Diagnostic Verification of the Climate Prediction Center Long-Lead Outlooks, 1995–98

D. S. WILKS

Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York

(Manuscript received 20 May 1999, in final form 24 September 1999)

ABSTRACT

The performance of the Climate Prediction Center's long-lead forecasts for the period 1995–98 is assessed through a diagnostic verification, which involves examination of the full joint frequency distributions of the forecasts and the corresponding observations. The most striking results of the verifications are the strong cool and dry biases of the outlooks. These seem clearly related to the 1995–98 period being warmer and wetter than the 1961–90 climatological base period. This bias results in the ranked probability score indicating very low skill for both temperature and precipitation forecasts at all leads. However, the temperature forecasts at all leads, and the precipitation forecasts for leads up to a few months, exhibit very substantial resolution: low (high) forecast probabilities are consistently associated with lower (higher) than average relative frequency of event occurrence, even though these relative frequencies are substantially different (because of the unconditional biases) from the forecast probabilities. Conditional biases, related to systematic under- or overconfidence on the part of the forecasters, are also evident in some circumstances.

1. Introduction

Forecasts of future weather conditions at long leads (months into the future) have the potential to provide very substantial economic value, even if the predictands are averages of meteorological quantities over months or seasons. Prominent among areas where decision making could be improved through use of such forecasts are energy (Knox et al. 1985); agribusiness (Mjelde et al. 1998); hydrology (Kim and Palmer 1997); and also the financial services industry, through the new financial instruments known as ''weather derivatives'' (Dischel 1998). However, in order for decision makers to realize full (or even positive) benefit from forecasts, a reasonably comprehensive understanding of the quality attributes of those forecasts is necessary (Katz and Murphy 1997).

One source of long-range forecasts for time-averaged conditions are the probabilistic long-lead outlooks produced operationally by the Climate Prediction Center (CPC) of the U.S. National Weather Service (O'Lenic 1994). These have been formulated and distributed in their present format since December 1994, at which time they replaced a similar but more limited forecast product that had been in operation since July 1982 (Epstein 1988; Wagner 1989). Building on earlier developments in long-range forecasting (Namias 1968), both of these systems have relied on subjective (i.e., human forecaster) probability judgments to reconcile the information from a collection of objective guidance products. The judgments expressed in the present outlooks give particular weight to relationships between time-averaged North American weather with the state and intensity of the El Niño–Southern Oscillation (ENSO) phenomenon (O'Lenic 1994).

This paper reports on the diagnostic verification (Murphy 1997; Murphy et al. 1989; Murphy and Winkler 1987, 1992) of the new CPC outlooks. Diagnostic verification of the previous generation of CPC long-lead forecasts was reported by Murphy and Huang (1991). Diagnostic verification differs from the more traditional ''measures-oriented'' approaches in that it examines the full joint frequency distributions of the forecasts under consideration and their corresponding observations, rather than computing and examining one or a few scalar measures of correspondence between forecasts and observations (e.g., mean squared error, or correlation between forecasts and observations). Diagnostic verification is clearly more elaborate and detailed than measures-oriented verification, but it offers important advantages. It allows identification of particular aspects of a collection of forecasts that may be strong, and others toward which efforts for improvement could be targeted. This information will generally be of interest to those responsible for producing the forecasts. In addition, the results of a diagnostic verification exercise can provide forecast users with sufficiently detailed information

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

about the quality of the forecasts to allow their optimal use (e.g, Murphy 1994; Wilks 1997; Winkler and Murphy 1985), and thus to derive maximum economic value from them. In addition, the validity of value-added transformations of the CPC outlooks (Briggs and Wilks 1996a,b; Croley 1996; and the interpretive products offered by CPC at http://www.cpc.ncep.noaa.gov/pacdir/NFORdir/HOME3.html) is predicated on the forecasts satisfying certain consistency relationships with the observations, which cannot be assessed through examination of only scalar measures of correspondence.

2. Diagnostic verification

A forecast verification dataset consists of a collection of forecasts, f, and the corresponding set of observations, o, to which the forecasts pertain. Generally each forecast and observation is rounded to one of a finite set of possible values; f_i, $i = 1, \ldots, I$; and o_j, $j = 1, \ldots, J$. The collective behavior of these forecasts and observations (apart from any time dependence that either or both may exhibit) is summarized fully by the joint frequency distribution of the forecasts and observations,

$$p(f_i, o_j) = \Pr\{f_i \cap o_j\}. \quad (1)$$

Here the intersection symbol $\cap$ can be read as ''and,'' and the distinction between sample relative frequency (left-hand side) and population probability (right-hand side) has been neglected. The joint distribution of forecasts and observations in (1) can be visualized as a table or matrix of dimension $I \times J$, each entry of which contains the number of occasions that both forecast f_i and observation o_j occurred divided by the total number of forecast–observation pairs in the dataset.

Diagnostic verification consists of describing and summarizing the statistical relationships in the joint distribution of forecasts and observations. However, for even modestly sized verification problems direct interpretation of (1) can be difficult. In practice it has been found that factored forms of (1) and, in particular, graphical representations of these factorizations, can be extremely illuminating. Two factorizations of the joint distribution (1) into a conditional distribution and a marginal distribution are possible (Murphy and Winkler 1987). Here, the factorization

$$p(f_i, o_j) = q(o_j | f_i) r(f_i), \quad (2)$$

called the calibration-refinement factorization, will be used. The calibration-refinement factorization expresses the joint distribution of the forecasts and observations as the product of the conditional distributions of the observations given each of the forecasts $q(o_j | f_i)$ (the ''calibration''), and the distribution $r(f_i)$ expressing the unconditional frequency of use of each of the I forecasts (the ''refinement'').

Because the forecasts of interest here are probabilistic in nature, the quantities f_i in Eqs. (1) and (2) are probabilities. The verification results in section 4 will be presented primarily in terms of the calibration-refinement factorization [(2)] and, in particular, using a graphical device for portraying the two distributions $q(o_j | f_i)$ and $r(f_i)$ called the reliability diagram (e.g., Wilks 1995). Reliability diagrams depict the I conditional relative frequencies $q(o_j | f_i)$ as a sequence of points that define a function of the forecasts f_i. For probability forecasts that are well calibrated (''reliable'') there is a close correspondence between each forecast f_i and the relative frequencies of the subsequent observations $q(o_j | f_i)$, so that the plotted points fall close to the 45° diagonal. The distribution $r(f_i)$, depicting the frequency of use of each of the I possible forecast values, is generally plotted as an inset bar chart or histogram in the larger diagram. The nature and, in particular, the dispersion, or variance, of the distribution $r(f_i)$ is also an important determinant of overall forecast quality. For a given degree of calibration, forecasts exhibiting higher-variance refinement distributions $r(f_i)$ deviate more frequently and more extremely from the climatological event probability, and thus express more confidence in aggregate. Given well-calibrated forecasts, a higher-variance refinement distribution implies greater accuracy or skill. In the extremes, the variance of $r(f_i)$ for perfect forecasts is a maximum [equal to $\pi(1 - \pi)$, where π is the sample climatological event frequency], while for the climatological forecast ($f_i \equiv \pi$) this variance is zero.

3. Data

a. Forecasts

The forecasts of interest here are the long-lead outlooks of the CPC for average temperature and total precipitation over the conterminous United States, in the format initiated in December 1994. These forecasts are constructed and disseminated monthly, near the middle of each month, and consist of 14 forecast maps each for temperature and precipitation. The first pair of maps pertains to temperature and precipitation outcomes for the next calendar month, that is, they relate to average monthly temperature and total monthly precipitation with a lead time of approximately 2 weeks. The remaining 13 map pairs pertain to overlapping 3-month ''seasons,'' the first of which also begins approximately 2 weeks after the forecast is issued. For example, the first of these forecasts, issued in December 1994, included temperature and precipitation forecasts for the month of January 1995; and also for the 3-month periods January–March 1995, February–April 1995, etc., through January–March 1996. The forecast dataset to be analyzed here includes forecasts issued each month from December 1994 through December 1998.

As mentioned previously, these are probabilistic forecasts, so the quantities displayed on the forecast maps are probabilities rather than particular temperature or precipitation values. The forecast probabilities relate to

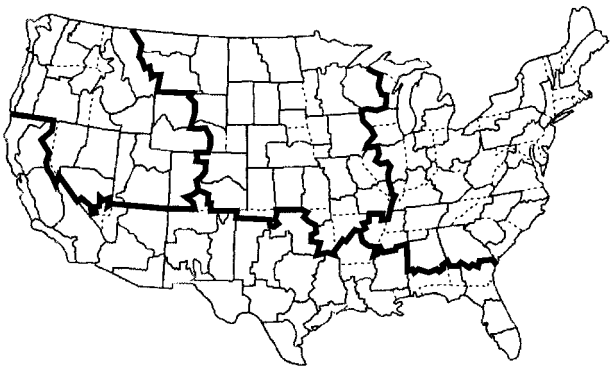

FIG. 1. The 102 divisions of the United States pertaining to the forecasts and observations (solid lines). Dashed lines indicate state borders that are not coincident with division boundaries. Heavy lines divide the country into four regions, defined according to a cluster analysis of the forecasts.

temperature or precipitation outcomes being in the lower ⅓ (below normal: cool or dry), middle ⅓ (near-normal), or upper ⅓ (above normal: warm or wet) of the (1961–90 base period) climatological distributions of average temperature or total precipitation for the appropriate month or season at particular locations. The forecast maps have been discretized by CPC to the 102 areas indicated in Fig. 1. These areas are coincident with, or are amalgamations of, the climatic divisions used by the U.S. National Climatic Data Center (NCDC). For each month, the forecast dataset thus includes ($2 \times 102 \times 14 =$) 2856 probabilities for the below-normal outcomes and equal numbers for the near-normal and above-normal outcomes. Note, however, that the forecasts are far from being mutually independent. Rather, they exhibit quite strong spatial correlation, as there are typically three or fewer major features on a forecast map for the United States; and they often exhibit time continuity (i.e., temporal autocorrelation) as well.

