

## RESULTS FROM A PROBABILISTIC FORECAST CONTEST WITH A PROPER SCORING METHOD

Harold E. Brooks

Department of Atmospheric Sciences  
University of Illinois at Urbana-Champaign  
Urbana, Illinois

Jack L. Powell

Department of Psychology  
University of Hartford  
West Hartford, Connecticut

### 1. INTRODUCTION

Many factors affect the performance of forecasters. The quality of numerical guidance, experience, and time taken to prepare the forecast are just a few of the possible influences. We wish to look other, more subtle, but possibly important factors. These include the influence of recent past performance and confidence in the ability to forecast. On the former point, anecdotal evidence exists that operational forecasters do, under some circumstances react to their previous performance. For instance, in cases where forecasters have failed to issue warnings for severe weather events which occur early on their shift or on a recent shift, occasionally they will issue warnings in less certain situations in the future. The desire to not be burned again can result in less accurate forecasts, with the forecaster making predictions not in accord with his true beliefs. Another possible response to previous unsatisfactory forecasts is to alter the strategy used in making the next forecast in an effort to improve.

On the second point, it appears on the surface that it is inherently better for a forecaster to have a high degree of confidence in his ability. There are circumstances, however, in which that is not necessarily true. A forecaster with less confidence may put more effort into preparing the forecast and may gain a truer picture of the situation and, thus, make a better forecast. The overestimation of one's understanding of a situation may be just as harmful as a lack of understanding.

We wish to look at these points, as well as the issue of what we actually think when we assign a probability to an event. Powell (1987) looked at how people responded when informed that there was a certain probability of an event occurring. People with little knowledge of a subject were heavily influenced by the probabilities assigned to events.

People who viewed themselves as having knowledge of a subject typically ignored information about the probability and made decisions based upon their own evaluations of the situation. However, in many cases these decisions reflected overconfidence in their abilities. Here we see potential application to public forecast problems. In many cases, the two groups of people may see the same data and put different interpretations on it. The issue of public perception of forecast products has been dealt with extensively. Perhaps just as important, though, is the *forecaster's* perception of the forecast. Here, we have the opportunity to look at the performance of a group of people who know (at least in theory) the limitations of the guidance and some understanding of the meaning of probability forecasting.

To examine these questions, we have taken results from a forecasting contest at the University of Illinois this spring. We have looked at how different forecast strategies worked and followed individual forecasters through the contest to see how they reacted to their previous performances. One result is that forecasters expressing greater confidence in their forecast (by putting narrower and more peaked probability distributions) generally performed worse than those forecasters who showed less confidence in their forecasts.

### 2. CONTEST PROCEDURE

Twenty-four members of the Department of Atmospheric Sciences at the University of Illinois participated in the contest, representing a wide range of experience, interests, and training. Five faculty, 16 graduate students, one post-doc, one research programmer (with an MA in atmospheric sciences), and the department computer system manager forecast in the contest. Forecasts were made every Tuesday and Thursday for the spring 1989 semester, with the exception of Spring Break, for a total of 29 forecasts. (The spring semester at Illinois ends on a Wednesday.) Quantities forecast were midnight-to-midnight high and lows and precipitation amount for each of the next two days. Participants indicated the probability of a given range of temperature and precipitation occurring. Each of the temperature forecasts were divided into 15 categories. The middle category was centered around the climatological temperature for that variable. The central 11 categories (3 through 13) were 3 K wide, with categories 2 and 14 being 4 K wide and the outermost categories containing anything more than 21 K from climatology. Precipitation forecasts were divided into six categories, 0-Trace, 0.01"-0.10", 0.11"-0.25", 0.26"-0.50", 0.51"-1.00", and anything more than 1.00". Each forecaster then had to enter his probabilities onto the department's HP computer using one of two interface programs, written for the contest when it was used at Saint Louis University. One program limited the number of categories which anyone could use for temperature forecasting to three, and the number of precipitation categories to four. The second program allowed any number of categories to be used, but forced the forecaster to enter all 66 categories for a given forecast into the program. As a result, another program was written and made available to participants to allow them to enter only the nonzero probabilities in their forecasts. Some continued to use only the three-category forecast method, which complicates interpretation of the results. Since this was the first time that a forecast contest using this scoring method has been used at Illinois, results from the first five forecasts of the semester will be disregarded in this paper. It is felt that this time was needed for participants to get a feel for the

