NOTES AND CORRESPONDENCE

Chance Behavior of Skill Scores

UWE RADOK

Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado

26 February 1987 and 1 August 1987

ABSTRACT

The skill score $S = (R - E)/(T - E)$ (representing R actual and E expected successful categorical forecasts in a total of T forecasts) remains a valid tool for assessing the overall quality of current probabilistic long-range forecasts, which start from categorical subdivisions of the forecast area. The skill score definition is modified to become a chi variate with one degree of freedom. Two sets of skill scores computed from forecasts of U.S. monthly precipitation and mean temperature are shown to have frequency distributions of similar shape with nonzero means and standard deviations generally corresponding to smaller independent numbers of verification points than those actually used. The largest skill scores of those examined were obtained for recent precipitation forecasts during a period when forecasts using only climatology were similarly skillful. This suggests that cooperation on part of the climate system remains an essential success ingredient in extended forecasting. A sequential procedure for monitoring the changing level of operational forecasting skill is described.

1. Introduction

The skill score has long been the standard tool for assessing the success of long-range forecasts, alongside more sophisticated methods involving pattern recognition (Somerville, 1977) and empirical orthogonal functions (Bettge et al., 1981). The new probabilistic forecasts which were started in midsummer 1982 (Gilman, 1986) add to the previous categorical subdivision of the forecast area extra lines indicating the forecasters' level of confidence that not only the right categories have been picked but also the regions of the larger anomalies. The simple categorical choices however remain the first steps in the forecasts. Moreover they provide the basis for historical assessments of changes in forecast skill since the beginning of long-range forecasts. This makes it worthwhile to examine the fluctuations that could occur in the skill score by chance. For this purpose a theoretical sampling distribution for the skill score is derived in this note and compared with the distributions of two sets of operational skill scores.

2. Definitions

The most common skill score definition is based on a contingency table of the form shown in Table 1. Denoting the sum of the diagonal, $\sum_{i=1}^{3} x_{ii}$, by R and its expected value by E, the skill score S is defined as

$$S=\frac{R-E}{T-E}. \tag{1}$$

Different definitions have been used for E, the total number of successful forecasts expected to be obtained by chance. In the original skill score proposed by Panofsky and Brier (1963)

$$E_1=\sum\frac{x_{i.}x_{.i}}{T}. \tag{2}$$

This has been shown by Livesey and Skilling (1985) to minimize the information of the table (maximize its "Shannon entropy"). The main operational definition now in use by NOAA's Climate Analysis Center is

$$E_2=0.3(x_{.1}+x_{.3})+0.4(x_{.2}). \tag{2'}$$

This has the (largely cosmetic) drawback that it does not give a unique minimum value to S; writing E_2 in the alternative form

$$E_2=0.3(T-x_{.2})+0.4x_{.2}=0.3T+0.1x_{.2}$$

shows that for $x_{.2}=0$, $E_{2_{\min}}=0.3T$, $S_{2_{\min}}=-3/7$; and for $x_{.2}=T$, $E_{2_{\min}}=0.4T$, $S_{2_{\min}}=-2/3$.

A modified form of E_2 that avoids this uncertainty and simplifies the theoretical results is

$$E_3=\frac{T}{3}. \tag{2''}$$

Then $S = 3R/2T - \frac{1}{2}$ and $S_{\min} = -\frac{1}{2}$, irrespective of the structure of the diagonal.

3. The sampling distribution of S

A statistical significance test appropriate for the full three-way contingency table in Table 1 with expected

Corresponding author address: Dr. Uwe Radok, Meteorology Department, University of Melbourne, Parkville, Victoria 3052, Australia.

value E_1 is given in most textbooks and involves the chi square distribution with 4 degrees of freedom. However, this definition of E places too much weight on the details of the failed forecasts which are of interest mainly for a detailed examination of the forecast method used. In the operational context the skill score essentially compares two observed numbers o_1, o_2 with two expected numbers e_1, e_2. This comparison defines a chi square with one degree of freedom

$$\chi^2_{(1)} = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2}. \tag{3}$$

