RP–73–17
Reprint

[Reprinted from BULLETIN OF THE AMERICAN METEOROLOGICAL SOCIETY, Vol. 55, No. 10, October 1974, pp. 1206–1216]
Printed in U. S. A.

# subjective probability forecasting experiments in meteorology: some preliminary results[1]

Allan H. Murphy [2]
National Center for Atmospheric Research [3]
Boulder, Colorado 80303

Robert L. Winkler
Indiana University
Bloomington, Indiana 47401

## Abstract

This paper describes the preliminary results of three experiments in subjective probability forecasting which were recently conducted in four Weather Service Forecast Offices (WSFOs) of the National Weather Service. The first experiment, which was conducted at the St. Louis WSFO, was designed to investigate both the ability of forecasters to differentiate among points in a forecast area with regard to the likelihood of occurrence of measurable precipitation and their relative ability to make point and area (including areal coverage) precipitation probability forecasts. The second experiment, which was conducted at the Denver WSFO, was designed to investigate the ability of forecasters to use credible intervals to express the uncertainty inherent in their temperature forecasts and to compare two approaches (variable-width intervals and fixed-width intervals) to credible interval temperature forecasting. The third experiment, which was conducted at both the Great Falls and Seattle WSFOs, was designed to investigate the effects of guidance (i.e., PEATMOS) forecasts upon the forecasters' precipitation probability forecasts.

For each experiment, some background material is presented; the design of the experiment is discussed; some preliminary results of the experiment are presented; and some implications of the experiment and the results for probability forecasting in meteorology and probability forecasting in general are discussed. The results of each of these experiments will be described individually and in much greater detail in a series of forthcoming papers.

## 1. Introduction

In a meteorological context and in other contexts as well, probability forecasts serve two basic purposes: 1) they provide forecasters with a means of expressing the uncertainty inherent in their forecasts and 2) they provide users of such forecasts with information needed to make rational decisions in uncertain situations. The recognition of these desirable characteristics of probability forecasts led the National Weather Service (NWS), after several years of experimentation, to institute their nationwide probability of precipitation (PoP) forecasting program in 1965. Under this program, precipitation probabilities are appended to public weather forecasts issued by the NWS (e.g., "the probability of precipitation today is 30%"). While PoP forecasts encountered some initial resistance on the part of both forecasters and the general public, the evidence presently available suggests that these forecasts are now considered to be an important and integral part of the public weather forecasting program (e.g., American Telephone and Telegraph Company, 1971; Bickert, 1967; and Murphy and Winkler, 1973).

We believe that the NWS's PoP forecasts represent a significant advance in the practice of weather forecasting. On the other hand, many aspects of probability forecasting, both in meteorology and in other contexts, are in need of further detailed investigation, both from the standpoint of theoretical studies and from the standpoint of practical and/or experimental studies.[4] The latter are of particular concern in this paper, and several areas in which experimental studies are needed can be readily identified.[5] First, since a PoP forecast relates to the probability of occurrence of measurable precipitation (i.e., $\geq 0.01$ inch) at a point in the forecast area (in general, at the official raingage), a PoP forecast, as such, is not a particularly "rich" forecast. That is, many potential users of such forecasts are concerned with a precipitation "threshold" other than 0.01 inch (or with several such thresholds). Moreover, many users are interested in the occurrence of precipitation, however defined,

[1] Supported in part by the National Science Foundation under Grant GA-31735. This paper represents a revised version of Murphy and Winkler (1974).

[2] On leave and visiting the International Institute for Applied Systems Analysis, Laxenburg, Austria (July 1973 to January 1974), and the Department of Statistics, Stanford University, Stanford, Calif. (February–August 1974).

[3] The National Center for Atmospheric Research is sponsored by the National Science Foundation.

[4] For a recent review of probability forecasting in meteorology, refer to Murphy (1972) or Julian and Murphy (1972).

