significant at the 5% level that is smaller than more than 10% of the results obtained from the shuffled temperature fields.

The results for seasonal precipitation in Fig. 4b are much weaker, with the most consistent statistical significance occurring only for September–October–November, and October–November–December. Comparing with Fig. 3, the areas over which local tests were significant at the 10% level (symbols ●, +, and •) were larger than those in Fig. 3c (for SON) in fewer than 5% of the results obtained from the shuffled precipitation fields. By contrast, the areas over which significant local tests occurred in Figs. 3a, 3b, and 3d were not atypical of those achieved using the 10 000 shuffled precipitation fields.

Figures 4c and 4d show that the statistical significance of the OCN monthly forecasts is considerably weaker. For temperature (Fig. 4c), the strongest results are for January, April, July, August, and October, although few of the tests achieve field significance at the 1% level. The results for monthly precipitation (Fig. 4d) are worst of all and indicate that performance of the OCNs is essentially indistinguishable from the results on the randomly reordered data. Evidently the OCN forecasts for monthly precipitation have no significant predictive value.

5. Possible physical basis of skill

It was noted in section 3a that OCN forecasts could exhibit skill relative to conventional 30-yr normals either if the data exhibited short-term cyclic or pseudocyclic behavior or if the observed data record were part of a longer-term (perhaps century-scale) consistent trend. In the former case, averaging lengths of some fraction of the period of the underlying cycle would capture much of the available signal. The latter case could result from a gradual warming or cooling trend, and the optimal averaging period would reflect a compromise between the strength of the trend (the most recent years should be the most representative of the next year) and the level of random variation around the trend (a longer averaging period would be required to smooth the noise around the trend).

The results of a preliminary look at this question are presented in Fig. 5. For the four standard climatological seasons, linear regressions were fit to each of the 344 divisional temperature series using the year as the independent variable. The data used in these regressions were the 33 years 1961–1993 *plus* the number of years specified as the OCN averaging period for each division according to the $COR_2(k)$ criterion. That is, the regression for each division was computed over the $n = 33 + k$ years beginning in 1961-k and ending in 1993. The vertical axes in the panels of Fig. 5 are the slope estimates for these regressions divided by the standard errors of those estimates, that is, the t-ratios. The t-ratios thus express the trade-off between trend strength (absolute value of the slope) and the scatter of points around the slope since the standard error of the slope estimate is proportional to the overall regression mean-squared error (see, e.g., Draper and Smith 1981). The horizontal axes in Fig. 5 are the corresponding values for $COR_2(k)$, with the dashed vertical lines indicating significance levels for local tests taken from Table 1a. Corresponding results for MSE*(k) are comparable but somewhat less distinct. The spatial distributions of these significance levels are broadly comparable to those in Fig. 2, showing results for local test significance using the MSE*(k) criterion, because the divisions generally exhibit comparable local significance levels for either of the two test statistics.

In all five panels of Fig. 5, low values of $COR_2(k)$ are associated with t-ratios near zero, indicating that any linear trends through time are small relative to shorter-period variations. In Fig. 5a, most locations exhibiting significant results for temperature according to $COR_2(k)$ in spring are associated with relatively large and positive t-ratios, suggesting that the apparently good OCN specifications for these developmental data result from a long-term warming trend. Similarly, most of the divisions with locally significant values for temperature in fall (primarily the north-central and northeastern locations in Fig. 2c) are associated with gradual cooling, although a few points (south Florida and the desert southwest) have high correlations associated with warming trends. (Note, however, that the spatial pattern in Fig. 2c does not achieve field significance.) The results for winter temperatures in Fig. 5d again indicate that most of the significant local values of $COR_2(k)$ are associated with temperature trends. Most of these are increasing temperature trends, although the divisions on the southwestern edge of the region of locally significant tests shown in Fig. 2d (generally the lower Mississippi Valley) appear to support OCN predictability as a result of gradually cooling winter temperatures.