Here $o_1 = R$, $o_2 = T - R$, $e_1 = E$, $e_2 = T - E$ so that

$$\chi^2_{(1)} = \frac{(T-E)(R-E)^2 + E(E-R)^2}{E(T-E)}$$

$$= \frac{T(R-E)^2}{E(T-E)}, \tag{4}$$

and since from (1)

$$R - E = S(T-E)$$

$$\chi^2_{(1)} = \frac{T-E}{E/T} S^2. \tag{5}$$

For the special case of $E = E_3 = T/3$ this takes the simple form

$$\chi^2_{(1)} = 2TS^2. \tag{5'}$$

Now the chi square distribution with one degree of freedom has the form

$$F(\chi^2_{(1)})d(\chi^2_{(1)}) = [2^{1/2}\Gamma(1/2)]^{-1}(\chi^2_{(1)})^{-1/2}e^{-\chi^2_{(1)}/2}d(\chi^2_{(1)}). \tag{6}$$

With $\Gamma(1/2) = \pi^{1/2}$ and $\chi_{(1)}$ as variables in place of $\chi^2_{(1)}$ (so that $d\chi^2_{(1)} = 2\chi_{(1)}d\chi_{(1)}$), and separating positive and negative values of $\chi_{(1)}$, this becomes

$$f(\chi_{(1)})d(\chi_{(1)}) = \pi^{-1/2}e^{-\chi^2_{(1)}/2}d\chi_{(1)}, \tag{7}$$

the standard Gaussian distribution with zero mean and unit variance. According to (5), therefore, chance skill scores can be expected to be normally distributed with a variance that depends on the total and expected forecast numbers, viz. $\sigma^2 = E/[T(T-E)]$. For $E = T/3$ this simplifies to $\sigma^2 = (2T)^{-1}$.[1]

4. Observed and chance skill scores

Two sets of operational skill scores have been used to test the results of the preceding section. The first set, kindly made available by Dr. D. L. Gilman, covered the period from winter 1974/75 through summer 1979 and consisted of 100 station scores and expected success numbers E_2 for each of 19 seasons. Each score was computed from six forecasts of the monthly mean temperature, starting the forecast period from the beginning and from the middle of each month. To reduce the effect of spatial coherence only half the stations were used here. The second set of skill scores, kindly made available by Dr. K. Hanson, consisted of 21 sets of monthly precipitation and mean temperature skill scores for the official CAC forecasts as well as for persistence and climatology ones, all of them computed with $E = E_3 = T/3$ from one precipitation and one temperature for each of the 48 contiguous states of the United States.

TABLE 1. Contingency table for the calculation of the skill score.

	Observed class numbers			
Forecast class numbers	1	2	3	Totals
1	x_{11}	x_{12}	x_{13}	$x_{\cdot 1}$
2	x_{21}	x_{22}	x_{23}	$x_{\cdot 2}$
3	x_{31}	x_{32}	x_{33}	$x_{\cdot 3}$
Totals	$x_{1\cdot}$	$x_{2\cdot}$	$x_{3\cdot}$	$T = \sum x_{i\cdot} = \sum x_{\cdot i}$

The analysis involved rearranging the scores or equivalent chi values in order of ascending magnitude. As chance results, the resulting cumulative frequencies would then be expected to approximate straight lines in probability coordinates, constructed to make $\int_{-\infty}^{\chi} e^{-\chi^2_{(1)}/2}d\chi$ a linear function of χ.

Figure 1 shows the chi values of the first set plotted in this manner separately for the four seasons and the year, with offset abscissa scales. The full straight lines are regression lines fitted to the nine ordinates representing cumulative probabilities from 0.1 through 0.9, while the broken straight lines are the expected theoretical distributions, N (0, 1). The observed distributions can be seen to have means (50% intersections) different from zero; they also have steeper slopes which imply that the variances σ^2 of the observed χ values are larger than their chance value, unity. A plausible explanation is that the six overlapping forecasts in each skill score are correlated with one another. An effective number of T_{eff} of independent forecasts can then be defined by $T_{\text{eff}}/T = 1/\sigma_\chi{}^2$, so that in the present case $T_{\text{eff}} = 6/\sigma_\chi{}^2$.

The mean values of χ for the four seasons and for the entire first set are given in Table 2, together with their standard deviations and independent forecast numbers T_{eff} which also appear in Fig. 1. The pairs of average skill scores $\bar{S}$ in Table 2 represent the extremes of E_2, which can range from $E_2 = 2.4$ (for $\chi_{\cdot 2} = T$) for $E_2 = 1.8$ (for $\chi_{\cdot 2} = 0$). As means of 250 scores (200 for the fall season) the $\bar{S}$ values are significantly positive for winter and spring only, according to the Student's t-tests (Hoel 1962, section 11.5) in the last two lines of the table. The numbers of independent forecasts range

[1] A reviewer has pointed out that, in this special case, S is a binomial variate which approaches the normal form for large T. The $\chi_{(1)}$ variate is not subject to such restrictions.