[5] The discussion which follows was purposely structured in such a way as to provide a framework for the experiments to be described in this paper, and this discussion was not intended to be comprehensive. In this regard, many other aspects of probability forecasting in meteorology (e.g., the use of feedback by forecasters, the formulation of a consensus among forecasters) are also in need of study from an experimental standpoint.

at points in the forecast area for which the probability of precipitation may be quite different than that at the official raingage. The present practice of issuing an average or uniform point probability forecast for the entire area clearly does not satisfy the requirements of such users (unless the point probability is indeed the same at every point in the area). In any case, a need exists 1) to investigate both the feasibility of making point precipitation probability forecasts for several points in a forecast area and the relationships between point and area probabilities and 2) to investigate the feasibility of making probability forecasts of precipitation amounts.

Second, uncertainty exists in the forecasts of all of the variables presently included in public weather forecasts; yet these forecasts, with the exception of those relating to precipitation occurrence, are still expressed in categorical terms. Clearly, experiments and other studies of a practical nature are needed to determine the feasibility of expressing forecasts of variables other than precipitation occurrence in probabilistic terms. These experiments should involve forecasts of continuous variables, such as temperature, as well as forecasts of discrete variables, such as the "present weather" classes.

Third, little if anything is presently known about the process by which a forecaster aggregates information in formulating probability forecasts (or forecasts in general for that matter). In view of 1) the desirability of reducing the amount of information that forecasters have to assimilate in formulating their forecasts and 2) the existence of normative models of information aggregation in probabilistic prediction (see Winkler and Murphy, 1973a), experimental studies of the aggregation process could lead to increases in the "efficiency" of the overall forecasting process.

Many different experiments could, of course, be formulated in each of the above-mentioned areas. We designed three experiments in subjective probability forecasting, one in each of these areas, and conducted these experiments in four Weather Service Forecast Offices (WSFOs) of the NWS. The first experiment involved point and area precipitation probability forecasts, the second experiment involved credible interval temperature forecasts, and the third experiment involved the effects of guidance (i.e., PEATMOS, which stands for "Primitive Equation and Trajectory Model Output Statistics") forecasts upon the forecasters' PoP forecasts. The three experiments are discussed in Sections 2, 3, and 4, respectively, of this paper. In each case, some background material is presented; the design of the experiment is discussed; some results of the experiment are presented; and some implications of the experiment and the results for probability forecasting in meteorology and probability forecasting in general are discussed.[6] Section 5 contains a brief summary and some concluding comments.

[6] We intend to describe the results of each of these experiments individually and in much greater detail in a series of forthcoming papers.

## 2. An experiment involving point and area precipitation probability forecasts

### *a. Point and area precipitation probability forecasts*

As indicated in Section 1, PoP forecasts are now issued on a regular basis by the NWS, and NWS forecasters have a considerable amount of experience at preparing such forecasts. The official definition of the probability which constitutes a PoP forecast is an average point probability of measurable precipitation for a forecast area (e.g., a metropolitan area). A point probability of precipitation is the probability of precipitation at a given point, and an average point probability of precipitation for a particular area is simply the average of the point probabilities of precipitation for all of the points in the area. In the forecasts formulated by NWS forecasters, the point probability is, in general, implicitly assumed to be uniform over the forecast area (i.e., the probability is the same for all of the points in the area). Under this assumption, the precipitation probability issued to the public applies to each point in the forecast area. The observation of precipitation, on the other hand, is taken at only one point (the official raingage). When the (point) probability of precipitation varies over the forecast area, forecasters *occasionally* issue two (or more) forecasts, each of which is applicable to a different part of the area. However, we believe that different PoP forecasts should be *routinely* provided for each portion of the area to which the forecaster is able to assign a different probability of precipitation, since the use of an average point probability for an entire forecast area will frequently be quite misleading.

Another potential problem concerns the interpretation of a precipitation probability by the public and by forecasters. Some members of the public may interpret a precipitation probability in terms of an area probability (the probability that precipitation will occur somewhere in the forecast area), an expected areal coverage (the expected fraction of the forecast area over which precipitation will occur), or yet some other definition. Moreover, some forecasters may have a definition other than the official definition in mind when making a precipitation probability forecast. In a recent questionnaire administered to almost 700 NWS forecasters (Murphy and Winkler, 1973), the responses indicated that different forecasters prefer different definitions of the event "precipitation" and of a precipitation probability, and, as a result, they often use definitions other than the official definitions in preparing their precipitation probability forecasts.[7]

The relationship between point and area precipitation probabilities has been studied theoretically (e.g., Epstein, 1966) but not experimentally. The experiment reported in this section was designed to investigate both the rela-

[7] For example, factors such as precipitation type, precipitation amount (i.e., a trace versus a measurable amount), and topography apparently cause forecasters to use different definitions in different situations.

tive ability of forecasters to make point and area (including areal coverage) probability forecasts and the ability of forecasters to differentiate among different points in a forecast area with regard to the likelihood of the occurrence of measurable precipitation.