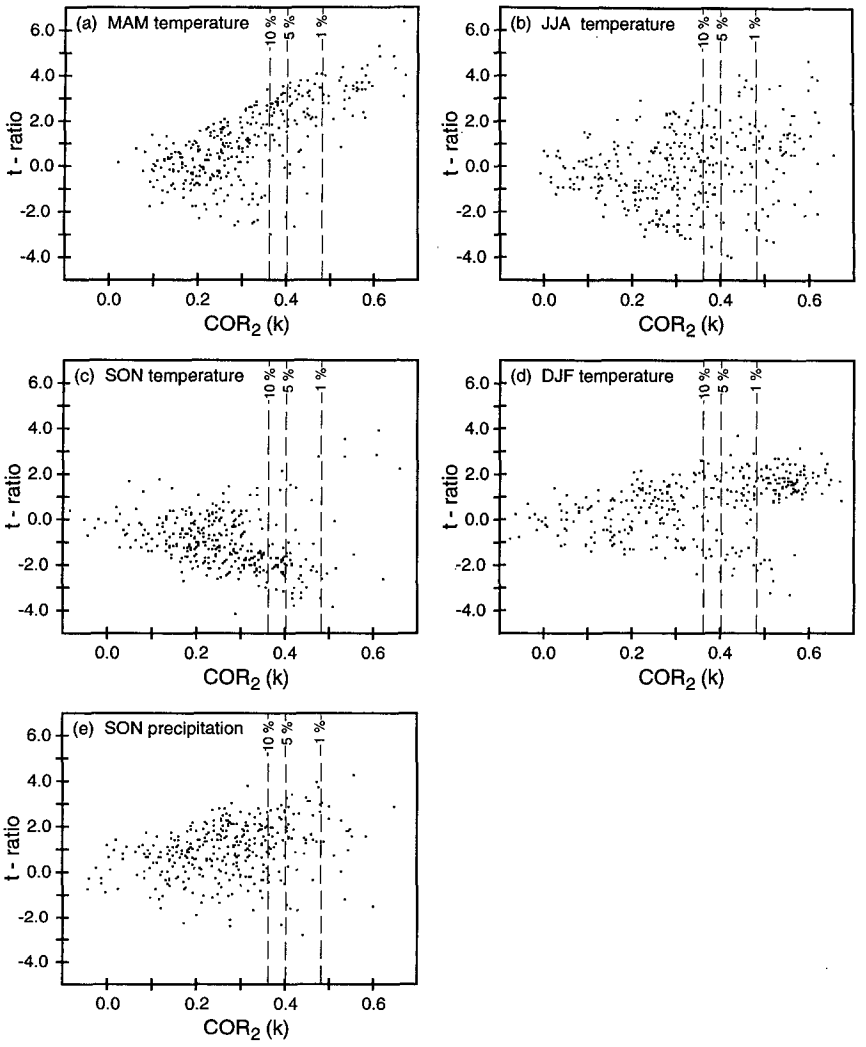

FIG. 5. Regression slopes normalized by their standard errors (t-ratios) as functions of local $COR_2(k)$ values for (a) spring, (b) summer, (c) fall, and (d) winter seasonal temperature forecasts; and (e) fall precipitation forecasts. Local significance levels from Table 1a are indicated by dashed vertical lines. Significant local tests associated with large absolute t-values indicate that OCN predictability results primarily from the existence of long-term trends.

The results for summer temperatures in Fig. 5b are qualitatively different, in that an appreciable fraction of significant local tests are associated with weak slopes, as indicated by relatively small t-ratios. Significant local tests associated with large positive t-ratios in Fig. 5b are located primarily in the west and southeast; while divisions having large $COR_2(k)$ and cooling are located mainly south of Lakes Erie and Ontario and in the center of the country. The remaining divisions with significant local tests in summer exhibit long-term temperature trends that are small in relation to shorter-period variations. Here the large values of $COR_2(k)$ suggest that the year to year temperature variations contain some predictive information.

Finally, Fig. 5e shows the results of the same analysis applies to the fall precipitation series. Here the nominal significance levels have again been taken from Table 1a, although these are approximate for the precipitation data. The precipitation results are more equivocal than those for temperature but indicate a tendency for significant OCN precipitation predictability to be associated with local climates becoming gradually wetter.

6. Operational modifications

It is noted in HVB that the use of separately fitted averaging intervals for each location and season leads

to operational difficulties. In particular, use of very different averaging periods for spatially or temporally adjacent forecasts leads to sharp gradients in the forecast fields, the short space scales or timescales of which are difficult to justify or accept.

Noting that forecast performance does not depend strongly on the averaging period for other than small k (see also Dixon and Shulman 1984; Sabin and Shulman 1985), HVB address this operational complication by finding the value of k maximizing forecast performance when averaged over all seasons and all climate divisions and conclude that $k = 10$ provides this maximum for seasonal temperature forecasts. While conducted using a very large dataset, this is another blind searching procedure, and as such is potentially subject to multiplicity problems of the kind discussed with respect to local OCN fitting in section 3. Its validity is also subject to investigation through methods similar to those used in sections 3 and 4.