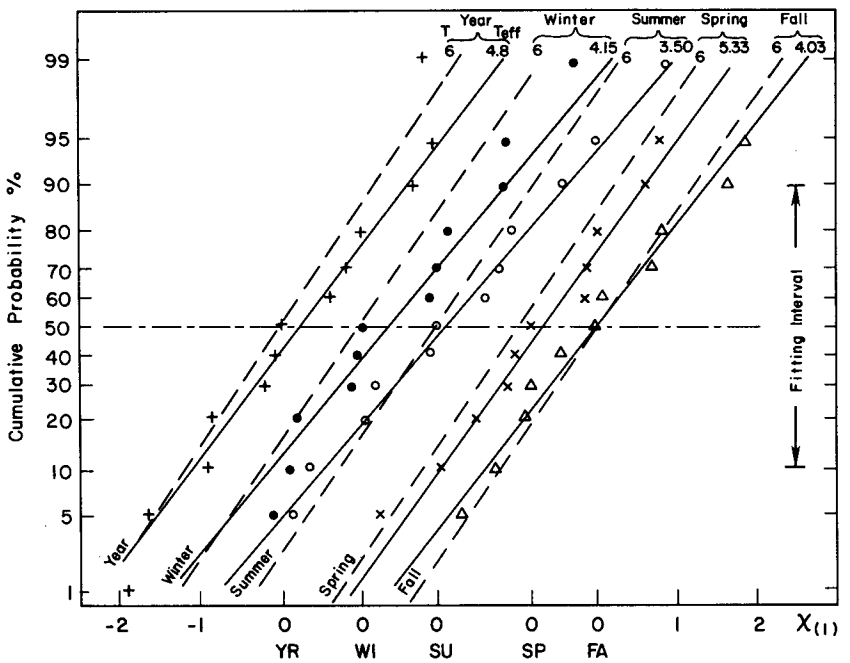

FIG. 1. Cumulative frequency distributions of chi variates derived from the skill scores of NOAA's monthly mean temperature forecasts for 50 U.S. stations during the period winter 1974/75 through summer 1979 (19 seasons; D. Gilman's data). The broken lines show the chance distribution of chi with one degree of freedom and six forecasts per score. The full lines represent normal approximations to the chi values computed from the actual scores and correspond to the smaller numbers of independent forecasts shown.

from a minimum of 3.5 in summer to a maximum of 5.3, only a little less than the actual number of forecasts (6), in spring.

The same type of representation has been used in Fig. 2 for the skill scores themselves of the official and climate forecasts in the second dataset. For economy of space, the cumulative temperature skill scores are plotted rising from right to left, and those of precipitation descending in that direction. Again the plots are reasonably straight; their means and standard devia-

TABLE 2. Characteristics of the distribution of skill scores for six monthly mean temperature forecast for 50 U.S. stations (D. Gilman's data). The skill scores have been converted to their $\chi_{(1)}$ equivalents, as described in the text, for the two extreme possibilities of the expected success number, $E_{max} = 2.4$, $E_{min} = 1.8$. The T_{eff} are the numbers of independent forecasts in the 6 making up each skill score. The last two lines test the statistical significances of the $\bar{\chi}$ with Student's $t = \bar{\chi}/\sigma_{\bar{\chi}}$, where $\sigma_{\bar{\chi}} = \sigma_\chi N^{-1/2}$.

	Spring	Summer	Fall	Winter	Year
Number of Scores (N)	250	250	200	250	950
$\bar{\chi}$	0.322	0.063	−0.038	0.356	0.261
$\bar{S}_{max} = \bar{\chi}\Big/\left(\frac{T-E_{max}}{E_{max}/T}\right)^{1/2} = \bar{\chi}/3$	0.107	.021	−0.013	0.119	0.087
$\bar{S}_{min} = \bar{\chi}\Big/\left(\frac{T-E_{min}}{E_{min}/T}\right)^{1/2} = \bar{\chi}/3.742$	0.086	.017	−0.010	.095	0.070
σ_χ	1.061	1.310	1.220	1.203	1.118
$T_{eff} = \frac{6}{\sigma_\chi^2}$	5.33	3.50	4.03	4.15	4.80
$t = \bar{\chi}N^{1/2}/\sigma_\chi$	4.80	0.76	−0.44	4.68	7.20
Chance probability of exceeding t	<0.001	>0.40	>0.60	<0.001	<0.001