*b. Design of the experiment*

The subjects in the experiment were 14 weather forecasters from the WSFO at St. Louis, Mo. Each time the forecasters were on public weather forecasting duty, they made point and area precipitation probability forecasts for the St. Louis metropolitan area. In particular, the forecasters were asked for 1) an average point probability of measurable precipitation for the entire forecast area; 2) point probabilities of measurable precipitation at five specific points (raingages) in the forecast area; 3) an area probability of measurable precipitation for the forecast area; and 4) the expected areal coverage of the forecast area by measurable precipitation. On each occasion, the forecasts were made for three different 12-hr periods (on the day shift, for "tonight," "tomorrow," and "tomorrow night"; on the midnight shift, for "today," "tonight," and "tomorrow"). The experiment was conducted from November 1972 to March 1973.

Observations from the Illinois State Water Survey network of raingages in the St. Louis area were used to verify the forecasts. The network included raingages at the five points for which point probabilities of precipitation were determined by the forecasters. In addition, a larger set of 20 raingages was used to verify the forecasts of area probability and expected areal coverage. Within the constraints imposed by the location of available raingages, the smaller set of five points and the larger set of 20 points were chosen to provide a reasonable coverage of the St. Louis metropolitan area (the set of 20 raingages defined a circular area with a radius of approximately 30 n mi centered at a point near the Arch).

*c. Some results of the experiment*

First, the point precipitation probabilities exhibited little variability over the forecast area. The sample variance was computed for each set of five point probability forecasts, and the average value of the variance was 0.001. This average sample variance is especially small considering that, with the exception of very small probabilities, any difference in probabilities must be of magnitude 0.10 at least.[8]

Next, the assessed average point probability and the average of the five individual point probabilities were compared. Since the average point probability was to be verified over a network of 20 raingages rather than just five raingages, this probability (denoted by $A$; for definitions of the symbols used in this section, refer to Table 1) could differ from the average of the five individual point probabilities (the individual probabilities are denoted by $B_1$, $B_2$, $B_3$, $B_4$, and $B_5$, and their average is denoted by $\bar{B}$), although we would not expect the difference to be large. In fact, the average value of $|A - \bar{B}|$ was only 0.005 (standard error = 0.0006), and the average value of $A - \bar{B}$ was 0.001 (standard error = 0.0007). In 663 (86.1%) of the cases, $A - \bar{B}$ was equal to zero, and the largest value of $|A - \bar{B}|$ was 0.24. In fact, $|A - \bar{B}|$ was larger than 0.05 in only 15 (1.9%) of the cases. Furthermore, a plot of $A - \bar{B}$ versus the sample variance of the five individual point probabilities reveals that no readily discernible relationship exists between these two variables (i.e., the variability of the five individual point probabilities is not related to the difference between the two average point probabilities).

TABLE 1. Definitions of symbols used in discussion of results of point and area precipitation probability forecasting experiment.

| Symbol | Definition |
|---|---|
| $A$ | Average point probability for forecast area |
| $B_1, \ldots, B_5$ | Individual point probabilities for five points in forecast area |
| $\bar{B}$ | Average of five individual point probabilities |
| $C$ | Area probability for forecast area |
| $D$ | Expected areal coverage of forecast area |

Another comparison of interest is that of average point probability and the expected areal coverage (denoted by $D$). Mathematically, $A$ and $D$ should be equal, since

$$A = (1/k) \sum_{i=1}^{k} p_i$$

and

$$D = E[(1/k) \sum_{i=1}^{k} \delta_i] = (1/k) \sum_{i=1}^{k} E(\delta_i) = (1/k) \sum_{i=1}^{k} p_i,$$

where $k$ represents the number of raingages, $p_i$ is the probability of precipitation at raingage $i$, and $\delta_i$ is an indicator variable that equals one if precipitation occurs at raingage $i$ and zero otherwise. From the experiment, $A - D = 0$ on 715 (92.9%) of the occasions, and the average value of $A - D$ was $-0.0005$ (standard error = 0.001). The average value of $|A - D|$ was 0.0007 (standard error = 0.001), and the largest value of $|A - D|$ was 0.30. In only 32 (4.2%) of the cases was $|A - D|$ larger than 0.05.