contest and to experiment with various forecasting philosophies. FOUS and FOUM data were provided, when available, for the forecasters, as well as maps from a DIFAX circuit. Verification of the forecasts were made by using data collected at the Illinois State Water Survey in Champaign. Forecasts were entered for MOS and consensus of the human forecasters for comparison to individuals.

The method used to score the forecasts is the ranked probability score from Epstein (1969). The equation for the score for any forecast which verifies in category  $j$  is given by

$$S_j = \frac{3}{2} - \text{Shape} - \text{Error}$$

where  $S_j$  is the score and Shape and Error are given by

$$\text{Shape} = \frac{1}{2(K-1)} \sum_{i=1}^K \left[ \left( \sum_{n=1}^i p_n \right)^2 + \left( \sum_{n=i+1}^K p_n \right)^2 \right]$$

and

$$\text{Error} = \frac{1}{K-1} \sum_{i=1}^K |i-j| p_i$$

where  $K$  is the number of categories and  $p_m$  is the probability in the  $m$ th category. Epstein discusses the behavior of this scoring rule in detail and we will briefly summarize the characteristics by examining each of the two variable terms. The minimum value for Shape is 0.5, which occurs when a probability of 1 is given to a single category. As the probabilities are spread out more, Shape decreases. In some sense, this is a measure of the uncertainty a forecaster assigns to his forecast. Error is quite simple. It ranges from 0, for a forecast when the probability given to the verifying category is 1, to 1, when the probability 1 is given to an event at one end of the range and the verification is at the other end. Note that even though Shape decreases as the probabilities are spread out, the fact that such a strategy results in a nonzero Error means that the  $S_j$  will always be less than 1 for a spread distribution.

The maximum value for a forecast is 1, as mentioned before, possible only when a single forecast category is used. The minimum value, however, is a function of the verifying category. For an extreme event, the minimum score is 0, while for an event in the center of the forecast region (in the case of temperature, a climatological high or low), it is 0.5. Again, these scores are achievable only by assigning 1 to a single category. Therefore, the forecaster is faced with the dilemma that the same strategy that results in the maximum possible score is the same strategy that results in the minimum score. A schematic of the bounds on the score as a function of category is shown in Figure 1.

For further examples of the scoring, we consider some possible simple forecast strategies. The score for a single category forecast is given by  $S_j = 1 - |i-j|/(K-1)$ . (For the 15 category temperature contest, this becomes  $1 - |i-j|/14$ .) If the probabilities are evenly distributed throughout the range, the formula for scoring becomes

$$S_j^* = \frac{2}{3} + \frac{1}{6K} + \frac{(K-j)(j-1)}{K(K-1)}$$

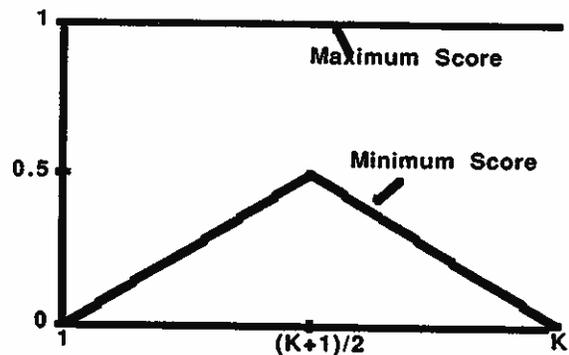


Figure 1—Bounds on the possible values on any forecast by the ranked probability score with  $K$  categories.

which reduces to  $(382 + 48j - 3j^2)/630$  for a 15 category forecast. The possible scores for this forecast range from 0.678 to 0.911 and the distribution is shown in Figure 2.

