Long-Range Weather Prediction: Limits of Predictability and Beyond*

EDWARD S. EPSTEIN

Climate Analysis Center, National Meteorological Center, National Weather Service, Washington, DC

(Manuscript received 4 August 1987, in final form 18 December 1987)

ABSTRACT

The details of the weather are not predictable beyond one to two weeks. At longer time ranges, averages of the weather over space and time can be usefully predicted only to the extent that the variations of the averages exceed the "noise" produced by the omnipresent but unpredictable transient weather. This margin of potential predictability is not large, but parts of it are being exploited in routinely issued monthly and seasonal forecasts. The format and utilization of these forecasts, the methods by which they are routinely produced, and prospects for improvements are discussed.

1. Predictability and the meaning of "long range"

Long-range prediction begins, by definition,[1] where synoptic predictability ends. By that same definition long-range prediction cannot deal with the details or the instantaneous states of the atmosphere, but must deal with collective information—such as means and variances. In a statistical sense, long-range *forecasts* necessarily deal with population values (expected values), but the *event* against which every forecast is verified is a single manifestation, a sample of one which includes the "noise" of the composite of unpredictable synoptic disturbances that actually occur. These unpredictable disturbances limit, at times substantially, the accuracy a long-range forecast can possibly have (Madden, 1976; Madden and Shea, 1978).

Actually, the situation is even worse. The "climatology" of the moment—the composite defining the population from which the instantaneous states are drawn—is itself interactive with the instantaneous states. Mixed in with the factors like variations in sea surface temperatures, soil moisture, snow cover, etc., that make the "climatology" differ from time to time (and therefore favor long-range predictability) are the net exchanges of heat, moisture, and momentum brought about by the synoptic waves whose details are not predictable.

Against this background of intrinsically limited potential skill, several groups attempt long-range forecasting. What follows is a brief description of the approach, methods, skill and prospects for improvement of the long-range forecasts issued by the Climate Analysis Center, an element of the National Weather Service's National Meteorological Center.

2. Forecast format and interpretation

The potential for skill in long-range prediction necessarily is very limited, but some potential does exist. We cannot expect—indeed we should reject as deceitful—forecasts claiming great skill or precision. But the value of forecasts of even meager skill can be great if the limited reliability of the forecast is made explicit by the forecaster and explicitly appreciated by the decision maker (e.g., Brown et al., 1986). Decisions such as how much fuel a public utility should store, how many or what mix of seasonal products to manufacture or stock, and what crops if any to plant, are examples of the type of significant economic decision that can hinge on modest variations in the probability of warmer or colder, wetter or drier than normal months or seasons ahead.

The primary format for the National Weather Service's long-range (monthly and seasonal) outlooks are probability statements defining the likelihood that the mean temperature or total precipitation will exceed or fall short of its climatological expectation. We define "warm" or "heavy" to include the highest 30% of the climatological frequency distribution for mean temperature or total precipitation, respectively, for any location and time of year; similarly "cold" or "light" include the lowest 30%, and "near normal" or "mod-

* Presented at the Seventh International Symposium on Forecasting, Boston, 27–29 May 1987.

[1] According to one authoritative definition (WMO, 1984), "long-range forecasting (LRF) is concerned with the prediction of the general behaviour of the larger scales of atmospheric circulation and weather over 'typically' a month or a season. The lower bound of LRF is currently provided by the practical limit of about 1–2 weeks to the predictability of individual synoptic-scale daily weather systems."