Another result of interest relates to the area probability (denoted by $C$). Theoretically, the area probability must be at least as large as any point probability, since precipitation at any point implies precipitation in the

[8] The numbers that could be used for point probability forecasts were limited to 0.00, 0.02, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, and 1.00.

area. A comparison of $C$ with $\max_i(B_i)$ yielded the following results: $C$ was smaller than the largest point probability on only 59 (7.7%) of the occasions, and, of the remaining 711 occasions, $C = \max_i(B_i)$ on 685 (89.0%) of these occasions. The average value of $C - \max_i(B_i)$ was actually slightly negative (−0.004, with a standard error of 0.0017), and the smallest value of $C - \max_i(B_i)$ was −0.30. These results indicate that the forecasters had misconceptions concerning the point probabilities or the area probability, or both. The consistency of the point probabilities, the average point probability, and the expected areal coverage indicates that these difficulties are most likely to be related to the area probability.

The final analysis to be described in this section is an investigation of the difference between $A$ and $CD$. According to the definitions given to the forecasters, $A$ should be greater than $CD$, with equality holding only when $C = 1$. In the experiment, $A$ was in fact greater than $CD$ for 702 (91.2%) of the forecasts. On the other hand, $A$ was less than $CD$ for only 10 (1.3%) of the forecasts. This result indicates that, as instructed, the forecasters thought of $D$ in a marginal sense rather than in a conditional sense. It is possible to consider a conditional expected areal coverage, which would be the expected areal coverage *given* that precipitation will occur somewhere in the forecast area. Such a conditional expected areal coverage must be at least as large as $D$, the marginal expected coverage. Specifically, the conditional expected areal coverage should equal $A/C$, whereas the marginal expected areal coverage, $D$, should equal $A$. Thus, the conditional measure should be larger than the marginal measure by a factor of $1/C$.

*d. Discussion*

The results presented in this section indicate that little variability existed among the point probability forecasts for the five points for which such forecasts were made. This result may be due to the fact that the St. Louis forecast area is not subject to any pronounced local effects (such as a topographical effect) or to the fact that the experiment was conducted in the winter, which tended to minimize the effects of mesoscale weather systems. On the other hand, perhaps the variability should have been greater, but the forecasters were simply unable to differentiate among the points more often (i.e., as often as they should). In this regard, the differences among the relative frequencies of precipitation at the five points of concern are of particular interest. An analysis of these data (as well as other analyses; see below) was delayed because the precipitation observations were not immediately available.

The forecasters were remarkably consistent when assessing the average point probability, the five individual point probabilities, and the expected areal coverage. Of course, this result may not generalize to more complex situations in which greater variability exists among the individual point probabilities, and this question can and should be investigated experimentally.[9] The area probability tended not to be consistent with the point probabilities; the former was frequently too low, even lower than some of the point probabilities, and this result is inconsistent. In general, the area probability should be greater than or equal to each individual point probability, with equality holding only when any precipitation that occurs in the entire area is certain to occur at the point in question.

The analyses presented in this section involved only the probabilities assessed by the forecasters. Further analyses along these lines are being conducted, including a more detailed investigation of the relationships among the different types of probabilities and a study of the effects of factors such as the individual forecaster, the forecast shift (i.e., day, midnight), and the forecast length. In this regard, the point precipitation probability forecasts of certain forecasters appear to be more variable than do those of other forecasters. In addition, we are analyzing the forecasts in light of the actual observations, which were not available at the time these initial analyses were undertaken. This portion of the analysis includes a study of the relationship between the probabilities and the observed relative frequencies for each type of probability determined in the experiment.