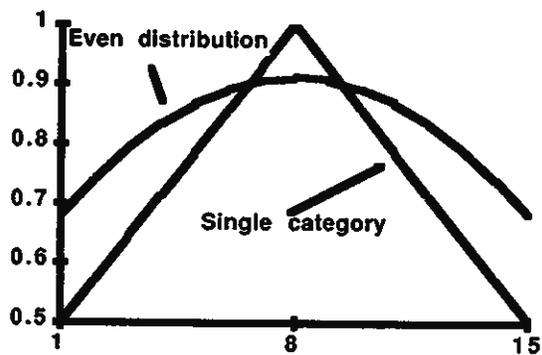


Figure 2—Possible scores from an even distribution of probabilities and a probability of 1 in a single category forecast versus verification category for a 15-category forecast. Note scale change from Figure 1.

We want to consider some more likely, but still simple, forecast strategies in order to look at the role of the two terms in the scoring equation. We choose three symmetric strategies covering three, four, and five categories. The first (A) is (0.1, 0.8, 0.1), the second (B) (0.1, 0.2, 0.4, 0.2, 0.1), and the third (C) is (0.05, 0.125, 0.2, 0.25, 0.2, 0.125, 0.05). A represents a high amount of confidence in the central value of the forecast, while C is a very uncertain forecast. We also look at one asymmetric forecast (D) (0.5, 0.3, 0.2), such as might be employed in the precipitation forecast, or in the case of frontal passage. [At this point, we point out that the maximum possible score for a given forecasts is achieved by the verification occurring in the category where the summed probability from either end of the distribution reaches 0.5. Thus, it is possible that, for some forecasts, such as (0.4, 0.2, 0.2, 0.2), the maximum score does not occur for a verification in the category that the forecaster puts the highest probability.] Figure 3 shows the dependence of these four forecasts on verification category, assuming that each its maximum in category 8. Table 1 gives the scores for the central seven categories for each method. We want to point out the fact that the most important factor for the symmetric distributions is that the center of the distribution is in the verification category. As you would expect, the only benefit to spreading the distribution occurs in cases where the verification is far from the center of the distribution. The forecaster is helped only by recognizing the possibility of the occurrence of rare events.

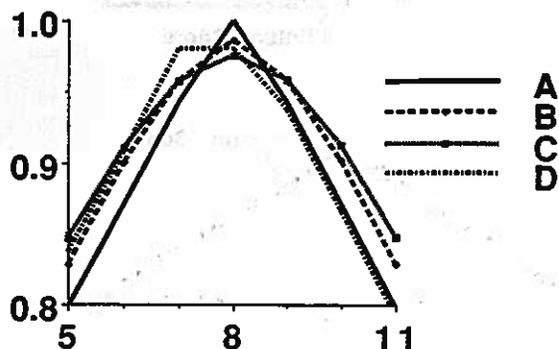


Figure 3—Scores from central portion of domain for 4 simple forecasts. See text for details.

Category	Forecast			
	A	B	C	D
5	.799	.829	.847	.836
6	.870	.900	.911	.908
7	.941	.957	.957	.979
8	.999	.986	.975	.979
9	.941	.957	.957	.936
10	.870	.900	.911	.865
11	.799	.829	.847	.794
Shape	.487	.457	.439	.471

Table 1—Scores for central region for forecasts shown in Figure 3. Shape of each forecast is also given. Note that more peaked forecast is better in category 8, but suffers in outlying areas.