Corresponding author address: Dr. Edward S. Epstein, NOAA/NWS/Climate Analysis Center, National Meteorological Center, Washington, DC 20233.

erate" include the middle 40% of the climatological frequency distribution. From experience we have learned that we can identify situations when the likelihood of noticeably abnormal conditions is somewhat greater or less than their climatological frequency of 30%, but we are unable to distinguish situations when the likelihood of the central categories differs from 40%. Thus the forecasts always carry the implicit message that the probability of near normal temperatures or moderate precipitation is 40%. The forecast maps that are issued (Fig. 1) indicate which of the two outer categories (e.g., warm or cold) is favored and what probability is assigned to that category. For example, if the probability of "warm" is 40%, then the probability of "cold" is, by inference, 20%: 0.4 + 0.4 + 0.2 = 1.0. (Variants on these charts, lacking the specific probabilistic information, often appear in the press and are available at local NWS offices. These are derived from the charts described here.)

The experience of being unable to distinguish situations in which the central category is more or less likely than its climatological frequency of 0.4 is quite consistent with the range of probabilities we have been able to assign reliably to the outer categories. If we represent the climatological distribution of, say, monthly mean temperatures by a normal distribution, as in Fig. 2a (not a bad approximation in general), then the interval that includes the middle 40% of the distribution is the mean ± 0.524 standard deviations. (Typically the standard deviation of monthly mean temperature is on the order of 3° or 4°F.) If we associate, with the conditional probability distribution of mean temperatures given the forecast, a normal distribution with the same standard deviation but a different mean, then a 40% probability for warm implies a shift of the mean by 0.27 standard deviations (see Fig. 2b). Such a shift also implies that the conditional probability of cold is 0.21, leaving a probability of 0.39 for the central, near normal category. Thus, the experience of not being able to alter the probability assigned to the middle category, while making relatively strong statements about the probabilities of the outer categories, is consistent with this model. This result (a 40% forecast statement implies a shift of the climatological distribution of about 0.27 standard deviations) further implies, in view of the limited strength of the probability statements that are generally made (typically of the order of 35%), that the average shift in the mean is no more than about 0.15 standard deviations (probably of the order of 0.10 for precipitation). Thus, present levels of skill are equivalent to reductions of variance of less than 2% (at most 1% for precipitation). Very modest skills nevertheless permit reasonably frequent changes in the odds on, say, warm versus cold, to 2:1, compared to the climatological odds of 1:1.

3. An example of skill, positive and negative

Figure 3 shows an example of a recently verified monthly outlook. Prepared in mid-March 1987, it covers the period from mid-March to mid-April. Along the heavy lines labeled "30" the forecaster is indifferent with regard to warm or cold, or to heavy or light precipitation. The temperature probabilities are 0.4 or greater for cold from the Gulf Coast to west Texas, and for warm over much of the Canadian prairie and