The experimental results presented herein have implications with regard to the importance of carefully defining variables in probability forecasting. If a forecaster uses a definition of a precipitation probability that differs from the official definition prescribed by the NWS, then he is likely to arrive at a different probability than he would if he used the official definition. Of course, this implication holds with respect to probability forecasting in general and is by no means limited to precipitation probability forecasting. In any case, even if the official definition is used for forecasting purposes, a better understanding of the relationships among the various types of probabilities may improve the forecaster's ability to formulate probability forecasts.

## 3. An experiment involving credible interval temperature forecasts

*a. Credible interval temperature forecasts*

Precipitation occurrence lends itself quite well to the use of probabilities since this variable is a simple dichotomy once the event "precipitation" is defined, and, as a result, only a single probability is needed to express

[9] An important consideration in the selection of the St. Louis WSFO for this experiment was the existence of a reasonably dense network of recording raingages in the St. Louis metropolitan area. We plan to undertake a similar experiment in a location in which prominent local effects exist and during a period of the year in which mesoscale systems are of greater importance. The availability of a network of recording raingages will, necessarily, remain a consideration in the selection of a suitable location for such an experiment.

a forecaster's uncertainty about the occurrence of this event. On the other hand, a continuous variable such as temperature requires a different type of probability forecast than does a dichotomous variable such as precipitation occurrence. Ideally, an entire probability distribution would be assessed, but such a distribution is not practical in terms of the time required of the forecaster or in terms of reporting to the general public. One way to summarize a probability distribution is in terms of one or more credible intervals, which are intervals of values of the variables of interest (here, maximum and minimum temperature) together with the probabilities associated with the intervals. The current operational procedure in forecasting temperature is to give either a point forecast or an interval forecast (for example, tomorrow's maximum temperature will be 65° or will be between 63° and 67°, respectively). However, a probability is *not* assessed when an interval forecast is made, so that potential users of the forecasts are presently unable to determine the uncertainty inherent in any particular temperature forecast. (In the above example, the probability associated with the interval from 63° to 67° could be 0.50, 0.99, or almost any other value. The usefulness of an interval forecast is greatly reduced if the probability associated with the interval is not provided. *Note:* As the unit in which all temperatures in this paper are expressed is the Fahrenheit degree, the unit symbol, F, will be omitted.)

Given that credible intervals are to be used in forecasting maximum (high) and minimum (low) temperature, the next question concerns the selection of particular intervals. In an earlier experiment, Peterson, Snapper, and Murphy (1972) used variable-width credible intervals in temperature forecasting. Variable-width credible intervals are intervals for which the probability is fixed in advance but the width of the interval will vary from situation to situation. For instance, if the probability is fixed at 0.50, then on some occasions a 50% credible interval for high or low temperature will be only 2° wide, while on other occasions the interval may be 5° wide.

An obvious alternative to variable-width forecasts is a forecast for which the width of the interval is fixed but the probability associated with the interval will vary from situation to situation. For instance, a forecaster might be asked to report a credible interval that is exactly 5° wide. In some situations the probability of such an interval might be 0.50, whereas in other situations the probability might be 0.90. Such a forecast will be called a fixed-width credible interval.

As a result of their experiment, Peterson, Snapper, and Murphy (1972, p. 969) concluded that "weather forecasters can use credible intervals to describe the uncertainty inherent in their temperature forecasts." The experiment reported in this section was designed to investigate further the ability of forecasters to use credible intervals in temperature forecasting and to compare two approaches (variable-width intervals and fixed-width intervals) to credible interval temperature forecasting.

*b. Design of the experiment*

The subjects in the experiment were four experienced weather forecasters from the WSFO at Denver, Colo. Each time the forecasters were on public weather forecasting duty, they made forecasts of high and low temperatures. On the day shift, the forecasts were for "tonight's low" and "tomorrow's high," whereas on the midnight shift the forecasts were for "today's high" and "tonight's low." Because the forecasters' schedules rotated them to other duties (e.g., aviation forecasting) on a regular basis and because of vacations and other leaves, more than six months were required to obtain 30 or more sets of forecasts from each participant. The data analyzed in this section were collected over a period from August 1972 to March 1973, and the four participants made 30, 31, 32, and 34 sets of forecasts, respectively.