This scoring rule is a "proper" scoring rule. This means that the optimal strategy for a forecaster is to assign exactly the probabilities to the events he truly believes will occur. Any attempt to "hedge" or be "overbold" will result, in the long term in a lower score. (The proof of the proper nature of the rule is given by Murphy (1969).) The main purpose of this is to force the forecaster to describe the forecast situation as scientifically accurate as he believes he can. Anything other procedure results in a lower score (Murphy and Epstein, 1967). This means that the forecaster cannot play any games with the scoring by adjusting his forecast to the rules. As an obvious example of a forecast scoring rule which can be played, consider either the Probability of Detection (POD) or False Alarm Rate (FAR) (or any combination of them) commonly used in severe weather operations. If one wishes to get the highest possible POD, he simply issues warnings all the time for his area of responsibility. If he wishes to minimize his FAR, he issues no warnings. In cases where, because of low population density or time of day, a forecaster does not think there is a good chance of verifying his warning, he may not issue a warning if he is concerned about his verification score.

Since there are six variables being forecast for each forecast period (two days—high and low temperature and precipitation amount), the maximum possible score on a forecast was 6. Since the range of scores is typically not large for an individual forecast day (high~5.8-5.9 and low~5.2-5.3), the scores were compared to a standard forecast and then multiplied by 100. Here the standard forecast is MOS, although that is not critical. Positive scores indicate forecasters who did better than MOS and negative scores indicate those who did worse.

### 3. GENERAL RESULTS

Of the 24 forecasters, 13 outperformed MOS for the length of the contest. Only one beat the consensus forecast. As a first look at the data, let us consider the performance of forecasters versus the spread they attached to their forecast, as determined by the average value of Shape for their temperature forecasts. (There was less variability in the value of Shape for precipitation forecasts due to the smaller number of categories and, more importantly, the relatively large chance assigned by most forecasters to no or little rain for most forecasters. The first two months of the contest averaged approximately 2/3 of the normal precipitation and most of the early forecasts verified with no precipitation and were forecast as such by most forecasters.) Figure 4 shows the average Shape versus performance compared to MOS over the course of the contest. Recall that the maximum value for Shape is 0.5, corresponding to a probability of 1 in a single category and that lower values indicate more spread out forecasts. Shape and performance are negatively correlated at a significance level of 99.99%. In other words, more peaked forecasts (which should signify an increased level of confidence) resulted in lower performance.

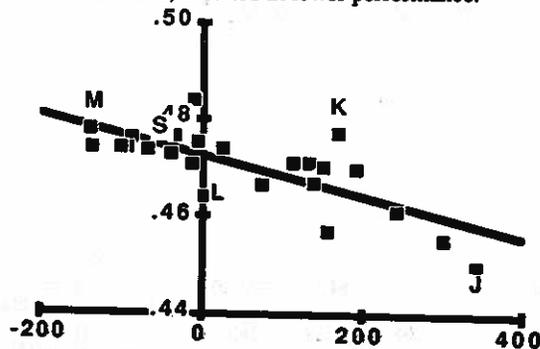


Figure 4—Average shape of temperature forecasts versus performance by human forecasters in contest. 0 line on abscissa is MOS score. Diagonal line is least squares fit to points. Small values on ordinate indicate less peaked forecasts. Letters indicate individual forecasters who will be studied in Section 4.

Another statistic of interest is the consistency of performance. We define a simple measure of this as the average forecast-to-forecast change in standing for an individual. In other words, a person finishing 5th one day and 12th the next would get a score of 7. The average change between forecasts, with MOS and consensus taken out of the standings, versus performance is shown in Figure 5. The better forecasters tended to be more consistent (significant at the 94% level), as would be expected. Three groups can be loosely defined from this chart. The first is the top three forecasters who also were consistent in their day-to-day performance. A second group is the next seven participants who also had intermediate consistency. The final group is made up of the lowest 14 forecasters who tended to move more than the 8 places that would occur from random placement. This is in agreement with past experience that the standings in forecast contests often reflect the number of poor forecasts made by individuals. The leading forecasters rarely made bad forecasts, thus limiting how far they could move from forecast to forecast.