Because the categories to which the forecasts pertain are mutually exclusive and collectively exhaustive (the outcome for a month or season must be either below, near, or above normal, as defined), the three probabilities for these outcomes at a given time and place must sum to 1. Thus at most two probabilities need to be specified to fully define a forecast. Operationally, further restrictions are placed on each forecast, which provide for full specification using only one probability, and thus allow easy display on a single map. The mapped quantities are ''probability anomalies'' for the one category of the three that is judged to be most likely. Since the climatological probability for each category is $\frac{1}{3}$, the forecast probability of the indicated outcome is obtained by adding $\frac{1}{3}$ to the mapped anomaly. If the most likely category is above or below normal, the probability of the near-normal category is regarded as equal to the climatological value of $\frac{1}{3}$ and the probability for the below- or above-normal category, respectively, is decreased by an amount equal to the probability anomaly. For example, if the probability for a below-normal outcome is $\frac{1}{2}$, then by implication the probability for the near-normal outcome is $\frac{1}{3}$ and the probability for the above-normal outcome is $\frac{1}{6}$. Occasionally the near-normal outcome is forecast as most likely, in which case the additional probability is removed equally from the below- and above-normal categories. Often the forecasters judge that they are unable to add information beyond the climatological probabilities, in which case the probability anomalies for all three categories are zero.

This current format for the CPC forecasts differs most from the previous (1982–94) forecasts with respect to lead time. These earlier forecasts provided a 1-month forecast and a single 3-month forecast, each with a lead time of 2–3 days. Other differences are that the older system defined the climatological probabilities of the three categories as 0.3–0.4–0.3 rather than ⅓–⅓–⅓, and that the probability for the near-normal category was always specified to be the climatological 0.4.

b. Observations

The basic observational data to be used are the monthly averaged temperatures and the monthly total precipitation for each of the 102 areas shown in Fig. 1 for the period January 1995–January 1999. These time series were constructed at CPC as weighted averages of monthly data from appropriate subsets of the 344 NCDC climate divisions (D. Unger 1999, personal communication). For each location, each monthly value in the series and each of the 3-month values from January–March (JFM) 1995 through November–January (NDJ) 1998/99 has been converted to the appropriate category (below, near, or above normal) according to its magnitude in relation to the corresponding 1961–90 climatological distribution. Because this base period is comparatively short, the two terciles defining the three categories have been obtained using smooth distribution functions fit to each set of 30 observations. The temperature data have been modeled using Gaussian distributions. The precipitation data have been modeled using gamma distributions, with individual months or seasons having zero precipitation treated as censored data (Wilks 1990).

4. Results

a. Scalar skill scores and lead-time stratification

In order to improve sample sizes, and to restrain the number of figures required in the following, the verification results will be aggregated over three groups of lead times. These stratifications are somewhat arbitrary, although they do appear as fairly natural groupings of Ranked Probability Scores (RPSs) (Epstein 1969; Wilks 1995) for the individual lead times. The RPS is computed as the averaged sum of squared differences be-

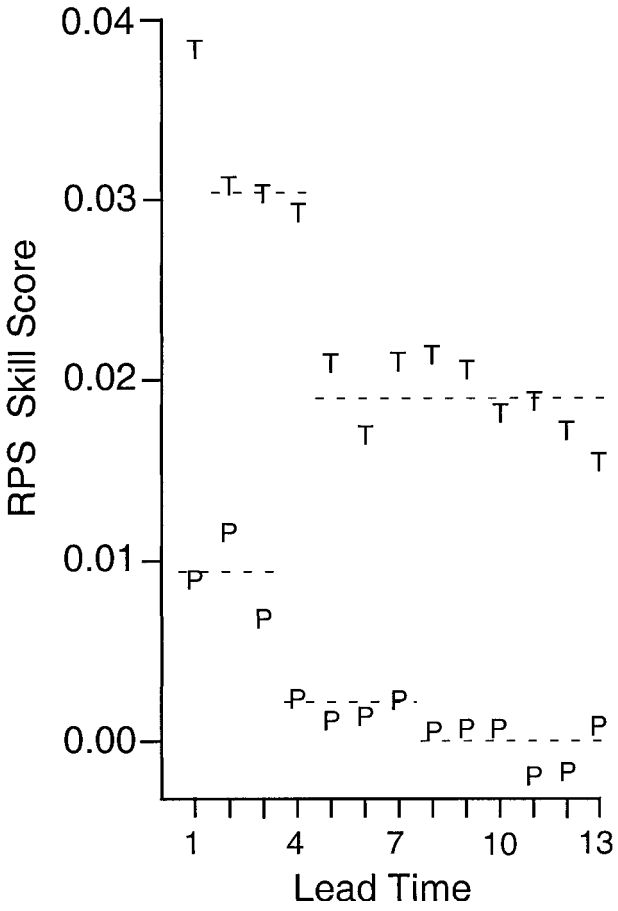

FIG. 2. Skill scores, based on the Ranked Probability Score in relation to the climatological probabilities, of the seasonal temperature (T) and precipitation (P) outlooks, as functions of the lead time (lead 1 = 0.5 month, lead 13 = 12.5 months). Dashed horizontal lines show skill scores averaged over the indicated ranges of lead times.

tween the *cumulative* distributions of the forecasts and observations,

$$\mathrm{RPS} = \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{3}\left[\left(\sum_{k=1}^{m} f_{i,k}\right) - \left(\sum_{k=1}^{m} o_{j,k}\right)\right]^2. \qquad (3)$$

Here the index k denotes the below-normal ($k = 1$), near-normal ($k = 2$), or above-normal ($k = 3$) outcomes, and T is the number of forecast–observation pairs being compared. The observation $o_{j,k}$ is a binary variable whose value is 1 for the observed category and 0 for the remaining two categories (e.g., $o_{j,1} = 0$, $o_{j,2} = 1$, and $o_{j,3} = 0$ for a near-normal outcome).

Figure 2 shows skill scores based on the RPS (i.e., RPS scaled and normalized by the RPS obtained using the climatological forecast ⅓–⅓–⅓; e.g., Wilks 1995) for the seasonal temperature (T) and precipitation (P) forecasts. It is immediately apparent that the temperature forecasts are more accurate than the precipitation forecasts according to this measure and that the accuracy of both decreases with increasing lead time. According to this skill score, the temperature forecasts appear to cluster in three lead-time groups, lead 1 (0.5 month), leads 2–4 (1.5–3.5 months), and leads 5–13 (4.5–12.5 months); while these results for the precipitation forecasts appear to cluster (although less markedly) at leads 1–3 (0.5–2.5 months), 4–7 (3.5–6.5 months), and 8–13 (7.5–12.5 months). These lead-time stratifications will be adopted in the following.

Note that the magnitudes of the skill scores in Fig. 2 are rather small. Conventionally this skill score is interpreted as a proportionate increase in accuracy from the reference (climatological) forecasts, in relation to the accuracy difference between the reference and perfect forecasts. While the CPC outlooks certainly do not approach the level of perfect forecasts, Fig. 2 nonetheless presents an overly pessimistic view of their potential usefulness and value. As a scalar measure of forecast performance, this skill score necessarily aggregates (in a somewhat arbitrary manner) many contributions to forecast accuracy into a single number. Put another way, many collections of forecasts with quite different error characteristics could receive exactly the same score according to a particular scalar measure, and distinguishing them according to that measure would clearly be impossible. A much fuller exposition of the performance of these forecasts is obtained below, through examination of their joint distributions with the corresponding observations.

b. Reliability diagrams

Figures 3 and 4 contain reliability diagrams for the 1-month and seasonal forecasts, respectively, with the seasonal results in Fig. 4 further stratified according to the lead times derived from Fig. 2. Results for both the below-normal (cool, ''C''; or dry, ''D'') and above-normal (warm or wet, ''W'') probability forecasts are plotted on the same figure in each case. Results for probabilities assigned to near-normal outcomes are not shown, primarily because in the comparatively rare instances where they are different from the climatological ⅓, they deviate very little from this value, and furthermore exhibit generally poor reliability.