Two of the forecasters worked within the framework of variable-width, fixed-probability forecasts, using 50% and 75% central credible intervals. (A "central" credible interval is defined as an interval for which the probabilities of being below and above the interval are equal.) The intervals were obtained by asking the forecaster to make a total of five indifference judgments at equal odds, thereby determining the median, the 25th percentile, the 12-1/2th percentile, the 75th percentile, and the 87-1/2th percentile, in that order. This process provided the forecaster with a systematic procedure for determining credible intervals. The forecaster then was asked to examine the resulting credible intervals to see if they seemed reasonable in the sense of adequately representing his judgments concerning the high or low temperature.

The other two forecasters in the experiment worked within the framework of fixed-width, variable-probability forecasts, using intervals of width 5° and 9°. First, the median of the forecaster's distribution was determined, just as in the case of the variable-width forecasts. Then, the forecaster was asked to assess probabilities for intervals of width 5° and 9° centered at the median. If the fixed-width intervals had not been centered in width at the median, then the forecasters would have had to report temperatures (i.e., at least one end point of each interval) as well as the probabilities. All intervals in the experiment were assumed to include their end points, and all temperatures were expressed to the nearest degree (thus, the interval from 63° to 67° included the five temperatures from 62.5° to 67.5°).

Prior to the start of the experiment, the authors met with the forecasters and discussed the concept of credible interval temperature forecasts. Following this meeting, lengthy sets of instructions were given to the participants, who were encouraged to read the instructions, to make several "practice" forecasts, and to discuss any difficulties with the experimenters. The instruction sets included

discussions of how credible intervals describe a forecaster's uncertainty when making temperature forecasts, careful definitions of the terminology to be used in the experiment, hypothetical dialogues between an "experimenter" and a "forecaster" to illustrate the procedures and to answer anticipated questions, and brief summaries of the procedures to insure understanding on the part of the forecasters. No difficulties arose after the instruction sets were distributed, and we believe that the participants understood the experimental procedures.

*c. Some results of the experiment*

For all of the participants in the experiment, the first task on each forecasting occasion was to determine a median. For the entire sample ($n = 254$), the average difference between the median and the observed temperature was −0.45° (standard error = 0.307°), and the average absolute difference, or average error, was 3.81° (standard error = 0.194°). Moreover, scatter diagrams suggest that the average error is not a function of the observed temperature. In general, then, the medians appear to be good point forecasts. For comparative purposes, the official forecast issued to the public was recorded on each occasion. The average difference between the official forecast and the observed temperature was −0.44° (standard error = 0.312°) and the average absolute difference was 3.91° (standard error = 0.195°). Therefore, the medians were, on the average, comparable to the official forecasts as point forecasts of high and low temperature. Of course, we would not expect the medians and the official forecasts to differ a great deal, since they were both determined by the same forecaster on almost all occasions.

For the variable-width credible intervals ($n = 132$), the observed temperature was inside the 50% credible interval 60 times (45% of the time), below the lower limit of the interval 34 times (26%), and above the upper limit of the interval 38 times (29%). These values are close to the expected percentages (50%, 25%, and 25%, respectively), and a goodness-of-fit test yields a small value of $\chi^2$ (1.333, with 2 degrees of freedom) even though the sample size is reasonably large. Similarly, for the 75% credible intervals, the observed temperature was inside the interval 97 times (73%), below the lower limit 14 times (11%), and above the upper limit 21 times (16%). These values, which are also close to the expected percentages (75%, 12.5%, and 12.5%, respectively), lead to a slightly larger value of $\chi^2$ (1.646, with 2 degrees of freedom). Thus, the observed relative frequencies are very close to the probabilities assigned to the intervals. Moreover, this result appears to be insensitive to the width of the credible interval.

The average error was expected to be an increasing function of the width of the 50% credible interval and the width of the 75% credible interval. The data presented in Table 2 indicate that a strong relationship does not exist, although a positive relationship does seem to hold for the range of widths for which a reasonable number of cases is available. The average widths are 6.2° (standard error = 0.11°) and 11.7° (standard error = 0.19°) for the 50% and 75% credible intervals, respectively.