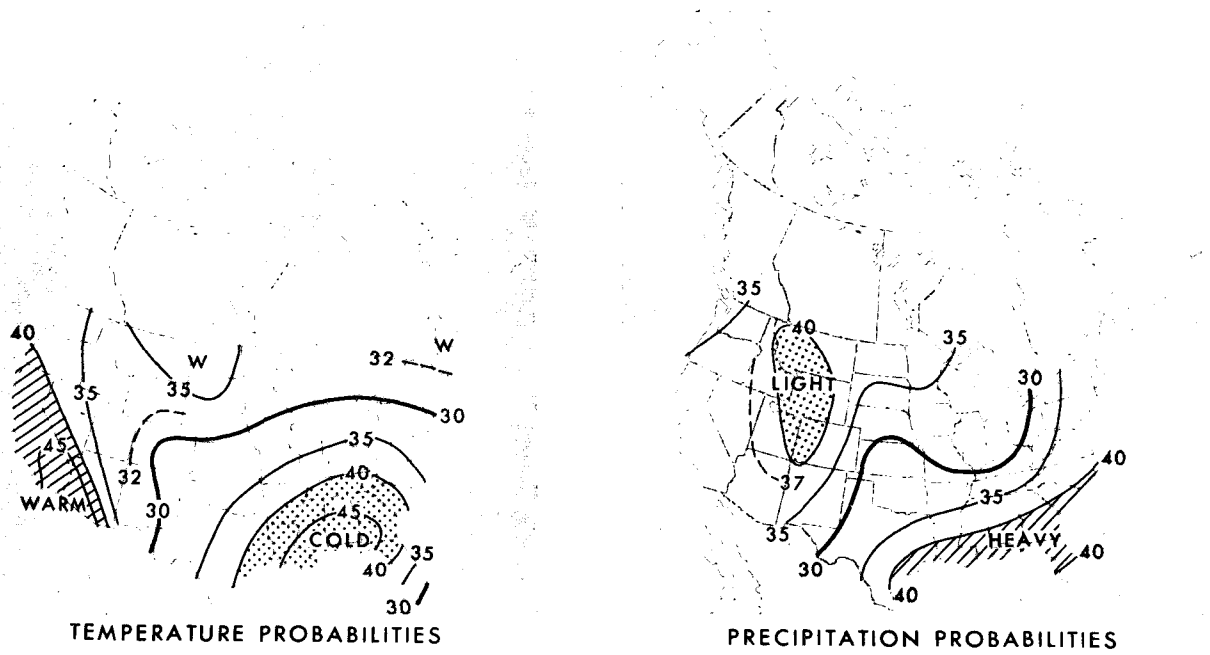

FIG. 1. Published National Weather Service 90-day outlook for January through March 1987.

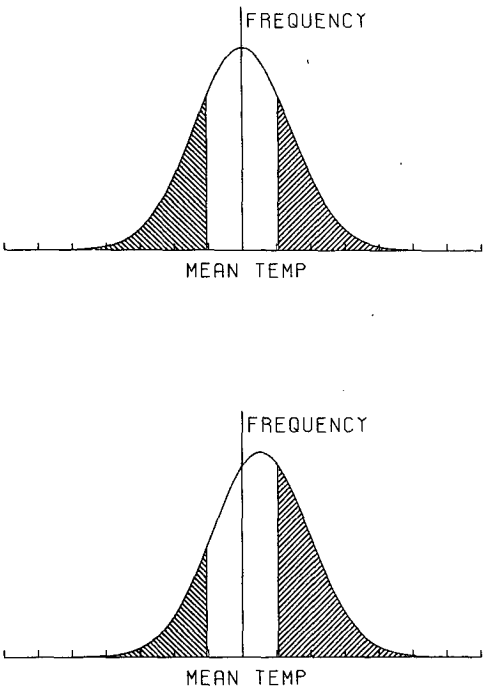

FIG. 2. Idealized climatological distribution of mean monthly or seasonal temperatures (upper graph), and conditional distribution implied by assigning 40% probability to upper temperature range of which the climatological probability is 30% (lower graph). The conditional distribution in the lower graph is shifted 0.270 standard deviations toward higher temperatures. The unshaded areas are within 0.524 standard deviations of the climatological mean, and contain 40% of the area in the upper and 39% in the lower graph. The corresponding probabilities of below normal temperatures are 30% and 21% respectively.

southeastward to parts of northern Wisconsin and northern Michigan. There are no places in North America where the forecaster is that confident with regard to precipitation.

Let us compare the forecast with what subsequently occurred (Fig. 4). The temperature outlook happens to have been relatively skillful. The overall pattern of warm over most of Canada, except for cold over the Canadian arctic archipelago, and cold in the southeastern United States, was correct. Nevertheless New England and coastal southern California were warm and western Nebraska was cold. Climatologically, 40% of the area should have been in the near normal category. In the region shown, this particular month was instead characterized by an unusually widespread occurrence of extremes.