For purposes of these plots, the forecasts f_i have been grouped in the 15 bins $f_1 \le 0.025$, $0.025 < f_2 \le 0.075$, . . . , $0.275 < f_7 \le 0.325$, $0.325 < f_8 \le 0.335$, $0.335 < f_9 \le 385$, . . . , $0.585 < f_{14} \le 0.635$, and $f_{15} > 0.635$. The central bin, containing the climatological forecast ⅓, is narrower than the others because this forecast is used most frequently. The calibration functions $q(o_j \mid f_i)$ in Figs. 3 and 4 are indicated by solid lines connecting the symbols ''C,'' ''D,'' or ''W,'' with the line thicknesses indicating the smaller of the two sample sizes for each pair of points: heavy lines for $n \ge 500$, medium lines for $500 > n \ge 50$, and light lines for $n < 50$. Points on the calibration functions with $n < 10$ have not been plotted. The light broken lines running through each calibration function are weighted (by sam-

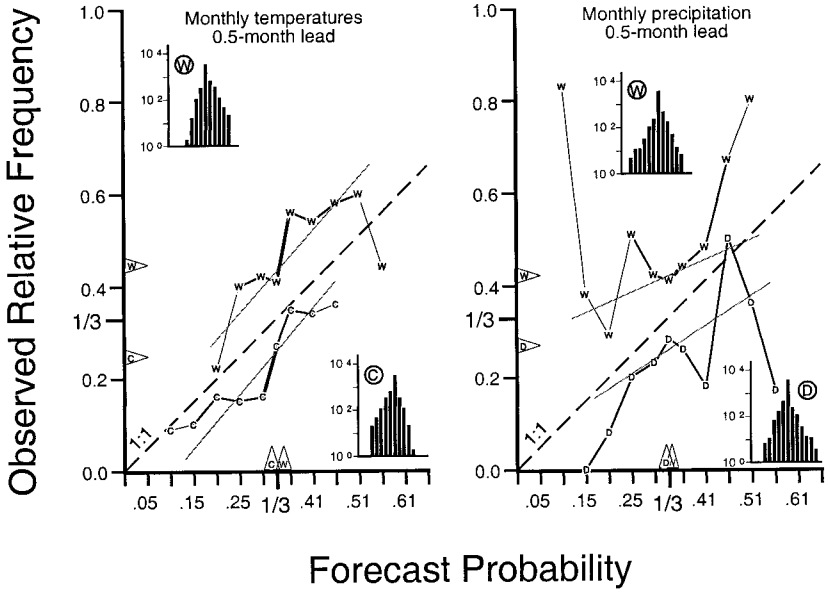

FIG. 3. Reliability diagrams [graphical depictions of (2)], for the 1-month forecasts of (left) temperature and (right) precipitation, made with a lead time of approximately 2 weeks. Calibration functions $q(o_j | f_i)$ for the warm or wet (above normal) outcomes are indicated by "W"; those for cool and dry outcomes are indicated by "C" and "D," respectively. Thickness of line segments connecting these symbols increases with sample size, and the light lines show weighted least squares regressions for each calibration function. Inset bar charts (note logarithmic vertical scales) indicate the refinement functions $r(f_i)$. Triangular symbols on the horizontal and vertical axes locate the average forecasts and average observations, respectively.

ple size, e.g., Draper and Smith 1981, p. 108 ff.) least squares fits to $q(o_j | f_i)$ as functions of f_i, as an aid to guiding the eye through the sometimes-considerable sampling variations (Murphy and Wilks 1998). The heavy dashed lines indicate the 1:1 relationship, onto which the $q(o_j | f_i)$ points for perfectly calibrated (i.e., fully reliable) forecasts would fall exactly.

The inset bar charts in Figs. 3 and 4 portray the refinement distributions, $r(f_i)$. Note that these are *not* histograms of $r(f_i)$ and, in particular, that the vertical scales on these insets are logarithmic in order that they can portray the huge range of sample size for the different forecast probabilities (forecasts with $n < 10$ are shown on these insets, so that the number of bars is larger than the number of plotted calibration-function points). In all cases the category containing the climatological probability ⅓ is by far the most frequently used, as noted above, even though this bin is much narrower than the others. The degree of dispersion of $r(f_i)$ around the average forecast $\overline{f}$ is indicative of the degree of confidence of the forecasters: distributions with nearly all their mass near the climatological value suggest that the forecasters have low confidence that they can discern deviations from the climatological probabilities for the three outcomes, while the sharper distributions containing more frequent use of the extreme probabilities reflect more confidence in aggregate. The average forecasts, $\overline{f}$, are indicated by the triangular symbols on the horizontal axes of each diagram. The corresponding average observations, $\overline{o}$ (i.e., the sample climatological relative frequencies), are indicated by the corresponding triangular symbols on the vertical axes.

The most prominent and consistent features of Figs. 3 and 4 are the strong cold bias in the temperature forecasts, and the strong dry bias in the precipitation forecasts, at all lead times. That is, while the average forecasts are quite near the climatological value of ⅓ in all cases, the observed event relative frequencies differ consistently from ⅓, reflecting the fact that the 1995–98 period to which the forecasts pertain was both warmer and wetter than the 1961–90 climatological base period. These differences were evidently not recognized in advance, in aggregate, by the forecasters during 1995–98. Clearly, these biases contribute negatively to the scalar skill scores shown in Fig. 2.

For the temperature forecasts at all leads, and for the monthly precipitation forecasts, the conditional probabilities $q(o_j | f_i)$ do increase steadily with increasing f_i, indicating that for these cases the forecasters were able to resolve subsets of the valid periods with different frequencies of the temperature and precipitation outcomes. Conventionally, resolution is summarized by the statistic

$$\text{RES} = \frac{1}{n} \sum_{i=1}^{I} n_i [q(o_j | f_i) - \overline{o}]^2, \qquad (4)$$

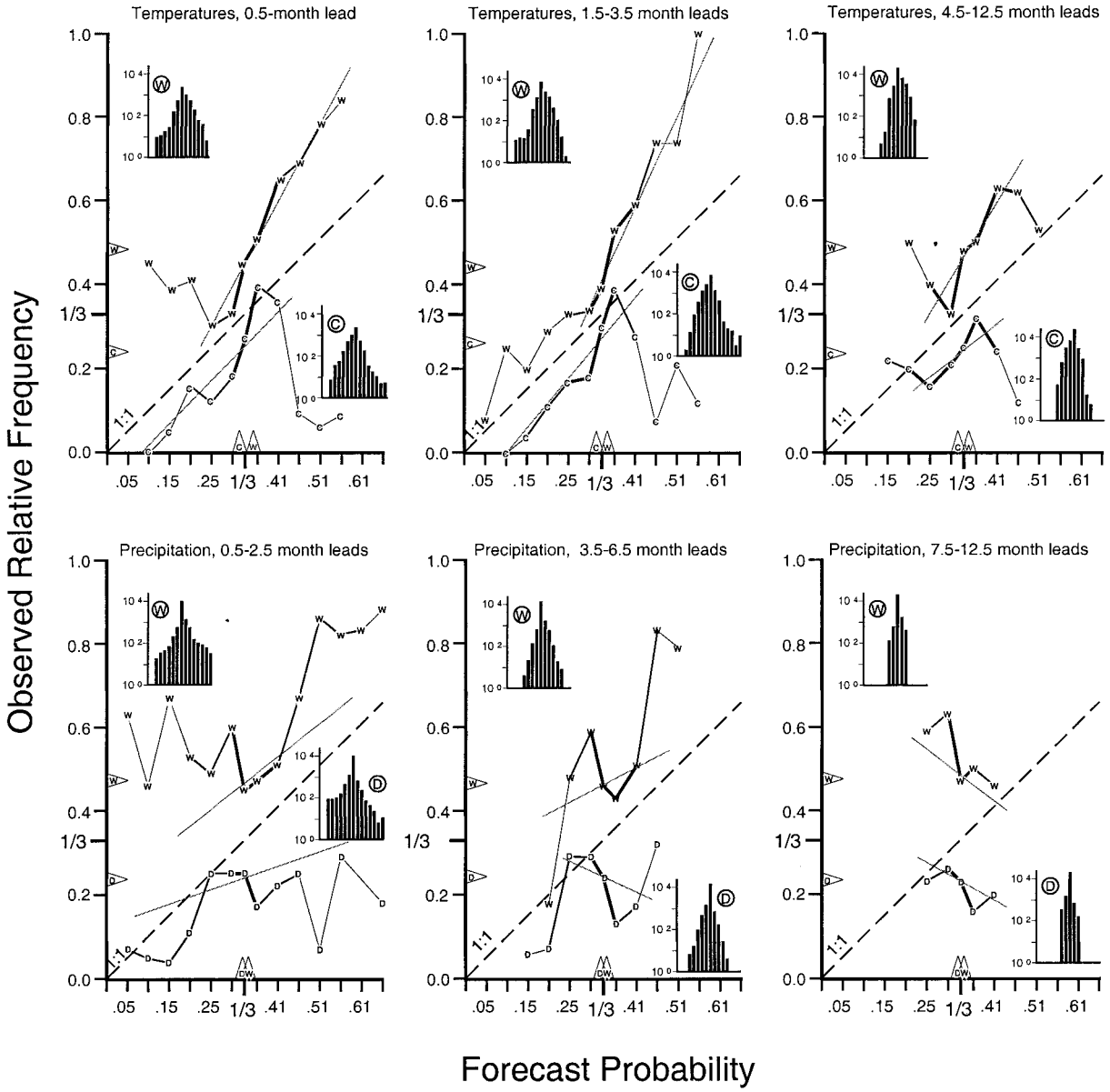

FIG. 4. As in Fig. 3 but for the seasonal temperature (top panels) and precipitation (bottom panels), stratified by lead time as indicated in Fig. 2.

where n_i is the number of forecasts in the ith category, and $n = \Sigma_i (n_i)$ is the total sample size. Geometrically, RES is the average squared distance between the points $q(o_j | f_i)$ and the average outcome $\overline{o}$ indicated by the triangular symbols on the vertical axes in Figs. 3 and 4. Values of the RES statistic for the reliability diagrams in Figs. 3 and 4 are given in Table 1. The ability to successfully resolve different outcomes is of course a necessary condition for the forecasts to be useful, and it is encouraging that even at the longest leads (4.5–12.5 months) the temperature forecasts do exhibit good resolution over the range of probabilities used. The resolution exhibited by the seasonal precipitation forecasts is clearly less good, although the forecasts for above-normal precipitation outcomes at the 0.5–2.5-month and 3.5–6.5-month leads appear also to exhibit some useful resolution.