Another question we wanted to look at was the influence of immediate past performance on the next forecast. (On Tuesday's forecasts, the verification period from the previous forecast had been over for three days. On Thursday, typically, 4 of the 6 variables had verified and everyone had a good idea as to how the others would come out.) We look at this in two ways. The first is the average placement of the forecaster in forecast  $n+1$  who finished in any given place in forecast  $n$ . This is shown in Figure 6. Most of the places in the standings from day  $n$  cluster

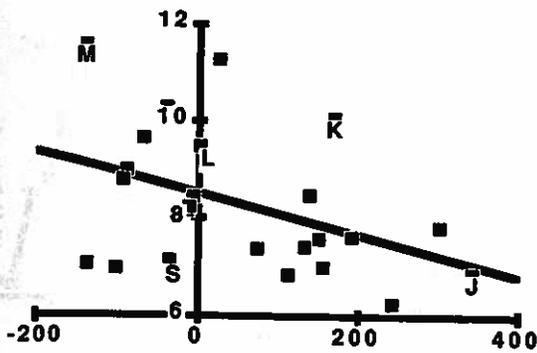


Figure 5—Same as in Figure 4 except for consistency of forecasts versus performance. Small values on ordinate indicate more consistent performance.

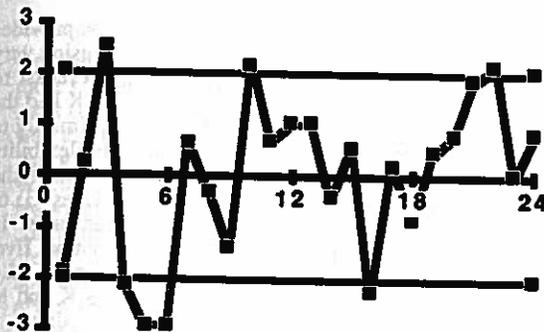


Figure 6—Performance on forecast  $n+1$  versus performance on forecast  $n$ . Average performance of 12.5 is subtracted from ordinate. Horizontal lines indicate 2 places above mean and 2 places below mean performance.

around 12.5, the mean standing, on day  $n+1$ . There is a tendency for finishers in the range 4 through 6 to do better than the mean on their next forecast, although this is not significant at this sample size. Some participants have speculated, however, that the effect was real and may result from a desire to improve the forecast technique just a little.

The second, and perhaps more meaningful, method of considering the influence of past forecasts is to look at how forecasters changed the shape of their forecasts with past performance. Five people showed correlations significant at the 90% level between their performance on day  $n$  and the shape of their forecast on day  $n+1$ . Four of the five were correlated in the sense that a poor forecast was followed by a more peaked (or bold) forecast and a good forecast was followed by conservative forecasting. (A total of 15 were positively correlated in that same sense, but not only four at the 90% level.) These four finished 1st, 4th, 11th, and 16th in the final standings. There are two possible explanations for this forecast strategy (assuming that any conscious or unconscious went into it), depending on how the last forecast went. If a person had a poor forecast, they may decide to attempt to make up ground quickly by becoming bolder. In the event of a good forecast, the forecaster might become more conservative in an effort to hold onto the gains made previously. Correlations of the other sign would indicate that a successful forecast would lead to more confidence and a bolder next forecast. It is interesting to note that of the top nine finishers, only one (2nd place) exhibited a strategy. The only significant correlation in that direction was from the 12th place finisher.

Place	Shape	Cons.	Flex.	
1	1	3	6	J
2	2	12	8	
3	4	1	4	
4	8	11	14	
5	22	21	19	K
6	3	5	7	
7	9	10	3	
8	7	16	10	
9	12	9	21	
10	11	2	23	
11	6	8	15	
12	15	23	5	
13	5	19	1	L
14	19	15	18	
15	24	14	2	
16	10	13	22	
17	21	7	9	S
18	13	22	12	
19	14	20	17	
20	20	18	16	
21	16	17	20	
22	18	4	24	
23	17	6	11	
24	23	24	13	M

Table 2—Three measures of forecast philosophy versus place in final standings. Shape is average shape of temperature forecast with low numbers indicating people who used less peaked distributions. Cons. is measure of consistency. Small numbers indicate more consistent forecast. Flex. is flexibility, with small numbers indicating more forecast-to-forecast change in shape parameter. The letter indicates the identifier for those mentioned in the individual performance section.