TABLE 2. Average error as a function of interval width for variable-width interval forecasts.

| 50% Credible intervals | | | 75% Credible intervals | | |
|---|---|---|---|---|---|
| Width (°F) | Number of forecasts $n$ | Average error (°F) | Width (°F) | Number of forecasts $n$ | Average error (°F) |
| 3 | 2 | 3.00 | 6 | 1 | 1.00 |
| 4 | 9 | 2.56 | 7 | 2 | 2.50 |
| 5 | 22 | 3.09 | 8 | 7 | 3.00 |
| 6 | 44 | 4.55 | 9 | 11 | 3.64 |
| 7 | 42 | 3.98 | 10 | 12 | 2.75 |
| 8 | 6 | 2.83 | 11 | 29 | 3.83 |
| 9 | 6 | 6.00 | 12 | 25 | 3.92 |
| 10 | 0 | — | 13 | 29 | 4.86 |
| 11 | 1 | 11.00 | 14 | 6 | 3.83 |
| | — | — | 15 | 4 | 2.00 |
| Total/average | 132 | 4.00 | 16 | 3 | 10.33 |
| | | | 17 | 0 | — |
| | | | 18 | 2 | 2.50 |
| | | | 19 | 0 | — |
| | | | 20 | 0 | — |
| | | | 21 | 1 | 11.00 |
| | | | | — | — |
| | | | Total/average | 132 | 4.00 |

Another result of interest relative to the variable-width intervals concerns their symmetry or asymmetry in terms of width. For the 50% credible intervals, the difference between the 75th percentile and the median was less than (equal to) (greater than) the difference between the median and the 25th percentile on 36 (67) (29) occasions. For the 75% credible intervals, the difference between the 87-1/2th percentile and the median was less than (equal to) (greater than) the difference between the median and the 12-1/2th percentile on 43 (41) (48) occasions. In both cases, equality implies an interval symmetric in width about the median. Thus, only 51% of the 50% intervals and 32% of the 75% intervals were symmetric. The preponderance of asymmetries among the central credible intervals suggests that fixed-width credible intervals, which are constrained to be symmetric in width, are not likely to be central credible intervals.

For the fixed-width credible intervals ($n = 122$), the average probability assigned to the 5° interval was 0.60 (standard error = 0.014) and the average probability assigned to the 9° interval was 0.80 (standard error = 0.010). The overall relative frequency with which the observed temperature was inside the 5° interval was 0.46, and the overall relative frequency with which the observed temperature was inside the 9° interval was 0.66. Therefore, the probabilities assigned to the intervals by the forecasters were, on the average, larger than they should have been according to the observations.[10] In Table 3 the relative frequency of inclusion of the observed temperature

TABLE 3. Average error and relative frequency of inclusion of observed temperature in interval as a function of probability of interval for fixed-width interval forecasts.

| 5°F intervals | | | | 9°F intervals | | | |
|---|---|---|---|---|---|---|---|
| Probability of interval | Number of forecasts $n$ | Average error (°F) | Relative frequency in interval | Probability of interval | Number of forecasts $n$ | Average error (°F) | Relative frequency in interval |
| 0.30 | 2 | 3.50 | 0.00 | 0.50 | 2 | 8.00 | 0.00 |
| 0.35 | 1 | 8.00 | 0.00 | 0.60 | 6 | 4.17 | 0.50 |
| 0.40 | 22 | 4.82 | 0.23 | 0.70 | 29 | 3.97 | 0.62 |
| 0.50 | 22 | 3.86 | 0.46 | 0.75 | 5 | 3.60 | 0.60 |
| 0.60 | 31 | 3.94 | 0.35 | 0.80 | 39 | 4.18 | 0.62 |
| 0.70 | 24 | 3.25 | 0.50 | 0.85 | 4 | 3.25 | 0.75 |
| 0.75 | 3 | 2.33 | 0.67 | 0.90 | 20 | 2.70 | 0.75 |
| 0.80 | 8 | 1.12 | 1.00 | 0.95 | 4 | 3.25 | 0.50 |
| 0.90 | 7 | 2.14 | 0.86 | 1.00 | 13 | 1.69 | 0.92 |
| 1.00 | 2 | 1.00 | 1.00 | | | | |
| Total/average | 122 | 3.60 | 0.46 | Total/average | 122 | 3.60 | 0.66 |

in these intervals is given as a function of the probability assigned to the intervals. If these values were graphed, many of the points would lie far from the "ideal" diagonal 45-deg line for which the observed relative frequency for each probability exactly equals that probability.