Figure 6a shows values of $COR_1(k)$ (symbol "1"), $COR_2(k)$ (symbol "2"), and MSE*(k) (symbol "M"), averaged (area weighted) over all climate divisions and over the 12 three-month seasons as functions of the averaging period. The three labels locate the value $k = 15$, which maximizes [minimizes, for MSE*(k)] these averages. The curve for $COR_1(k)$ shows some differences in magnitude and shape from that presented by HVB; which probably results from area weighting and, to a lesser extent, from the use of all 344 divisions here. While best performance appears to occur for $k = 15$, nearly comparable local optima occur at $k = 10$ and 11 and $k = 25$, which result is also consistent with those obtained by HVB.

The statistical significance of the apparent optima at $k = 15$ in Fig. 6a can be evaluated using a procedure similar to that used in section 4. Here the time ordering of underlying data is repeatedly (10 000 times) shuffled, while preserving both the spatial structure of the data in a given season and the membership of each of the 12-month seasons within a given year. That is, there are $i = 63$ individual data objects, each consisting of 344 spatial dimensions and 12 seasonal dimensions, and 10 000 of the possible 63! permutations of these are randomly selected. For each permutation, the performance of each of the three measures $COR_1(k)$, $COR_2(k)$, and MSE*(k) is determined for all values of k between 1 and 30 for each location and season and then averaged over the seasons and (with area weights) locations. The value of the best averaged performance statistic for each permutation, regardless of the value of k that produced it, is then saved as a contribution to its respective null distribution.

The critical levels for the best of 30 possible averaging periods are indicated in Fig. 6a using line weights. The heaviest lines indicate significance at the 1% level, the medium solid lines indicate significance at the 5% level, the thin black lines indicate significance at the 10% level, and the thin gray lines show the range

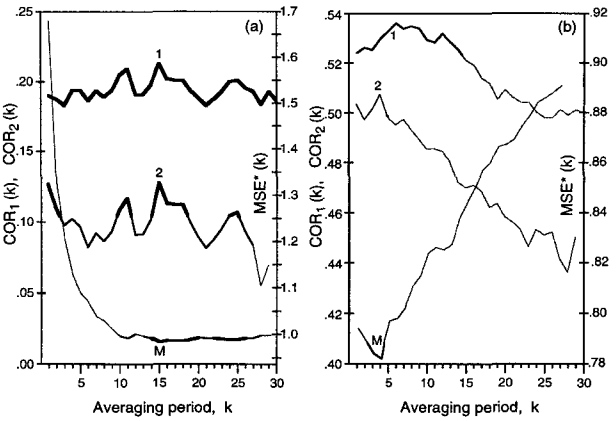

FIG. 6. Area-weighted average values of $COR_1(k)$, $COR_2(k)$, and MSE*(k) as a function of the averaging period, k, for seasonal temperature forecasts. Line weights indicate statistical significance of the results for the averaging period yielding best performance (located by the symbols 1, 2, and M, respectively): heavy solid lines, $p \leq 0.01$; medium solid lines, $0.01 < p \leq 0.05$; light solid lines, $0.05 < p \leq 0.10$; faint lines, $p > 0.10$. (a) Averages over all divisions and seasons, (b) averages over only those divisions and seasons for which local tests were significant at the 5% level.

of values not significant at even the 10% level. Note that these values pertain to the single best performance of the 30 possible averaging periods and thus are conservative test levels for, say, the second- and third-best values. For all three of the performance measures, the best values at $k = 15$ are better than would be expected by chance at better than the 1% level (actually, better than the 0.1% level), as are the secondary optima at $k = 10$ and 11 and $k = 24$ and 25.

Results in section 2 indicate that there are large areas in the United States over which OCN forecasts for particular seasons show no significant skill. Including these areas in the search for a "globally" optimal k would thus seem to dilute any predictive signal that may be present and/or produce biased specifications. Figure 6b shows the results of a search similar to that represented in Fig. 6a, except that only local values of the performance measures that are significant at the 5% level are included in the averages. There is again substantial agreement among the resulting best "global" averaging periods for the three measures $COR_1(k)$, $COR_2(k)$, and MSE*(k), which yield $k = 6$, $k = 4$, and $k = 4$, respectively. Of course the magnitudes of the results in Fig. 5b are substantially higher [lower, for MSE*(k)] than in Fig. 6a because values not satisfying local statistical significance are not included in the averages.