The final point we would like to look at about the behavior of the group in general concerns the "flexibility" of the forecasters. By this, we mean how much they were willing to change the average shape of their forecasts from day to day, instead of following the same forecast each time out. One would expect that the leading forecasters would show a greater day-to-day change, indicating a tendency to evaluate each situation on its own merits and show more confidence in "easier" forecast situations. This is the case in this sample, at least to a certain extent. Table 2 summarizes the flexibility of the forecasters (as well as the shape and the consistency as a function of place in the final standings). Of the eight most flexible forecasters, five of them finished in the top seven in the standings.

Our composite picture of the best forecaster then is one who assigns probabilities (even small ones) over a wide range, is consistent in his performance, and who is flexible in terms of how he approaches each forecast. A poor forecaster would be expected to display none of these traits. In the next section, we will examine how five individuals performed, including ones who fit this conceptual model of forecasting and some who do not.

#### 4. INDIVIDUAL PERFORMANCES

Figure 7 shows the forecast-to-forecast performance of the contest winner, J. J was particularly strong in the early and late portions of the contest. As seen in Table 2, J made the most smooth (least confident?) forecasts in the contest and was among the most consistent and flexible forecasters. J fits the description of the good forecaster that we proposed earlier. J finished in the top five in 13 out of the 24 forecasts, five more times than the second most frequent person. As a sidebar, J's worst performance occurred when a cold frontal passage just before the beginning of the day 1

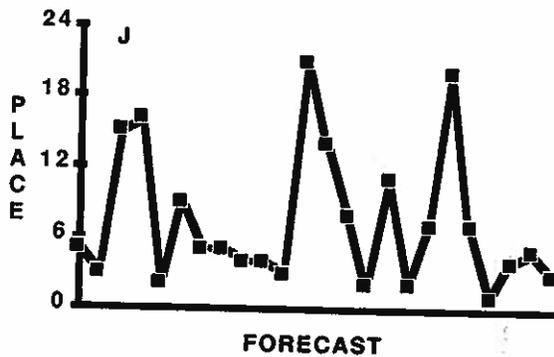


Figure 7—Finish on each forecast by Forecaster J. There are 24 forecasts and the standings do not include MOS or consensus.

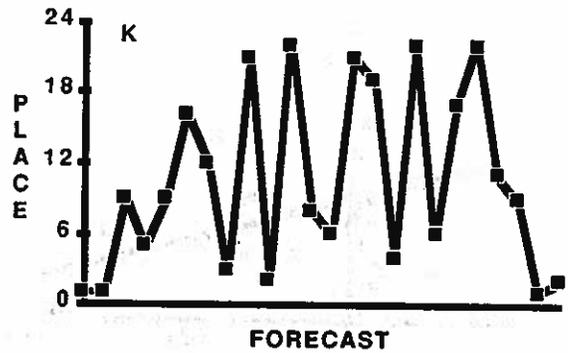


Figure 9—Same as Figure 7 for Forecaster K.

forecast period dropped the temperature 15 degrees in the hour before verification of the high began. For that forecast, the last-place forecaster in the contest won the daily game. J's other bad forecast came when the NGM had Champaign inside the 1 inch accumulated rain contour for 2 consecutive 12 hour periods and, officially, only a trace fell. J typically followed poor forecasts with good forecasts, finishing in the lower half of the daily game consecutively only twice. By making smooth forecasts, thereby accounting for the possibility of rare events, J was cushioned from the possibility of extremely poor forecasts.