The inset bar charts portraying the refinement distributions $r(f_i)$ show that, particularly for the seasonal forecasts at the shortest lead times, probabilities quite near the extremes of the allowable range are used at least occasionally. The dispersion of these distributions is markedly greater than found by Murphy and Huang (1991) for the earlier generation of these outlooks. This increase in apparent forecaster confidence is especially striking, considering that the older forecast format in-

TABLE 1. Four-parameter summaries of the reliability diagrams in Figs. 3 and 4, following Murphy and Wilks (1998) for (a) temperature forecasts and (b) precipitation forecasts. The parameters b_0 and b_1 are the intercept and slope, respectively, of the weighted least squares line fits to the calibration functions $q(o_j|f_i)$, which are indicated as light broken lines in the figures. The parameters $\bar{f}$ and s_f are the mean and standard deviation, respectively, of the refinement distribution $r(f_i)$ shown in the insets. Also included are the unconditional biases [(5)] and values of the RES statistic [(4)]. Results for 1-month forecasts and seasonal forecasts aggregated by lead time according to results in Fig. 2 are shown.

(a) Temperature forecasts.

Valid	Lead (month)	Cool						Warm					
		b_0	b_1	$\bar{f}$	s_f	Bias	RES	b_0	b_1	$\bar{f}$	s_f	Bias	RES
Month	0.5	−0.13	1.19	0.32	0.042	0.069	0.0036	0.06	1.13	0.34	0.042	−0.104	0.0046
Season	0.5	−0.08	1.02	0.32	0.054	0.075	0.0073	−0.16	1.85	0.35	0.054	−0.138	0.0121
Season	1.5–3.5	−0.12	1.19	0.32	0.044	0.059	0.0056	−0.31	2.16	0.35	0.044	−0.096	0.0113
Season	4.5–12.5	−0.02	0.79	0.32	0.033	0.087	0.0013	−0.09	1.68	0.34	0.034	−0.141	0.0046

(b) Precipitation forecasts.

Valid	Lead (month)	Dry						Wet					
		b_0	b_1	$\bar{f}$	s_f	Bias	RES	b_0	b_1	$\bar{f}$	s_f	Bias	RES
Month	0.5	0.06	0.62	0.33	0.037	0.065	0.0019	0.28	0.44	0.34	0.037	−0.090	0.0030
Season	0.5–2.5	0.12	0.35	0.33	0.047	0.094	0.0014	0.20	0.78	0.34	0.048	−0.125	0.0047
Season	3.5–6.5	0.39	−0.46	0.33	0.022	0.092	0.0011	0.29	0.52	0.34	0.022	−0.127	0.0020
Season	7.5–12.5	0.42	−0.56	0.33	0.013	0.095	0.0002	0.74	−0.78	0.33	0.013	−0.153	0.0008

volved essentially zero lead time. Overall impressions of dispersion in the refinement distributions that can be obtained from the bar charts in Figs. 3 and 4 are quantified by the standard deviations of these distributions, which are included in Table 1. The dispersion of these distributions for the temperature forecasts are greater than for the precipitation forecasts, as the latter quantity is generally regarded as more difficult to forecast. Also, the standard deviations of $r(f_i)$ for the seasonal forecasts decrease with increasing lead time, again reflecting the decrease in forecaster confidence. Surprisingly, the standard deviations of $r(f_i)$ for the monthly forecasts are smaller than for the seasonal forecasts at the same (or, in the case of precipitation, longer) lead.

Table 1 shows four-parameter summaries for each of the reliability diagrams in Figs. 3 and 4, following Murphy and Wilks (1998). The two parameters b_0 and b_1 are the intercept and slope, respectively, of the weighted least squares lines through the calibration function $q(o_j|f_i)$. While a linear fit is not necessarily the best functional form for this purpose in all cases, these lines do serve both to smooth the sampling variations in the conditional event relative frequencies and to summarize the dominant character of each calibration function. The parameters $\bar{f}$ and s_f are the mean and standard deviation, respectively, of the refinement distributions $r(f_i)$ shown in the insets. Since the sample climatological relative frequencies (i.e., the average observation) can be recovered from these parameters as $\bar{o} = b_0 + b_1\bar{f}$, the unconditional (i.e., overall) bias in each case can be computed as

$$\text{bias} = \bar{f} - \bar{o} = (1 - b_1)\bar{f} - b_0. \qquad (5)$$

Apart from these biases, regression slopes b_1 near 1 (e.g., for monthly temperature forecasts) indicate that aggregate forecaster confidence as reflected by s_f is appropriate for their state of knowledge. Regression slopes substantially greater than 1 (e.g., seasonal forecasts for the warm temperature outcomes) suggest that greater confidence on the forecasters' part is warranted: issuing sharper forecasts through expanded use of more extreme probabilities (increasing s_f) would rotate the calibration function closer to unit slope. Conversely, slopes appreciably smaller than 1 (e.g., for the monthly precipitation forecasts) indicate that the forecasters are overconfident, and that better forecasts overall would result from more conservative deviations from the climatological probability. Regressions with very shallow or negative slopes (e.g., dry seasonal precipitation outcomes at all leads) indicate that the forecast events are not resolved well if at all overall, and thus cannot be considered useful in aggregate.

Also shown in Table 1 are the values of the RES statistic [(4)] for each case. Forecast resolution as reflected by RES can be understood here as a combination of the effects of the regression slope b_1 and the refinement standard deviation s_f. Other things equal, a large b_1 implies greater deviations between the extremes of $q(o_j|f_i)$ and the average observation $\bar{o}$, and thus a large RES. However, the dispersion s_f controls how intensely populated $q(o_j|f_i)$ will be at the extremes (i.e., the magnitudes of n_i for large or small i), so that a large b_1 in combination with a small s_f will result in only a modest RES.

c. Stratified results

It may be of interest for many purposes to see verification results disaggregated more finely than according to lead time (e.g., Livezey 1990). In this section verification statistics for each lead time are presented separately for geographic, seasonal, and ENSO stratifica-

TABLE 2a. Disaggregation of Table 1 according to the geographic stratification defined by the heavy lines in Fig. 1. Temperature forecasts (cool outcome).

Region	Valid	Lead (month)	b_0	b_1	$\bar{f}$	s_f	Bias	RES
NW	Month	0.5	0.09	0.25	0.31	0.049	0.142	0.0025
	Season	0.5	−0.06	0.59	0.31	0.053	0.187	0.0035
	Season	1.5–3.5	−0.15	0.90	0.31	0.042	0.181	0.0061
	Season	4.5–12.5	0.17	−0.25	0.31	0.033	0.218	0.0002
NC	Month	0.5	−0.05	1.13	0.33	0.036	0.007	0.0023
	Season	0.5	−0.11	1.36	0.33	0.055	−0.009	0.0185
	Season	1.5–3.5	0.06	0.94	0.34	0.038	−0.040	0.0087
	Season	4.5–12.5	0.41	−0.23	0.34	0.028	0.008	0.0008
NE	Month	0.5	−0.36	2.00	0.33	0.029	0.030	0.0055
	Season	0.5	−0.06	1.24	0.32	0.040	−0.017	0.0084
	Season	1.5–3.5	−0.03	1.25	0.32	0.037	−0.050	0.0085
	Season	4.5–12.5	0.20	0.52	0.32	0.029	−0.046	0.0017
S	Month	0.5	−0.10	0.97	0.31	0.046	0.109	0.0048
	Season	0.5	0.07	0.32	0.31	0.062	0.141	0.0030
	Season	1.5–3.5	0.01	0.55	0.31	0.051	0.130	0.0020
	Season	4.5–12.5	0.05	0.36	0.31	0.037	0.148	0.0014

tions of the data. In order to conserve space, only the summary parameters b_0, b_1, $\bar{f}$, and s_f, and bias and RES are reported in tabular form. While yet finer divisions of the data (e.g., lead-time, seasonal, and ENSO stratification simultaneously) might also be of interest, it is doubtful whether there is sufficient data available for meaningful analysis, considering the strong space and time correlation present in both forecasts and observations.