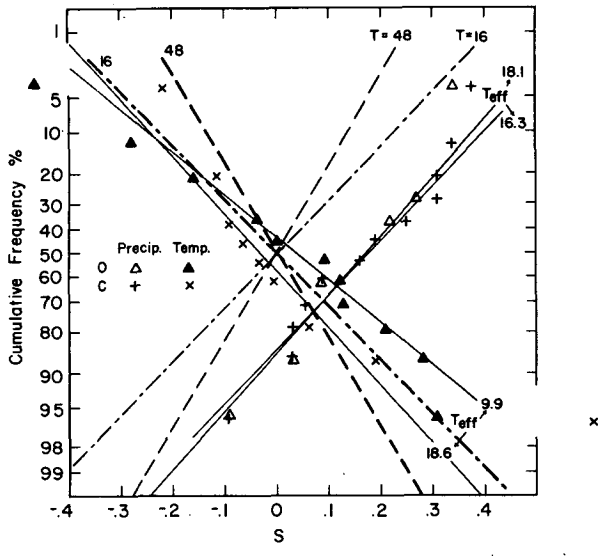

FIG. 2. Cumulative frequency distribution of the skill scores of precipitation (P) and monthly mean temperature (T) forecasts skill scores for December 1983 through September 1985 (K. Hanson's data). O: official NOAA forecasts, C: climatology forecasts. The broken lines are chance distributions of scores computed from 48 and 16 verification points. The full lines are normal approximations to the distributions of the actual scores and correspond to the smaller verification numbers shown.

tions are given in Table 3. They suggest statistically significant skill scores of 0.17 and 0.18 as means of 21 values for precipitation, and no skill for the temperature forecasts. In this case we expect $\sigma_s^2 = (2T)^{-1}$; the actual standard deviations are equivalent to independent forecast numbers $T_{\text{eff}} = (2\sigma_s^2)^{-1}$ ranging from 18 to as low as 4, compared to the actual value of T (48), presumably due to the spatial coherence of the forecasts. The large scatter (low T_{eff} values) of the persistence forecasts points to especially large patterns as responsible for successes or failures.

5. The operational assessment of skill scores

The second dataset has also been used to simulate an operational test for possible trends in forecasting skill. The test was designed to discriminate between several alternative levels of skill. In the operational context these are expressed more conveniently in terms of the forecast success ratio which for $E_3 = T/3$ has the form

$$R/T = \frac{2}{3}S - \frac{1}{3} \tag{8}$$

Corresponding values of R/T and S are given in Table 4. The procedure used is the "sequential probability ratio test" (e.g., see Hoel, 1962, section 14.1). For discriminating with optimum efficiency (minimal sampling) between alternative means of a normal distribution with unit variance, progressive sums of the variable [$(2T_{\text{eff}})^{1/2}S$, in the present case] are calculated and plotted as function of time in relation to pairs of paralled straight lines, representing two mean skill scores $\bar{S}$ (i.e., $\Sigma\ (2T_{\text{eff}})^{1/2}\bar{S}_1 < [\Sigma\ (2T_{\text{eff}})^{1/2}\bar{S}_2]$. If the higher of these mean scores ($\bar{S}_1$) is true the cumulative sum should move above that line in all but a small number β of cases; if the lower mean score is true, the cumulative sum should fall below the $\bar{S}_2$ line in all but a small number α of cases. No decision between the two mean scores can be made while the cumulative sum remains between the two lines; instead further scores must be awaited until one of the lines is crossed, although a cumulative sum rising faster (slower) than the two lines can be taken as suggesting an upward (downward) trend in skill.

For the present case the limits are defined by the inequalities (derived in Hoel, 1962, p. 355)

$$\frac{1}{2T_{\text{eff}}\Delta\bar{S}}\log_e\frac{\beta}{1-\alpha} + \frac{m}{2}\Sigma\,\bar{S} < \sum_1^m S < \frac{1}{2T_{\text{eff}}\Delta\bar{S}}\log_e\frac{1-\beta}{\alpha} + \frac{m}{2}\Sigma\,\bar{S}, \tag{9}$$

where $\Delta\bar{S}$ is the difference between the two prescribed mean scores and $\Sigma\ \bar{S}$ their sum; m is the progressive number of skill scores tested.