The primary verification tool we use is a slight variant of the ranked probability score (Epstein, 1969; Murphy and Daan, 1985, p. 422), a quadratic scoring rule that accounts for the ordering of the predictand (as from cold to near normal to warm; see Appendix). Scores are calculated routinely for 100 sites across the conterminous United States. The best possible score of 1.0 at each location would require the assignment everywhere of probabilities of 1.0 to the category that subsequently occurs. If the forecast always contains climatological probabilities (0.3, 0.4, 0.3) and the climatological frequencies occur, the average score at each station is 0.790. On this occasion (see Table 1), because of the relative infrequency of the near normal category, a forecast everywhere of the climatological probabilities gave an average score of 0.778 ($=C$). The score for the "official" forecast was 0.808 ($=F$). This yields a "skill"

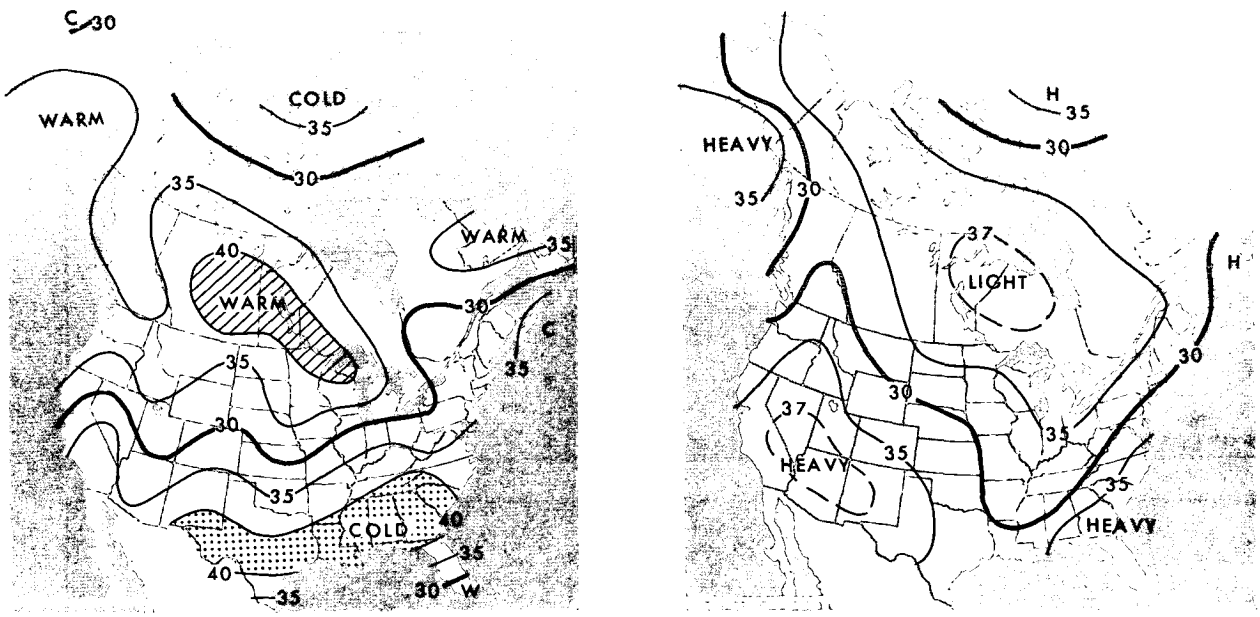

FIG. 3. Published National Weather Service 30-day outlook for mid-March to mid-April 1987.

FIG. 4. Observed mean temperatures, mid-March to mid-April 1987.

FIG. 5. Observed precipitation, mid-March to mid-April 1987.

score relative to climatology $[(F - C)/(1.0 - C)]$ of 0.137, a relatively good score. More typically, values near 0.10 are achieved.

A very high skill score cannot be expected when one is unable to use, reliably, probabilities near 0 or 1. We can calculate the best score achievable given just the frequency of use of particular probability statements (here, none as large as 0.45 or less than 0.15) and the frequencies of the three possible events. In this case the "best" mean score was 0.822, which becomes a "best" skill score of 0.197.