Tables 2a and 2b show the results for temperature forecasts from Table 1a, disaggregated according to the geographic stratification indicated by the heavy lines in Fig. 1. These four regions were defined following a cluster analysis of the forecasts (temperature and precipitation forecasts simultaneously), without regard to the corresponding observations. The reliability diagrams for temperature forecasts summarized in Table 2a are similar overall to those stratified only by lead time, although the results for the 1-month forecasts in the northwest region are surprisingly poor in that the summary calibration slope b_1 is small for the cool outcomes and actually negative for the warm outcomes. The standard deviation s_f indicates the least confidence for forecasts pertaining to the northeast region, although for most of these forecasts other than at the longest leads, b_1 is greater than 1, suggesting that this apparent reticence may be misplaced. Forecasts for the southern region generally exhibit the most confidence (largest s_f), but especially for probabilities of the warm outcomes the large slope values indicate that sharper forecasts (even higher confidence) would be warranted.

The precipitation forecasts summarized in Tables 2c and 2d show much stronger geographic differences. In particular, the precipitation forecasts for the north-central and northeastern regions are quite poor (predominantly negative calibration slopes) at all leads, and are probably not useful for any purpose. By contrast, the precipitation forecasts for the northwestern and southern

TABLE 2b. Temperature forecasts (warm outcome).

Region	Valid	Lead (month)	b_0	b_1	$\bar{f}$	s_f	Bias	RES
NW	Month	0.5	0.72	−0.60	0.35	0.049	−0.160	0.0065
	Season	0.5	0.16	1.21	0.36	0.053	−0.236	0.0081
	Season	1.5–3.5	0.05	1.45	0.35	0.042	−0.208	0.0059
	Season	4.5–12.5	0.67	−0.21	0.35	0.033	−0.246	0.0014
NC	Month	0.5	0.02	0.96	0.33	0.036	−0.007	0.0028
	Season	0.5	−0.50	2.46	0.33	0.055	0.018	0.0259
	Season	1.5–3.5	−0.52	2.37	0.33	0.038	0.068	0.0185
	Season	4.5–12.5	0.12	0.67	0.33	0.028	−0.011	0.0025
NE	Month	0.5	−0.44	2.49	0.34	0.029	−0.067	0.0088
	Season	0.5	0.16	0.83	0.35	0.040	−0.100	0.0052
	Season	1.5–3.5	0.02	1.11	0.35	0.037	−0.058	0.0070
	Season	4.5–12.5	0.18	0.74	0.34	0.030	−0.092	0.0022
S	Month	0.5	−0.10	0.97	0.31	0.046	−0.170	0.0081
	Season	0.5	0.07	0.32	0.31	0.062	−0.213	0.0114
	Season	1.5–3.5	0.01	0.55	0.31	0.051	−0.189	0.0101
	Season	4.5–12.5	0.05	0.36	0.31	0.037	−0.220	0.0077

TABLE 2c. Precipitation forecasts (dry outcome).

Region	Valid	Lead (month)	b_0	b_1	$\bar{f}$	s_f	Bias	RES
NW								
	Month	0.5	0.00	0.61	0.33	0.035	0.129	0.0013
	Season	0.5–2.5	−0.29	1.40	0.33	0.034	0.158	0.0044
	Season	3.5–6.5	−0.23	1.25	0.33	0.019	0.148	0.0026
	Season	7.5–12.5	0.15	0.08	0.34	0.016	0.163	0.0009
NC	Month	0.5	0.28	−0.07	0.33	0.034	0.073	0.0023
	Season	0.5–2.5	0.19	0.05	0.33	0.034	0.124	0.0026
	Season	3.5–6.5	0.72	−1.53	0.33	0.015	0.115	0.0012
	Season	7.5–12.5	0.35	−0.43	0.33	0.012	0.122	0.0004
NE	Month	0.5	0.61	−0.99	0.33	0.018	0.047	0.0016
	Season	0.5–2.5	0.71	−1.39	0.34	0.033	0.103	0.0048
	Season	3.5–6.5	1.18	−2.85	0.33	0.017	0.090	0.0038
	Season	7.5–12.5	0.78	−1.64	0.33	0.010	0.091	0.0011
S								
	Month	0.5	−0.07	1.16	0.33	0.049	0.017	0.0072
	Season	0.5–2.5	0.04	0.86	0.31	0.067	0.003	0.0052
	Season	3.5–6.5	0.08	0.73	0.32	0.030	0.006	0.0017
	Season	7.5–12.5	0.12	0.52	0.33	0.016	0.038	0.0003

regions show quite good resolution even through the 3.5–6.5-month lead, although the dry bias evident in the spatially aggregated results is still present, especially in the northwest. The precipitation forecasts for the southern region exhibit the highest confidence (s_f), but for the wet outcome probabilities even sharper forecasts would be warranted.

Table 3 shows disaggregation of the summary reliability diagram parameters according to winter [December–February for monthly forecasts, and NDJ–JFM for seasonal forecasts], spring (March–May, and February–April to April–June), summer (June–August, and May–July to July–September, and fall (September–November, and August–October to October–December) seasons. For the temperature forecasts in Tables 3a and 3b these are broadly similar to the aggregated results in Table 1a, although the winter 1-month temperature forecasts exhibit surprisingly weak calibration. Generally the forecasts for summer and fall show lower confidence, although particularly for the warm-outcome probabilities the large regression slopes b_1 suggest that greater confidence would be justified. The overall cold bias is evident in all seasons except spring, where the relative frequencies of the cold, near-normal, and warm outcomes for 1995–98 are fairly close to the climatological distributions defined by the 1961–90 normals.

The seasonally stratified precipitation results in Tables 3c and 3d are again notably different from the aggregated results in Table 1b. All seasons again show a strong dry bias. The results for 1-month and 0.5–3.5-month lead seasonal forecasts in winter show generally good resolution. The results for summer precipitation forecasts at all leads indicate quite strong resolution, which would justify much more frequent use of probabilities away from the climatological ⅓, even though forecasters were evidently least confident about precip-

TABLE 2d. Precipitation forecasts (wet outcome).

Region	Valid	Lead (month)	b_0	b_1	$\bar{f}$	s_f	Bias	RES
NW	Month	0.5	0.30	0.52	0.34	0.035	−0.137	0.0033
	Season	0.5–2.5	0.01	1.71	0.34	0.034	−0.251	0.0084
	Season	3.5–6.5	−0.11	2.01	0.33	0.019	−0.443	0.0051
	Season	7.5–12.5	0.73	−0.52	0.33	0.016	−0.228	0.0030
NC	Month	0.5	0.86	−1.28	0.33	0.034	−0.952	0.0063
	Season	0.5–2.5	0.27	0.61	0.34	0.035	−0.137	0.0058
	Season	3.5–6.5	1.05	−1.73	0.33	0.015	−0.149	0.0031
	Season	7.5–12.5	1.20	−2.18	0.33	0.012	−0.151	0.0013
NE	Month	0.5	0.39	0.04	0.33	0.018	−0.073	0.0022
	Season	0.5–2.5	1.04	−1.84	0.33	0.033	−0.103	0.0056
	Season	3.5–6.5	1.52	−3.25	0.33	0.017	−0.118	0.0051
	Season	7.5–12.5	1.69	−3.72	0.33	0.010	−0.132	0.0016
S	Month	0.5	0.01	1.13	0.34	0.050	−0.054	0.0082
	Season	0.5–2.5	−0.06	1.39	0.35	0.068	−0.076	0.0123
	Season	3.5–6.5	−0.37	2.30	0.34	0.030	−0.072	0.0058
	Season	7.5–12.5	−0.16	1.78	0.34	0.016	−0.105	0.0021

TABLE 3a. Disaggregation of Table 1 according to seasons. Temperature forecasts (cool outcome).

Season	Valid	Lead (month)	b_0	b_1	$\bar{f}$	s_f	Bias	RES
Winter	Month	0.5	−0.09	0.56	0.32	0.041	0.231	0.0016
	Season	0.5	−0.05	0.54	0.30	0.065	0.188	0.0103
	Season	1.5–3.5	−0.05	0.60	0.31	0.041	0.174	0.0036
	Season	4.5–12.5	−0.03	0.48	0.32	0.028	0.196	0.0004
Spring	Month	0.5	−0.14	1.75	0.31	0.052	−0.092	0.0121
	Season	0.5	0.04	0.98	0.31	0.064	−0.034	0.0171
	Season	1.5–3.5	0.07	0.95	0.31	0.057	−0.054	0.0124
	Season	4.5–12.5	0.03	1.05	0.31	0.040	−0.046	0.0062
Summer	Month	0.5	−0.35	1.92	0.32	0.038	0.056	0.0101
	Season	0.5	−0.01	0.78	0.32	0.044	0.080	0.0059
	Season	1.5–3.5	−0.13	1.17	0.32	0.035	0.076	0.0031
	Season	4.5–12.5	0.10	0.43	0.32	0.031	0.082	0.0011
Fall	Month	0.5	−0.20	1.33	0.33	0.024	0.091	0.0030
	Season	0.5	−0.50	2.23	0.33	0.036	0.094	0.0096
	Season	1.5–3.5	−0.64	2.77	0.33	0.033	0.056	0.0164
	Season	4.5–12.5	−0.36	1.82	0.33	0.030	0.089	0.0054

itation for this season. Results for spring precipitation forecasts are similar to the aggregated results, while the fall precipitation forecasts are uniformly poor and apparently not useful.