The limits corresponding to (9) have been calculated with $\alpha = 0.05$ and $\beta = 0.1$ for the forecast skill scores

TABLE 3. Characteristics of the distribution of skill scores for 21 forecasts of the mean monthly temperature (T) and precipitation (P) for the 48 contiguous states (K. Hanson's data). The T_{eff} are the number of independent forecasts in the 48 making up each skill score. For other details see text.

	Forecasts					
	Official		Climate		Persistence	
	T	P	T	P	T	P
$\bar{S}$	0.04	0.17	−0.03	0.18	0.05	0.08
σ_S	0.225	0.166	0.164	0.175	0.237	0.352
$T_{\text{eff}} = \frac{1}{2\sigma_s^2}$	9.9	18.1	18.6	16.3	8.9	4.0
$t = \bar{S}(21)^{1/2}/\sigma_s$	0.81	4.69	−0.08	4.71	0.97	1.04
Probability exceeding t with 20 degrees of freedom	>0.30	<0.001	>0.90	<0.001	>0.30	>0.30

TABLE 4. Skill Score (s) and forecast success ratio (R/T) equivalents.

S	0	0.1	0.25	0.40	0.55	0.70	0.85
R/T	0.33	0.4	0.5	0.6	0.7	0.8	0.9

of the second dataset, transformed to chi variates with $T_{\text{eff}} = 48$ and $T_{\text{eff}} = 16$. In Figs. 3 and 4 the progressive sums of the χ equivalents of S are shown in relation to their limits for the entire period of record. In practice a test would become conclusive (with the assurance set by α and β) when the χ plots cross one of the limits. Such crossings happened for both the official and climatology forecasts of precipitation when the full number (48) of forecasts was used (Fig. 3). This is shown more clearly by a numerical version of the test (Table 5). In a truly operational context the test would end with a crossing and begin anew from that point; since the limits are parallel straight lines this simply means transferring the origin to the time (m) of the crossing. Treated in this manner the last four climatology forecasts of temperature (x) suggested a significant increase in success ratio to over 50%.

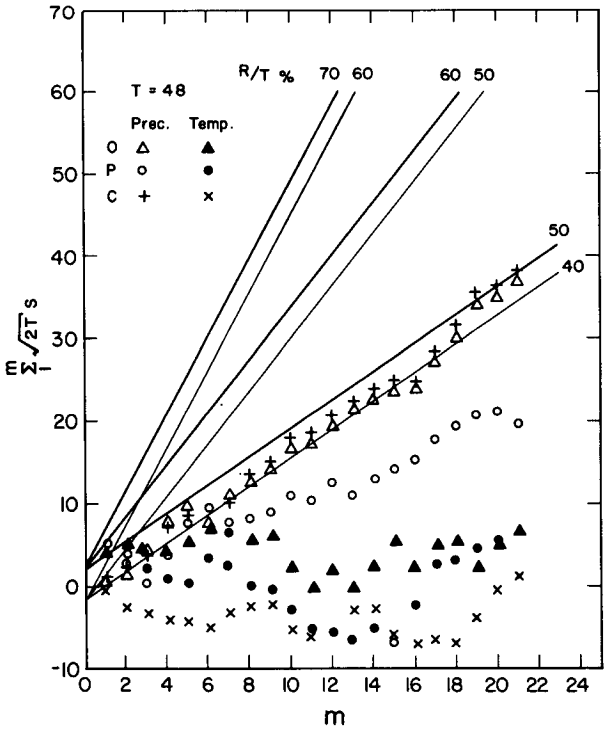

FIG. 3. Cumulative sums of chi values computed from the skill scores of Fig. 2. The parallel lines are sequential limits for discriminating between the mean forecast success percentages indicates, as explained in the text, assuming each score to represent $T = 48$ independent verification points, O: official NOAA forecasts, P: forecasts based on persistence, C: forecasts based on climatology.