The precipitation forecast did not fare as well when the observations came in (Fig. 5). The more "chopped-up" appearance of the precipitation verification map is typical. Precipitation events occur on relatively small scales, so precipitation is inherently less predictable. One heavy thunderstorm, not at all related to the monthly mean flow pattern, can dominate the monthly total at a station. In this case the skill score of the official forecast was very near zero (Table 2), in fact slightly negative. More usually it is positive although generally less than 0.10.

This level of forecast skill has a reasonably firm empirical base with regard to forecast covering the United States, but can probably be extrapolated more generally to middle latitudes of the Northern Hemisphere. On the basis of predictability studies it should be possible to achieve better scores for tropical locations, but the experience in this regard is very limited.

4. Making long-range forecasts

In practice, until the last couple of years, all of our operational long-range forecasts—the National Weather Service issues monthly outlooks 24 times a year and seasonal (three-month) outlooks 12 times a year—have been empirically based. Increasingly in the

TABLE 1. Scores for forecasts of mid-March to mid-April mean temperatures.

Mean ranked probability scores at 100 U.S. stations	
Forecast (F)	= 0.8084
Climatology (C)	= 0.7780
Skill score (S)*	= 0.137
Best score given probabilities assigned	
Forecast (F)	= 0.8218
Climatology (C)	= 0.7780
Skill score (S)	= 0.197

* Skill score = $(F - C)/(1 - C)$.

TABLE 2. Scores for forecasts of mid-March to mid-April mean precipitation.

Mean ranked probability scores at 100 U.S. stations	
Forecast (F)	= 0.7901
Climatology (C)	= 0.7926
Skill score (S)	= −0.012
Best score given probabilities assigned	
Forecast (F)	= 0.8216
Climatology (C)	= 0.7926
Skill score (S)	= 0.139

last several years numerical model output has become a major factor in determining the monthly outlooks.

Construction of the outlooks is a three-step process. First, the forecaster tries to forecast the mean tropospheric flow pattern for the period in question at an elevation of about 3 km (10 000 ft, the 70-kP level), expressed as a departure from the long-term climate mean for that period. The second step is to infer from the anomalies in this flow pattern the patterns of temperature and precipitation anomaly. The final step is the assignment of probabilities.

a. Monthly outlooks

For the monthly outlook the forecaster tries to take advantage, as much as possible, of the last vestiges of "dynamic" predictability. Even though the numerical models are unable to predict reliably the synoptic daily weather features of the atmosphere beyond 5 to 7 days, the very largest (hemispheric scale) features retain some predictability to day 10 and perhaps beyond. The forecaster examines very closely the output of the model runs produced at the National Meteorological Center (NMC) and also those from the European Centre for Medium-Range Weather Forecasting (ECMWF). Also considered are the sequences of patterns of mean flow anomalies over recent months, and a statistical forecast that involves point-by-point regression on the preceding month's mean heights of the 70-kP pressure surface, in effect a forecast blending climatology and persistence. Also available are maps of recent anomalies of sea surface temperature and snow cover. (The models have, to the extent to which they are capable, already incorporated the effects of observed sea surface temperature anomalies. A point of interest is that the models are global; a global model and global initial data are required to produce good results beyond 4 or 5 days. However, almost all the other tools used by the forecaster refer only to the Northern Hemisphere.) The forecaster assimilates this information into a prognostic map. The major decisions involving the forecaster's judgment and experience are 1) which model to trust for the first half-month or so, and 2) whether to expect the latter part of the month to resemble best the climatology, the models' versions of the early part of the month, the pattern of most recent past month, or the statistical indicators.

The primary tool for estimating the temperature and precipitation anomalies is a set of statistical (multiple regression) relationships that were derived from a large dataset relating precipitation and temperature to contemporaneous values of the 70-kP height field over much of the hemisphere (Klein, 1985). The forecaster also has a set of general rules as to which temperature and precipitation patterns should follow from certain flow fields, but these are used mainly in regions for which regression relations have not been derived. Finally, he or she may examine previous cases with observed flow patterns similar to those forecast and note the corresponding observed temperatures and precipitation.