Statistical significance of the three optima located by the symbols "1," "2," and "M" in Fig. 6b is computed in the same was as described above with respect to Fig. 6a, except that the averages comprising the respective null distributions are comprised only of those local values in each shuffled data series that are nom-

inally significant at the 5% level. Line weights indicating the significance levels are the same as those used in Fig. 6a, with the maximum of $COR_1(6)$ being significant at the 5% level ($p = 0.017$), the maximum of $COR_2(4)$ being significant at the 10% level ($p = 0.061$), and the minimum of MSE*(4) being significant at the 5% level ($p = 0.038$). These results in Fig. 6b also agree that the search for a global averaging period produces valid results for the seasonal temperature forecasts. The choices for best averaging period are reasonably consistent among the three forecast performance measures but are quite different from those indicated in Fig. 5a.

Figure 7 presents the corresponding analyses for the seasonal precipitation forecasts. When all climate divisions and seasons are averaged (Fig. 7a), the correlation measures indicate maxima for short averaging times ($k = 2$ and $k = 4$) and other maxima near $k = 15$ and $k = 27$. All three of these are statistically significant at the 5% level for the $COR_1(k)$ measure, while the maximum for $COR_2(2)$ achieves significance only at the 10% level ($p = 0.082$). The best result for the MSE*(k) statistic is for $k = 29$ and is not significantly better than could be expected by chance. These results are considerably weaker than those for the seasonal temperature forecasts in Fig. 6a and provide little support for the estimation of a "global optimum" averaging period, at least in the way being evaluated here. Figure 7b provides the seasonal precipitation forecast results for averages taken only over divisions and seasons for which local tests are significant at the 10% level. None of the three maxima are significant even at the 10% level, further indicating that there is little promise in this approach for seasonal precipitation forecasts.

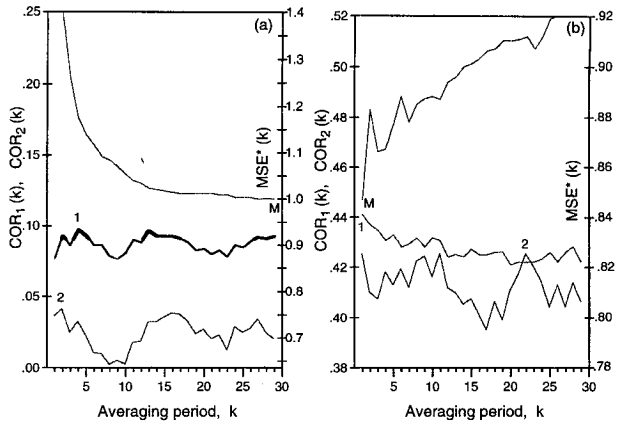

FIG. 7. As Fig. 6, for seasonal precipitation. Averages in (b) are over divisions and seasons for which local tests were significant at the 10% level.

7. Summary and conclusions

Two multiplicity problems arise in the evaluation of OCNs at many locations over a large area such as the United States, using only dependent data. First, the fitting procedure is allowed to choose the best averaging period, among many, over which to calculate the OCNs for a given location. The resulting averaging period will thus be "tuned" to some extent to the random sampling variations of the data series at hand. Second, simultaneous comparison of the performance of the OCNs over a network of locations will yield apparently positive results by chance for some locations, even if the forecasts possess no real predictive skill. This second problem is of particular concern because choosing the best of 30 averaging periods at each location enhances the magnitude of dependent-sample artificial skill and because the spatial correlation of the data will tend to produce spatially coherent regions exhibiting that artificial skill.

The extent to which the apparent predictive capacity of seasonal and monthly OCN hindcasts of temperature and precipitation is larger than would be expected by chance has been examined here using nonparametric hypothesis tests. At the local level, the sampling distributions of several forecast performance statistics under the null hypothesis were constructed by repeatedly reordering the data. Values of these performance statistics computed from the actual data that are better than all but a small proportion of those computed for the random series are judged to be sufficiently good that they are unlikely to have occurred by chance.

The field significance of these local tests was approached in the same way, by reordering the underlying fields consisting of data at 344 individual climate divisions and computing the areas exhibiting nominally significant local tests. Field significance results for seasonal temperature forecasts indicate strongly that, with the exception of the fall and early winter seasons SON through NDJ, significant local skill is sufficiently widespread that it is unlikely to have occurred by chance. Significant results for seasonal precipitation are much less widespread in space and convincingly achieve field significance only for SON and OND. Results for monthly temperature forecasts are somewhat better than those for seasonal precipitation but less impressive than those for seasonal temperature. Results for monthly precipitation forecasts appear to be no better than would be expected by chance.