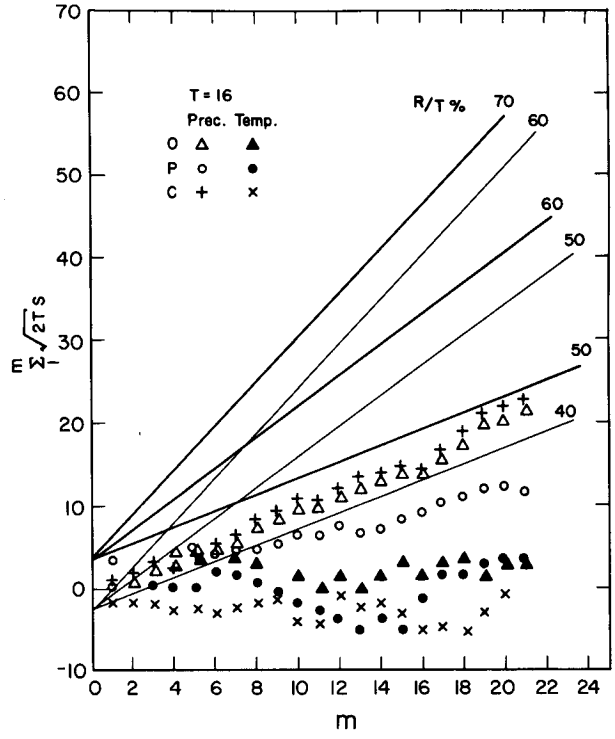

FIG. 4. As in Fig. 3, but assuming each score to represent only 16 independent verification points.

With a smaller effective number Fig. 4 shows that the decision limits move farther apart; the forecast sequences then remained in limbo between the 40% and 50% success rates for the entire period of 21 months.

6. Conclusion

The sequential analysis represents one way of continually monitoring changes in skill and assessing their statistical significance. From a physical point of view the concurrent behavior of the official and climate skill scores for precipitation in the second dataset examined is quite revealing. These precipitation skill scores had comparable positive mean values. One possible interpretation is that the official forecasts gave a good deal of weight to climatic mean values, and succeeded because the climate during the period studied "played the game", with small anomalies. At the current state of the art of extended forecasting such conformism or persistence may be prerequisites for more than occasional successes. Viewed in this way, the forecasts follow a fixed track while climate weaves around their predictions, approaching and following them for a time now and then before diverging again. If this is a realistic appraisal, a monitoring procedure revealing sudden changes in skill, in the manner here outlined, ought to become a stock in trade of long-range predictions.

A modified procedure for probabilistic forecasts

TABLE 5. Sequential testing for 10% differences in forecast success ratios. The observed sums have been computed from the precipitation scores of the official forecasts for May through September 1985.

m	Observed $\sum_{1}^{m} \sqrt{2TS}$	Hypothetical R/T limits %, $T = 48$							
		40 vs 50			50 vs 60			60 vs 70	
1	3.72	0.18	3.68⊕	1.65	×	5.15	3.12	×	6.62
2	6.76	1.90	5.40⊕	4.84	×	8.34	⊕7.78		11.27
3	10.78	3.61	7.11⊕	8.02	×	11.52	⊕12.43		15.93
4	11.60	5.33	8.83⊕	11.21	×	14.70	⊕17.08		20.58
5	13.62	7.04	10.54⊕	⊕14.39		17.89	⊕21.74		25.24

Interpretation: $R/T = 50\%$ if the alternative is chosen to be 40%; the test remains undecided between 50% and 60% until $m = 5$ when 50% is favored. The alternative of 70% is rejected at $m = 2$.

might use a larger number of classes and make the score give also some credit for forecasts which miss by only one class (e.g., predicted "much above" when "above" occurred). It would be worthwhile to carry out a detailed comparison of the distribution of the predicted probabilities with that of the verified frequencies, so far made available only as a three-year summary by Gilman (1986).

Acknowledgment. A second reviewer's comments have helped greatly to clarify the intended meaning of this note.

REFERENCES

Bettge, A. G., D. P. Baumhefner and R. M. Chervin, 1981: On the verification of seasonal climate forecasts. *Bull. Amer. Meteor. Soc.* **62,** 1654–1665.

Gilman, D. L., 1986: Expressing uncertainty in long range forecasting. *Namias Symposium,* Roads, J. O. Ed., 174–187. SIO Ref. 84-17. La Jolla, CA 92037, 198 pp.

Hoel, P. G., 1962: *Introduction to Mathematical Statistics.* 2nd ed., Wiley, 427 pp.

Livezey, A. K., and J. Skilling, 1985: Maximum entropy theory. *Acta Crystallogr.,* **A41**(2), 113–122.

Panofsky, H. A., and G. W. Brier, 1963: *Some Applications of Statistics to Meteorology.* Pennsylvania State College, 224 pp.

Somerville, R. C. J., 1977: Pattern recognition techniques for forecast verification. *Contrib. Atmos. Phys.,* **50**(3), 403–410.