For the assignment of monthly probabilities the forecaster relies largely on a verification history since 1974, including the seasonal and regional variations in forecast success, and also importantly on the consistency among the various forecast tools. It is particularly noteworthy whether or not the NMC and ECMWF prognostic charts agree in particular regions, and whether they are consistent between today's output and yesterday's.

Most of the skill of the monthly forecast is the skill of dynamic predictions out to ten days. Our models have no demonstrable, reliable skill in the last third of the month and are routinely run to only ten days, although a number of experimental 30-day forecasts have been produced.

b. Prospects for improvements in the monthly outlooks

Efforts to improve the models continue, not only to improve long-range predictions, but also to improve predictions at shorter ranges. Some of this involves larger computers and greater resolution in space and time, some involves better representation of physical processes (which may be feasible only with bigger computers).

Within only the last few years there has been a growing effort to develop methods to produce a priori assessments of the skill of individual predictions. Certain types of atmospheric flow patterns seem to be more persistent and more predictable than others and on occasion our models have done remarkably well further into the future than can easily be explained by chance. Can we identify the times and places where this is likely to occur? How can we use our models, and the physical and mathematical skills they represent, to predict their own skill?

c. Seasonal outlooks

While monthly predictions are treated as initial value problems, seasonal predictions of the atmosphere are more in the nature of boundary-value problems, but with inconstant interactive boundaries. A model that has a little skill out to ten days does not help much in a prediction that extends 90 days. The lower boundary of the atmosphere does not change excessively in a few weeks but there can be dramatic changes over a season at the sea surface. (It is now believed that the variation of equatorial sea surface temperature may be the most important aspect of the atmosphere's lower boundary which we must consider.) What is needed for seasonal forecasting are interactive atmosphere–ocean–land surface models. Since such models are not yet available we must rely for our seasonal predictions on empirical tools.

The empirical tools available for predicting the seasonal mean 70-kP heights and flow field are similar to those used for monthly predictions. An important addition is an array of statistical estimates of persistence at lags of one month and one, two, four and eight seasons. These are probably the most frequently cited sources of forecasters' judgments, although statistical relationships involving equatorial Pacific and North Pacific sea surface temperatures also may be significant. With the help of extensive charts of teleconnections—correlations of height anomalies in one location with anomalies elsewhere in the hemisphere—forecasters construct skeletal prognostic seasonal mean maps of the 70-kP major anomaly locations.

The next step is to interpret these maps in terms of seasonal temperature and precipitation anomalies. For this purpose the forecasters search through the archives for analogs, past months (from the same part of the year) or seasons whose mean 70-kP heights resemble their prognostic maps. From these analogs they produce a composite picture of expected temperature and precipitation anomalies.

For the last year or so a new automated statistical/analog technique (Livezey and Barnston, 1987) whose input is recent and current data on an array of atmospheric and oceanic parameters, and whose output is temperature and precipitation directly, has also been available to forecasters.

One of the important differences between the monthly and seasonal forecasts is that the monthly outlook is primarily the output of a single person. The process has been made sufficiently objective so that different forecasters will come to essentially the same conclusions. The seasonal projection, however, generally involves four forecasters working independently. They then compare their conclusions and (allowing room to make modifications) explain the reasons for their decisions. The lead forecaster then integrates the individual forecasts into one, assigning probabilities again on the basis of experience with verification (in this case since 1959) but also dependent on the strength of the consensus.

d. Prospects for improving seasonal outlooks

Several avenues are being followed to improve the seasonal outlook. One is an effort to increase the lead time of the forecast. The forecasts are currently released about three days before the beginning of the time when they are valid, but there is reason to believe that under certain conditions (specifically the presence of an active El Niño) the forecast can be provided as much as a season in advance with little loss of skill. Are there other circumstances of significant slowly evolving atmosphere/ocean or atmosphere/land surface vacillations that will at least temporarily make enhanced predictability possible?