Last, Table 4 stratifies the reliability diagram summaries according to the temperature anomaly in the Niño-3.4 region of the eastern tropical Pacific during the valid month or season, with anomalies less than −0.4°C regarded as cold, and anomalies greater than +0.4°C regarded as warm. Again the reliability is compromised in all cases by the cold bias evident in the aggregated results, but good resolution is maintained in all cases through the 1.5–3.5-month lead. Forecaster confidence as reflected by s_f is generally highest for the warm Niño-3.4 temperatures and least for the near-normal Niño-3.4 temperatures, which is consistent with the expectation that much of the skill at these leads derives from forecasts of the ENSO phenomenon. However, the quite steep calibration slopes at all leads during near-normal Niño-3.4 months and seasons indicates that sharper temperature forecasts in these cases would be well justified.

The precipitation results in Tables 4c and 4d suggest that a large portion of the aggregate ability to forecast precipitation is contributed by the warm subset. In contrast, all calibration slopes for cool Niño-3.4 cases are negative. Precipitation forecasts during neutral Niño-3.4 temperature conditions are modestly successful through the 0.5–2.5-month lead period, although as before their reliability is compromised by the dry bias resulting from comparatively wet conditions during 1995–98.

5. Post hoc recalibrations

The results presented in section 4 show that in many cases the long-lead CPC forecasts are able to resolve in advance subsets of the valid periods with outcome relative frequencies that are quite different from the (both 1961–90, and sample) climatological values. While event resolution is a necessary condition for users to be

TABLE 3b. Temperature forecasts (warm outcome).

Season	Valid	Lead (month)	b_0	b_1	$\bar{f}$	s_f	Bias	RES
Winter	Month	0.5	0.51	0.12	0.35	0.041	−0.202	0.0041
	Season	0.5	0.25	1.11	0.36	0.065	−0.290	0.0117
	Season	1.5–3.5	−0.02	1.71	0.35	0.042	−0.228	0.0073
	Season	4.5–12.5	0.38	0.70	0.35	0.028	−0.275	0.0030
Spring	Month	0.5	0.08	0.78	0.36	0.052	−0.001	0.0044
	Season	0.5	−0.28	1.77	0.35	0.064	0.010	0.0227
	Season	1.5–3.5	−0.34	1.92	0.35	0.057	0.018	0.0186
	Season	4.5–12.5	−0.41	2.17	0.36	0.040	−0.011	0.0105
Summer	Month	0.5	−0.33	2.25	0.34	0.038	−0.095	0.0130
	Season	0.5	−0.34	2.36	0.34	0.044	−0.122	0.0147
	Season	1.5–3.5	−0.40	2.51	0.34	0.036	−0.113	0.0111
	Season	4.5–12.5	−0.31	2.28	0.34	0.032	−0.125	0.0066
Fall	Month	0.5	−0.36	2.30	0.33	0.024	−0.069	0.0082
	Season	0.5	−0.46	2.85	0.33	0.036	−0.150	0.0127
	Season	1.5–3.5	−0.59	2.99	0.33	0.033	−0.067	0.0152
	Season	4.5–12.5	−0.23	2.15	0.33	0.030	−0.150	0.0072

TABLE 3c. Precipitation forecasts (dry outcome).

Season	Valid	Lead (month)	b_0	b_1	$\bar{f}$	s_f	Bias	RES
Winter	Month	0.5	−0.07	1.03	0.33	0.047	0.060	0.0049
	Season	0.5–2.5	0.04	0.42	0.32	0.075	0.146	0.0053
	Season	3.5–6.5	0.38	−0.61	0.33	0.032	0.151	0.0029
	Season	7.5–12.5	0.53	−1.09	0.33	0.022	0.160	0.0017
Spring	Month	0.5	0.05	0.66	0.33	0.028	0.062	0.0022
	Season	0.5–2.5	0.15	0.25	0.32	0.050	0.090	0.0025
	Season	3.5–6.5	0.34	−0.29	0.33	0.029	0.086	0.0015
	Season	7.5–12.5	0.90	−1.97	0.33	0.010	0.080	0.0008
Summer	Month	0.5	−0.51	2.42	0.33	0.021	0.041	0.0057
	Season	0.5–2.5	−0.36	1.92	0.33	0.013	0.056	0.0013
	Season	3.5–6.5	−0.01	0.85	0.33	0.010	0.060	0.0003
	Season	7.5–12.5	−1.26	4.60	0.33	0.007	0.072	0.0016
Fall	Month	0.5	0.46	−0.61	0.33	0.043	0.071	0.0046
	Season	0.5–2.5	0.49	−0.69	0.33	0.033	0.068	0.0023
	Season	3.5–6.5	0.95	−2.08	0.33	0.013	0.066	0.0015
	Season	7.5–12.5	0.23	0.03	0.33	0.013	0.090	0.0006

able to derive value from forecasts, it is by no means sufficient. In particular, the presence of strong cool and dry biases in the CPC long-lead forecasts imply that these forecasts were miscalibrated, and that taking the forecast probabilities at face value would not be justified. The problem is analogous to assessing the probabilities of outcomes in a hypothetical dice game in which one of two six-sided dice is normal, but the other has two ''6s'' but no ''3.'' A gambler (user) would not be well served by probability calculations (forecasts) based on the assumption that both dice are normal and fair (the forecasts are unbiased), because the average throw will sum to 7.5 rather than 7 (the average temperature or precipitation is higher than in the 1961–90 base period). The reliability, or calibration, of the CPC forecasts is further degraded in some cases by conditional biases, which reflect their being either overconfident (regression slopes $b_1 < 1$) or underconfident ($b_1 > 1$), apart from any overall, unconditional bias.

Retrospectively, one can define transformations of the forecast probabilities that yield unbiased forecasts with maximum correspondence between forecasts and the subsequent event probabilities, at least for the data sample at hand. If it could be assumed that future CPC long-lead forecasts would have the same properties as those from 1995 to 1998, then these transformations would generally lead to a substantial enhancement of forecast value for those future forecasts (analogous to informing the gambler about the nature of the nonstandard die). Alternatively, such recalibration functions could provide useful guidance to the forecasters themselves in producing better-calibrated forecasts in the future.

A simple adjustment of this kind is the linear transformation

$$f_{\text{adj}} = \alpha(f - \bar{f}) + \bar{o}, \tag{6}$$

where the scaling parameter α either expands or contracts the dispersion of the forecasts f around the mean $\bar{f}$ of the distribution $r(f_i)$, and $\bar{o}$ is the sample climatological relative frequency. The transformation in (6)

TABLE 3d. Precipitation forecasts (wet outcome).

Season	Valid	Lead (month)	b_0	b_1	$\bar{f}$	s_f	Bias	RES
Winter	Month	0.5	−0.16	1.76	0.34	0.047	−0.098	0.0094
	Season	0.5–2.5	0.21	0.94	0.35	0.076	−0.189	0.0141
	Season	3.5–6.5	0.41	0.38	0.34	0.032	−0.199	0.0067
	Season	7.5–12.5	0.94	−1.21	0.34	0.022	−0.189	0.0019
Spring	Month	0.5	0.31	0.36	0.34	0.028	−0.092	0.0032
	Season	0.5–2.5	0.20	0.86	0.34	0.050	−0.152	0.0097
	Season	3.5–6.5	0.12	1.02	0.34	0.029	−0.127	0.0039
	Season	7.5–12.5	0.50	−0.14	0.33	0.010	−0.124	0.0012
Summer	Month	0.5	−0.45	2.49	0.33	0.021	−0.042	0.0058
	Season	0.5–2.5	−0.42	2.56	0.34	0.013	−0.110	0.0022
	Season	3.5–6.5	−0.27	2.08	0.33	0.011	−0.086	0.0007
	Season	7.5–12.5	−2.08	7.56	0.33	0.007	−0.085	0.0039
Fall	Month	0.5	1.04	−1.87	0.33	0.043	−0.093	0.0125
	Season	0.5–2.5	0.82	−1.13	0.33	0.033	−0.117	0.0036
	Season	3.5–6.5	1.27	−2.45	0.34	0.013	−0.097	0.0040
	Season	7.5–12.5	1.16	−2.08	0.34	0.013	−0.113	0.0013

TABLE 4a. Disaggregation of Table 1 according to Niño-3.4 temperature anomalies at forecast valid time. Temperature forecasts (cool outcome).