At the other extreme, forecaster M finished last in the contest. M's forecast-to-forecast performance is shown in Figure 8. M made one of the most peaked forecasts of anyone in the contest and, as a result, was the least consistent forecaster in the contest. M won two daily forecasts (only one person, forecaster K won as many as three) and finished in the top three 4 times (only three people did so more often), but finished in the bottom three 6 times. M's two firsts were sandwiched around a 21st place forecast. Over half of the time, M finished more than 10 places away from the previous finish. This is the sort of behavior one would typically expect from a peaked, "confident" forecast. If the forecast is right, it is very good, but if it is wrong, it is very bad. There was a large difference between M's performance on Tuesdays and Thursdays. On Tuesday, M average finish was 9.7, while on Thursday it was 18.7. (M was actually 4th in the standings if only Tuesday forecasts were counted.) It is possible that this is due to some effect of completely knowing the previous results. There are also a large number of other causes possible for this occurrence.

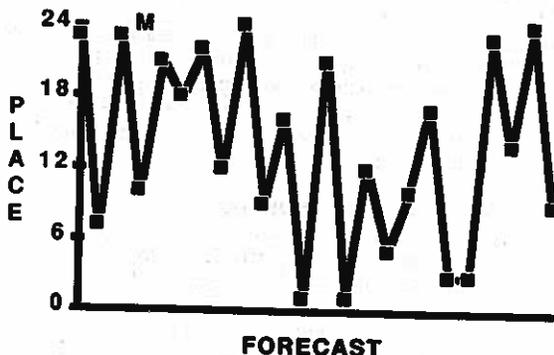


Figure 8—Same as Figure 7 for Forecaster M.

K, whose performance is shown in Figure 9, provides us with an example of a forecaster who, despite using very peaked forecasts and being inconsistent from forecast to forecast, still finished near the top of the contest. K had the 3rd most peaked forecasts in the contest. Due primarily to this, K's finishes were the 4th least consistent. In flexibility of forecast strategy, K was also near the bottom of the contest, finishing 19th. K had the most first places (3) of anyone, but also finished in the bottom three four times. In half of the forecasts, K was either in the top or bottom five. On Tuesdays, K finished 7.4 (2nd) and on Thursday 12.3 (12th) for an overall 5th place. No one besides K and M was more than 3.1 places different between the two days.

The fourth person of interest is Forecaster L (Figure 10), who made unpeaked forecasts on average (5th smoothest), but was still inconsistent (6th most inconsistent) and did not perform as well as would be predicted by shape of forecast (13th overall). L, however, was among the least experienced forecasters and experimented with different strategies more than most. L was by far the most flexible forecaster, reflecting efforts to find a strategy. Interestingly, L was one of the forecasters most influenced by immediate previous performance. Even though it was not significant at the 90% level, L made more peaked forecasts after a good performance, in general. L's case will be subjected to more careful analysis since L represents a subclass in the contest with the combination of lack of experience and attempts to experiment with strategy. Surprisingly, some of the other inexperienced forecasters made few forecast-to-forecast changes (this is not preclude the possibility that they made significant changes over longer time scales.)

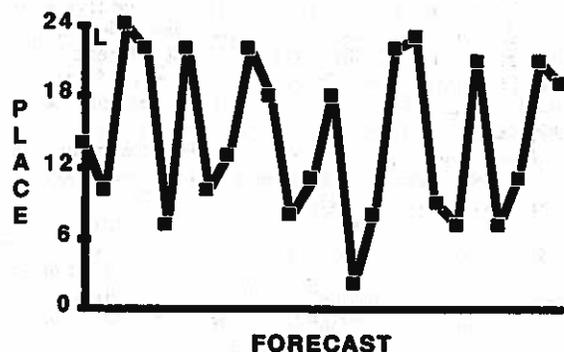


Figure 10—Same as Figure 7 for Forecaster L.