Examination of the relationship between OCN performance and temperature trends for four seasons indicates that, in most cases, statistically significant local OCN specifications result from gradual temperature changes that can be captured as linear trends in regression analyses. However, a substantial number of locally significant tests for summer occur in the absence of such a trend, suggesting that, for these, the OCNs are exploiting shorter-term temperature variations. Of course, whether these patterns will persist into the com-

ing decades is an open question that cannot be addressed by a study of this kind.

Operationally, the use of OCNs presents the problem that allowing k to vary freely produces highly irregular forecast fields. The approach proposed by HVB to alleviate this problem is to choose a single best averaging period for all locations and seasons. Appropriately designed significance tests indicate that this approach is justified for the seasonal temperature forecasts but is dubious for the seasonal precipitation forecasts.

The significance testing approach adopted here is less than ideal but has been resorted to because cross validation does not seem practicable given the available data. The testing framework does allow identification of levels of local forecast performance that are too great to have plausibly arisen by chance and provides reassurance that many of the spatial patterns of significant forecast performance are unlikely to have resulted only from sampling variations. However, this analysis does not yield estimates of those forecast performance measures that could be expected for independent (i.e., future) data. Rather, the critical values used here (e.g., in Table 1a) are better than can be expected for forecasts of future independent data, precisely because they have been constructed to reproduce the factors that lead to artificial skill for hindcasts. On the other hand, the critical values in Table 1b, which are based on randomly chosen averaging periods, are probably overly conservative. Respective pairs of values in Tables 1a and 1b might be expected to bracket future forecast performance.

The performance of the OCN temperature and precipitation forecasts is significantly better than that of 30-yr climate normals for some times and places but is quite modest overall. Even so, however, it is likely to remain a viable contribution to these difficult forecasts for the near future.

Acknowledgments. This work was supported by the Northeast Regional Climate Center under NOAA Grant NA16CP-0220-02.

REFERENCES

Barnston, A. G., and H. M. van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate,* **5,** 963–977.

Dixon, K. W., and M. D. Shulman, 1984: A statistical evaluation of the predictive abilities of climatic averages. *J. Climate Appl. Meteor.,* **23,** 1542–1552.

Draper, N. R., and H. Smith, 1981: *Applied Regression Analysis.* 2d ed., Wiley and Sons, 709 pp.

Efron, B., 1982: *The Jackknife, the Bootstrap, and Other Resampling Plans.* Society for Industrial and Applied Mathematics, 92 pp.

Elsner, J. B., and C. P. Schmertmann, 1994: Assessing forecast skill through cross validation. *Wea. Forecasting,* **9,** 619–624.

Huang, J., H. M. van den Dool, and A. G. Barnston, 1996: Long-lead seasonal temperature prediction using optimal climate normals. *J. Climate,* **9,** 809–817.

Kunkel, K. E., and A. Court, 1990: Climatic means and normals—a statement of the American Association of State Climatologists (AASC). *Bull. Amer. Meteor. Soc.,* **71,** 201–204.

Lamb, P. J., and S. J. Changnon, Jr., 1981: On the "best" temperature and precipitation normals: The Illinois situation. *J. Appl. Meteor.,* **20,** 1383–1390.

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.,* **111,** 46–59.

Mielke, P. W., Jr., K. J. Berry, and G. W. Brier, 1981: Application of multi-response permutation procedures for examining seasonal changes in monthly mean sea-level pressure patterns. *Mon. Wea. Rev.,* **96,** 540–548.

National Climatic Data Center, 1994: Time bias corrected divisional temperature-precipitation drought index. Documentation for dataset TD-9640. NOAA/NCDC, Asheville, NC, 12 pp.

Ropelewski, C. F., and J. B. Jalickee, 1983: Estimating the significance of seasonal precipitation amounts using approximations of the inverse gamma function over an extended range. Preprints, *Eighth Conf. on Probability and Statistics in the Atmospheric Sciences,* Amer. Meteor. Soc., 125–129.

Sabin, T. E., and M. D. Shulman, 1985: A statistical evaluation of the efficiency of the climatic normal as a predictor. *J. Climatol.,* **5,** 63–77.

von Storch, H., 1982: A remark on Chervin–Schneider's algorithm to test significance of climate experiments with GCMs. *J. Atmos. Sci.,* **39,** 187–189.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 464 pp.

——, and K. L. Eggleston, 1992: Estimating monthly and seasonal precipitation distributions using the 30- and 90-day outlooks. *J. Climate,* **5,** 252–259.