Niño-3.4	Valid	Lead (month)	b_0	b_1	$\bar{f}$	s_f	Bias	RES
Cool	Month	0.5	−0.23	1.23	0.33	0.034	0.154	0.0036
	Season	0.5	−0.14	0.99	0.32	0.041	0.143	0.0055
	Season	1.5–3.5	0.01	0.68	0.32	0.036	0.092	0.0032
	Season	4.5–12.5	0.26	−0.27	0.33	0.033	0.159	0.0016
Normal	Month	0.5	−0.05	1.15	0.32	0.035	0.002	0.0042
	Season	0.5	−0.61	2.84	0.32	0.035	0.021	0.0138
	Season	1.5–3.5	−0.60	2.81	0.32	0.034	0.021	0.0148
	Season	4.5–12.5	−0.56	2.61	0.32	0.031	0.045	0.0111
Warm	Month	0.5	−0.19	1.41	0.32	0.052	0.059	0.0075
	Season	0.5	0.04	0.57	0.31	0.077	0.093	0.0082
	Season	1.5–3.5	0.05	0.61	0.32	0.057	0.075	0.0047
	Season	4.5–12.5	0.14	0.35	0.32	0.036	0.068	0.0009

includes both an overall bias adjustment to the extent that $\overline{f} \neq \overline{o}$, and a correction for forecaster over- ($\alpha < 1$) or under- ($\alpha > 1$) confidence that yields slope parameters b_1 for the adjusted forecasts that are near 1. Table 5 shows values of the scaling parameter α, obtained by minimizing the REL statistic (Murphy 1973),

$$\text{REL} = \frac{1}{n}\sum_{i=1}^{I} n_i[q(o_j|\, f_i) - f_i]^2. \qquad (7)$$

Retrospective application of (6) to the forecast data yields reliability diagrams similar to those in Figs. 3 and 4, except that the points are translated either left or right to correct the unconditional bias, and are either dilated or shrunk relative to the mean forecast to alleviate the conditional biases. The particular sampling variations (excursions around the fitted regression lines) evident in Figs. 3 and 4 will be retained, but scatter around the 1:1 line. For example, Fig. 5 shows the reliability diagrams obtained from the recalibrated seasonal temperature forecasts for the 1.5–3.5-month leads (cf. top middle panel of Fig. 4). It is important to emphasize that because the values α, $\overline{f}$, and $\overline{o}$ have been derived from the samples that are themselves being recalibrated, these results are better than could be obtained using independent data. However, these recalibrations will approximate the characteristics of the best forecasts of this kind that could be formulated given the present level of scientific knowledge.

Recall that results for probability forecasts of the near-normal outcomes have not been reported in the analyses above, primarily because the overwhelming majority of the forecasts involve adjustments only to probabilities for the above- and below-normal categories, as a consequence of the convention adopted within CPC for the forecast format. Furthermore, in the few cases where CPC has forecast increased probability of the near-normal class, it has done so with poor reliability and resolution, and over a very restricted range of probabilities. Because the probabilities for the below-, near-, and above-normal categories must sum to 1 for any forecast, applying adjustments such as (6) to both the below- and above-normal probabilities will often imply probabilities for the near-normal category that are different from the climatological ⅓. Figure 6 shows reliability diagrams for the forecasts of near-normal seasonal temperatures that are implied by recalibration of the below- and above-normal forecasts using (6) and the parameters in Tables 1 and 5. Here the symbols ''1,'' ''2,'' and ''3'' indicate the 0.5-, 1.5–3.5-, and 4.5–12.5-month leads, respectively, and the line weights have the same meanings as in the foregoing figures. The range of probabilities for the near-normal temperature out-

TABLE 4b. Temperature forecasts (warm outcome).

Niño-3.4	Valid	Lead (month)	b_0	b_1	$\bar{f}$	s_f	Bias	RES
Cool	Month	0.5	0.08	1.34	0.34	0.034	−0.196	0.0071
	Season	0.5	0.29	0.97	0.34	0.041	−0.280	0.0069
	Season	1.5–3.5	0.32	0.64	0.34	0.036	−0.198	0.0036
	Season	4.5–12.5	0.73	−0.27	0.34	0.033	−0.298	0.0074
Normal	Month	0.5	−0.27	1.78	0.35	0.035	−0.003	0.0102
	Season	0.5	−0.70	3.11	0.35	0.035	−0.038	0.0136
	Season	1.5–3.5	−0.66	2.99	0.35	0.035	−0.036	0.0130
	Season	4.5–12.5	−0.70	3.17	0.34	0.032	−0.038	0.0124
Warm	Month	0.5	0.09	1.01	0.35	0.052	−0.094	0.0055
	Season	0.5	−0.15	1.81	0.35	0.076	−0.134	0.0258
	Season	1.5–3.5	−0.34	2.29	0.35	0.057	−0.112	0.0207
	Season	4.5–12.5	−0.34	2.30	0.35	0.036	−0.115	0.0116

TABLE 4c. Precipitation forecasts (dry outcome).

Niño-3.4	Valid	Lead (month)	b_0	b_1	$\bar{f}$	s_f	Bias	RES
Cool	Month	0.5	0.32	−0.02	0.34	0.046	0.027	0.0054
	Season	0.5–2.5	0.84	−1.51	0.33	0.038	−0.012	0.0103
	Season	3.5–6.5	2.38	−6.20	0.33	0.016	−0.004	0.0132
	Season	7.5–12.5	1.80	−4.48	0.33	0.017	0.008	0.0080
Normal	Month	0.5	−0.25	1.53	0.33	0.012	0.075	0.0011
	Season	0.5–2.5	−0.14	1.04	0.33	0.017	0.127	0.0019
	Season	3.5–6.5	0.18	0.09	0.33	0.014	0.120	0.0009
	Season	7.5–12.5	−0.14	1.00	0.33	0.012	0.140	0.0017
Warm	Month	0.5	−0.09	1.00	0.32	0.041	0.090	0.0026
	Season	0.5–2.5	0.03	0.54	0.32	0.071	0.117	0.0024
	Season	3.5–6.5	−0.04	0.74	0.33	0.032	0.126	0.0011
	Season	7.5–12.5	−0.68	2.67	0.33	0.014	0.129	0.0017

come here is much greater than in the original forecasts, and the good reliability indicates that the separate recalibrations for the below- and above-normal outcomes are reasonably consistent with one another.

6. Summary and conclusions

This paper has presented a diagnostic verification analysis of the CPC long-lead forecasts for the years 1995–98. These forecasts were found to successfully resolve different relative frequencies of the temperature outcomes throughout the full range of lead times, which extend to more than a year. As would be expected, precipitation events are less well resolved, and only precipitation forecasts with lead times of 2.5 months or less could be considered successful in aggregate. The most prominent sources of forecast inaccuracy were the quite strong (unconditional) forecast biases that may have resulted from the 1995–98 period being substantially warmer and wetter than the 1961–90 reference period. Less prominent but still important in some cases were conditional biases, or the systematic miscalibration of the probability forecasts after accounting for differences between the average forecast and average observation. A much broader range of probabilities are used, with good event resolution and at much longer lead times, than in the earlier generation of these forecasts (Murphy and Huang 1991), indicating substantial improvement from the previous forecasts.

Stratification of the forecasts according to season, geographic region, and tropical Pacific sea surface temperature yielded differences that were most prominent for the precipitation forecasts. In particular, precipitation forecasts for all lead times were quite poor, and probably not useful in aggregate, for the north-central and northeastern regional stratification, the fall seasonal stratification, and the cool Niño-3.4 temperature stratification. Perhaps surprisingly, temperature forecasts for periods with near-normal Niño-3.4 conditions show quite strong resolution for both warm and cool forecasts at all lead times, suggesting that forecasters would be justified in being less conservative regarding temperature probabilities in such conditions.

The utility of a diagnostic verification exercise derives from its ability to identify particular aspects of the forecasts, both good and bad, that contribute to overall performance. After the strong unconditional biases noted above, the most important weakness of these forecasts are the conditional biases manifested by weighted-regression slopes b_1 different from 1, and deriving fundamentally from systematic over- or underconfidence for particular predictands and lead times. Although cor-

TABLE 4d. Precipitation forecasts (wet outcome).

Niño-3.4	Valid	Lead (month)	b_0	b_1	$\bar{f}$	s_f	Bias	RES
Cool								
	Month	0.5	0.59	−0.59	0.33	0.046	−0.065	0.0068
	Season	0.5–2.5	0.81	−1.25	0.33	0.037	−0.068	0.0089
	Season	3.5–6.5	2.14	−5.15	0.34	0.016	−0.049	0.0094
	Season	7.5–12.5	2.01	−4.77	0.33	0.017	−0.106	0.0086
Normal	Month	0.5	−0.23	1.97	0.34	0.012	−0.100	0.0012
	Season	0.5–2.5	0.23	0.82	0.34	0.017	−0.169	0.0031
	Season	3.5–6.5	0.82	−0.99	0.34	0.014	−0.143	0.0015
	Season	7.5–12.5	0.86	−1.06	0.33	0.012	−0.180	0.0029
Warm	Month	0.5	−0.05	1.43	0.34	0.041	−0.096	0.0045
	Season	0.5–2.5	0.10	1.13	0.35	0.072	−0.146	0.0120
	Season	3.5–6.5	−0.25	2.19	0.34	0.033	−0.155	0.0071
	Season	7.5–12.5	−0.91	4.20	0.33	0.014	−0.146	0.0047

TABLE 5. Adjustment parameters α for (6), that minimize the resulting REL statistic [(7)].