Among the empirical tools being brought to bear on this question are some sophisticated statistical methods. Some of these methods are new and untested; others, very computer intensive, were entirely impractical only ten years ago, but are now feasible. The question is, are they productive?

Another methodological improvement that needs further study is the means—currently entirely judgmental—by which different sources of information are combined. In particular, the statistical persistence estimates appear to be largely independent of the analog results, although each has its own spatial interdependence. It is not obvious how to add the two inputs together.

Work is also progressing on ocean models that should have the capability of translating past and projected information on surface winds, and up-to-date information on ocean temperatures, into skillful forecasts of the temperatures at the ocean surface, at least in tropical latitudes, where such information is most critical to the monthly as well as the seasonal forecast. Other research is concerned with the inclusion in numerical models of other aspects of the surface boundary conditions, like soil moisture, ice and snow, and sea ice. There is much to learn and probably some years to wait, however, before models and physical predictions will be able to play the significant role in the seasonal outlook that they now have in the monthly outlook.

Acknowledgments. I am indebted to the members of the Prediction Branch of the Climate Analysis Center, and especially to their Chief, Dr. Donald Gilman, not only for advice and comments on this manuscript, but also for allowing me to observe and question them as they carried out their long-range forecasting responsibilities.

APPENDIX

The Ranked Probability Score as a Verification Tool

The ranked probability forecast as defined by Murphy and Dann (1985), $\mathrm{RPS_{MD}}$, ranges from 0 for a perfect forecast to $N-1$ for a forecast, with probability one, of one extreme of N categories when the other extreme is observed. In the present application $N = 3$ and we want the score to range from 0 for the worst possible forecast to 1 for a perfect forecast. Thus

$$\mathrm{RPS_{CAC}} = 1 - \frac{1}{2}\,\mathrm{RPS_{MD}}$$

$$= 1 - \frac{1}{2}\sum (R_i - D_i)^2$$

where the R_i are the cumulative forecasts of the ith category and the D_i are the cumulative observations. Thus, for a (climatological) forecast of (0.3, 0.4, 0.3),

$R = (0.3, 0.7, 1.0)$. There are three possible observations. For an observation of the first (lowest) category $D = (1.0, 1.0, 1.0)$ and for an observation of the last category $D = (0.0, 0.0, 1.0)$. Given a forecast of climatology and the observation of the first category, $\mathrm{RPS_{CAC}} = 1 - \frac{1}{2}(0.7^2 + 0.3^2 + 0.0) = 0.71$. When the middle category is observed, $\mathrm{RPC_{CAC}} = 1 - \frac{1}{2} \times (0.3^2 + 0.3^2 + 0.0) = 0.91$. By symmetry, or by applying the formula again, the score when the last category is observed, $\mathrm{RPS_{CAC}}$, is 0.71. If the relative frequencies of the three categories are their climatological values then the average score will be $(0.3)(0.71) + (0.4)(0.91) + (0.3)(0.71) = 0.79$.

REFERENCES

Brown, B. G., R. W. Katz and A. H. Murphy, 1986: On the economic value of seasonal precipitation forecasts: The fallowing/planting problem. *Bull. Amer. Meteor. Soc.*, **67**, 833–841.

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

Klein, W. H., 1985: Space and time variations in specifying monthly mean surface temperature from the 700 mb height field. *Mon. Wea. Rev.*, **113**, 277–290.

Livezey, R. E., and A. G. Barnston, 1987: An operational multifield antilog prediction system for United States seasonal temperatures. Submitted to *J. Geophys. Res.*

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

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

Murphy, A. H., and H. Daan, 1985: Forecast Evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, Eds., Westview Press, 545 pp.

World Meteorological Society, 1984: Report of the session of the commission for atmospheric sciences working group on long-range weather forecasting research. *Long-Range Forecasting Research Publications Series, No. 4*, WMD, Geneva.