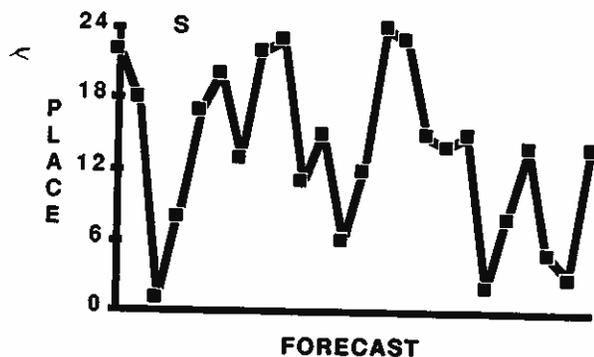


Figure 11—Same as Figure 7 for Forecaster S.

The final individual we will consider is S (Figure 11), represents yet another rare breed, that of the peaked (4th most peaked) forecaster who is nonetheless consistent (7th most consistent). S finished 17th overall in the contest and was a relatively good forecaster the last five weeks, after being a relatively poor forecaster the preceding five. Whether this represents a change in the understanding of the individual synoptic situations, thereby justifying the inherent confidence of the forecaster, or a more subtle effect is not clear. Certainly there was not a significant change in approach to distributing probabilities. It is possible that, given the small sample size, that this simply represents another arrangement of the wildly volatile performance shown by M. A longer data record would be necessary to resolve that. Perhaps the most remarkable forecast performance of any was S's high temperature forecast on the cold frontal passage day mentioned earlier. S put 40% in category 4, and 30% each in category 5 and 10, reasoning that if the front made it to Champaign before midnight, the high would be cold and, if it didn't it would still be warm. In fact, the high (if verification would have begun 75 minutes earlier) would have been in category 10 and it ended up verifying in category 5. This was the only split distribution seen in temperature forecasting for the contest.

##### 5. CONCLUDING REMARKS

Perhaps there is nothing particularly startling about these results. We would like to raise some of the issues mentioned earlier and hope to continue looking at them ourselves. The effects of forecaster confidence and method of making the forecast are not entirely clear. From this small data set, it would appear that forecasters tend to become more conservative (smoother forecasts) after making a good forecast. We have certainly seen that, within the range of this scoring method, more spread out forecasts, acknowledging the existence of unlikely events tend to do better than more narrowly defined forecasts. They also tend to perform a little more consistently, rarely being wildly wrong. This, we feel is an important point for forecasting that has been made in the past: A forecaster needs to recognize the limitations of his knowledge, the data, and the science in general to make the best possible forecast. Often, the most critical forecast (or nowcast decisions) are based on the edge of the knowledge and data. An extreme observation, for instance a large pressure drop at a station, may either be bad data or the single clue to a rapidly developing severe situation. Failure to accept the latter possibility, even though the former may be more likely, could have serious ramifications.

Finally, the forecast contest in this form represents a stage in the public forecasting procedure. Certainly, we would not wish to deliver to the public a statement on the

order of "the high will be between 50 and 70", but the possibilities of the high being in that range should be evaluated in the process of making the forecast. Ideally, a forecaster should not ever have an event occur to which he has assigned a probability of 0. When that occurs, he has in effect said that the verification was impossible from the previous situation, which is a very different thing than saying it is unlikely. We feel that this procedure of assigning probabilities is something that has to happen every time a forecast is made, at least at the level of being aware of all the possibilities.

##### 6. ACKNOWLEDGMENTS

We would like to thank the participants in the contest for their effort, time, and comments on the contest. In particular, we thank Ali Tokay and Block Andrews for administering the contest. Discussions with John Walsh and William Chapman were extremely helpful. This work is supported in part by NSF grant ATM 87-00778.

##### 7. REFERENCES

- Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, 8, 985-987.
- Murphy, A. H., 1969: On the "Ranked Probability Score." *J. Appl. Meteor.*, 8, 988-989.
- \_\_\_\_\_, and E. S. Epstein, 1967: A note on probability forecasts and "hedging." *J. Appl. Meteor.*, 6, 1002-1004.
- Powell, J. L., 1987: A test of two factors influencing the decision rules in judgment tasks. Ph. D. thesis, University of Missouri—Saint Louis, 153 pp.