Predictand	Valid	Lead	α (below normal)	α (above normal)
Temperature	Month	0.5	1.6	1.4
Temperature	Season	0.5	1.2	2.1
Temperature	Season	1.5–3.5	1.4	2.6
Temperature	Season	4.5–12.5	0.9	1.8
Precipitation	Month	0.5	0.9	0.8
Precipitation	Season	0.5–2.5	0.5	0.8

relation measures relating forecasts and observations are sometimes used in forecast verification, such statistics are blind to both the unconditional and systematic conditional biases that are prominent features of these forecasts, and thus would portray potential rather than actual skill (Murphy and Epstein 1989; Murphy 1995). The use of correlation measures for verification of the present forecasts would be completely uninformative about the two primary contributions to error (and opportunities for their improvement). Indeed, for these data, correlation scores would measure primarily the lack of fit of the weighted regression lines in Figs. 3 and 4 and in Tables 1–4. In contrast, scalar verification measures such as the ranked probability score (Fig. 2) reflect the forecast biases very strongly and thus portray an unduly pessimistic picture of forecast performance.

In order for decision makers to make best rational use of forecasts, it is essential that outcome probabilities corresponding to each possible forecast be known, at least approximately. For reliable, or well-calibrated, probability forecasts it is possible to take the forecasts at face value: close correspondence between the forecasts f_i and the conditional outcome relative frequencies $q(o_j | f_i)$ [cf. (7)] implies that the forecasts ''mean what they say.'' Direct use of interpretive procedures based on the forecasts (e.g., Briggs and Wilks 1996a,b; Croley 1996) assumes good reliability. The previous generation of these forecasts possessed this property (Murphy and Huang 1991), although over a much more restricted range of forecast probabilities. The two major sources of bias in the current forecasts both contribute to degrading their reliability. Forecasters at CPC have recognized the unconditional bias ($\overline{f} \neq \overline{o}$) problem and have taken steps to incorporate changing climatic mean conditions into the forecast construction process (Monastersky 1999; R. E. Livezey 1999, personal communication). The second source of bias is conditional; that is, the tendency to underforecast large probabilities and overforecast small probabilities in situations where the forecasters could be more confident, or the tendency to overforecast large probabilities and underforecast small probabilities when more conservative probability assessment would be beneficial. It is hoped that identification of these biases will help to refine forecast performance in the future (cf. Murphy and Daan 1984).

Last, a simple recalibration scheme was introduced

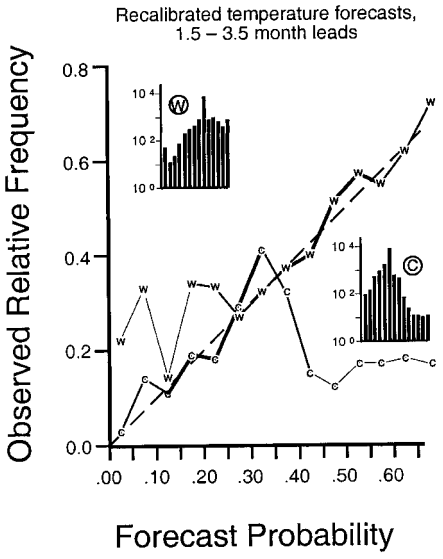

FIG. 5. Reliability diagrams for recalibrated temperature probability forecasts for the 1.5–3.5-month leads.

that corrects the unconditional and conditional biases based on in-sample statistics. Given that, at minimum, the unconditional biases have been recognized and their correction is being attempted by CPC, it should not be expected that these corrections would be valid for future forecasts. They might be useful, however, in retrospective studies of potential forecast value (Wilks 1997). To the extent that the CPC forecasters could successfully incorporate adjustments to future forecasts in the spirit of those in Table 5, which imply well-calibrated forecasts over a broad range of probabilities for near-normal

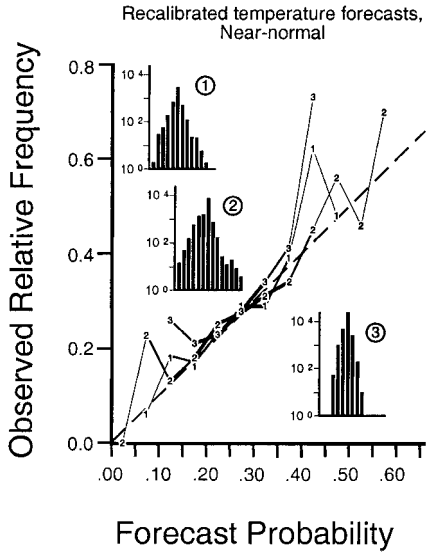

FIG. 6. Reliability diagrams for recalibrated probability forecasts for near-normal seasonal temperature outcomes. Here ''1'' indicates 0.5-month leads, ''2'' indicates 1.5–3.5-month leads, and ''3'' indicates 4.5–12.5-month leads.

conditions (Fig. 6), it might well be possible to relax the current convention of enforcing the climatological probability for the near-normal category in most cases.

Acknowledgments. I thank Dave Unger of the NOAA Climate Prediction Center for providing the archived forecasts and corresponding verifying observations, and Bob Livezey for useful discussions. This work was supported by the NOAA Economics and Human Dimensions program, under Grant NA86GP0555.

REFERENCES

Briggs, W. M., and D. S. Wilks, 1996a: Estimating monthly and seasonal distributions of temperature and precipitation using the new CPC long-range forecasts. *J. Climate,* **9,** 818–826.

——, and ——, 1996b: Extension of the CPC long-lead temperature and precipitation outlooks to general weather statistics. *J. Climate,* **9,** 3496–3504.

Croley, T. E., II, 1996: Using NOAA's new climate outlooks in operational hydrology. *J. Hydrol. Eng.,* **1,** 93–102.

Dischel, B., 1998: The fledgling weather market takes off. Part 1: Weather sensitivity, weather derivatives and a pricing model. *Applied Derivatives Trading.* [Available online at http://www.adtrading.com/adt32/weather1.hts.]

Draper, N. R., and H. Smith, 1981: *Applied Regression Analysis.* John Wiley and Sons, 790 pp.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

——, 1988: Long-range weather prediction: Limits of predictability and beyond. *Wea. Forecasting,* **3,** 69–75.

Katz, R. W., and A. H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts.* Cambridge University Press, 222 pp.

Kim, Y.-O., and R. N. Palmer, 1997: Value of seasonal flow forecasts in Bayesian stochastic programming. *J. Water Res. Planning Management,* **123,** 327–335.

Knox, J. B., H. Moses, and M. C. MacCracken, 1985: Summary report of the workshop on the interactions of climate and energy. *Bull. Amer. Meteor. Soc.,* **66,** 174–183.

Livezey, R. E., 1990: Variability of skill of long-range forecasts and implications for their use and value. *Bull. Amer. Meteor. Soc.,* **71,** 300–309.

Mjelde, J. W., H. S. J. Hill, and J. F. Griffiths, 1998: A review of current evidence on climate forecasts and their economic effects in agriculture. *Amer. J. Agric. Econ.,* **80,** 1089–1095.

Monastersky, R., 1999: When meteorologists see red. *Sci. News,* **155,** 188–192.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

——, 1994: Assessing the economic value of weather forecasts: An overview of methods, results and issues. *Meteor. Appl.,* **1,** 69–73.

——, 1995: The coefficients of correlation and determination as measures of performance in forecast verification. *Wea. Forecasting,* **10,** 681–688.

——, 1997: Forecast verification. *Economic Value of Weather and Climate Forecasts,* R. W. Katz and A. H. Murphy, Eds., Cambridge University Press, 19–74.

——, and H. Daan, 1984: Impacts of feedback and experience on the quality of subjective probability forecasts: Comparison of results from the first and second years of the Zierikzee experiment. *Mon. Wea. Rev.,* **112,** 413–423.

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

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

——, and J. Huang, 1991: On the quality of CAC's probabilistic 30-day and 90-day forecasts. *Proc. 16th Annual Climate Diagnostics Workshop,* Los Angeles, CA, Amer. Meteor. Soc., 390–399.

——, and R. L. Winkler, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting,* **7,** 435–455.

——, and D. S. Wilks, 1998: A case study in the use of statistical models in forecast verification: Precipitation probability forecasts. *Wea. Forecasting,* **13,** 795–810.

——, B. G. Brown, and Y. S. Chen, 1989: Diagnostic verification of temperature forecasts. *Wea. Forecasting,* **4,** 485–501.

Namias, J., 1968: Long-range weather forecasting—History, current status and outlook. *Bull. Amer. Meteor. Soc.,* **49,** 438–470.

O'Lenic, E., 1994: Operational long-lead forecast for the climate outlook. Technical Procedures Bulletin 418, NOAA/NWS/CPC, 30 pp. [Available from NOAA/CPC, 5200 Auth Rd., Camp Springs, MD 20746.]

Wagner, A. J., 1989: Medium- and long-range forecasting. *Wea. Forecasting,* **4,** 413–426.

Wilks, D. S., 1990: Maximum likelihood estimation for the gamma distribution using data containing zeros. *J. Climate,* **3,** 1495–1501.

——, 1995: *Statistical Methods in the Atmospheric Sciences: An introduction.* International Geophysics Series, Vol. 59, Academic Press, 464 pp.

——, 1997: Forecast value: Prescriptive decision studies. *Economic Value of Weather and Climate Forecasts,* R. W. Katz and A. H. Murphy, Eds., Cambridge University Press, 109–145.

Winkler, R. L., and A. H. Murphy, 1985: Decision analysis. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview, 493–524.