The average error was expected to be a decreasing function of the probability assigned to the 5° central credible interval and the probability assigned to the 9° central credible interval. Although the number of forecasts available for some probabilities is limited, Table 3 indicates that the average error does tend to decrease as the probability increases.

*d. Discussion*

The results presented above indicate that the medians determined by the participants were good forecasts of observed high and low temperatures. The credible intervals also seemed to fit the observations well in an overall sense, with the variable-width intervals being better in this respect than the fixed-width intervals. In further analyses, we are investigating the effects of such factors as the temperature variable of concern (i.e., high, low), the individual forecaster, the forecast shift, and the forecast length. The relationships among some of the variables considered in the analysis presented in this section are also being examined in greater detail.

The experimental results have obvious implications for temperature forecasting. The use of probabilities, via credible intervals, in temperature forecasting allows the forecaster to express his degree of uncertainty concerning the high or low temperature. Point forecasts do not describe uncertainty, and interval forecasts without probabilities only describe uncertainty in a vague, informal manner. To the extent that these experimental results indicate that credible interval temperature forecasting is feasible and that the procedures investigated in this experiment yield reasonable results, these procedures could be very useful in temperature forecasting in practice.

Although the experiment has been oriented toward temperature forecasting, the procedures are quite general and can be used to determine credible interval forecasts of other continuous variables. Thus, the implications of the experiment extend far beyond temperature forecasting to forecasts of other meteorological variables (e.g., wind speed or the wind components) and to forecasts of variables of interest in other fields (e.g., economic indicators).

## 4. An experiment involving the effects of guidance forecasts on precipitation probability forecasts

*a. The aggregation of information in probability forecasting*

In formulating a subjective probability forecast, a forecaster intuitively assimilates information from a variety of sources and formulates judgments, in probabilistic terms, about future events, such as the occurrence of precipitation tomorrow. The responses to a questionnaire (Murphy and Winkler, 1971) indicated that the relative importance and the order of examination of information sources vary among forecasters and among weather situations, and the results of a more recent and more extensive survey of NWS forecasters (Murphy and Winkler, 1973) have provided additional evidence regarding this point. In order to study the information aggregation process experimentally, some controls on the order of examination of information sources are needed (see Winkler and Murphy, 1973a). Ideally, controls concerning all information sources would be use-

[10] While these intervals are too "tight," they are not as tight as the distributions obtained in many experiments involving probability assessments in other contexts (e.g., Alpert and Raiffa, 1969; Stael von Holstein, 1972). We attribute the forecasters' performance in this experiment to the degree of their substantive (i.e., meteorological expertise. For the most part, the participants in these other experiments were not experts in the areas of concern.

ful, but this ideal situation is very difficult to attain in an operational setting.

Guidance forecasts prepared by the NWS using a procedure called PEATMOS (Klein and Glahn, 1974) represent an information source of particular interest with regard to the precipitation probability forecasting process because the guidance forecasts themselves are expressed in probabilistic terms. PEATMOS (*P*rimitive *E*quation *a*nd *T*rajectory *M*odel *O*utput *S*tatistics) is a combination of a numerical (i.e., physical-mathematical) model and a statistical technique. This "objective" forecasting procedure determines the weather-related statistics of the output of the numerical model (e.g., the percent of the time that measurable precipitation occurs when the model predicts 80% relative humidity). The probabilities provided by PEATMOS, then, represent a source of information that is available to the forecaster in determining a precipitation probability.

Although the responses to the questionnaires mentioned above have provided some information concerning the relative importance of various information sources to forecasters, as well as the order in which they examine these sources in the process of formulating their PoP forecasts, no experimental investigations of this process have been conducted in an operational setting. The experiment reported in this section was designed to investigate the effects of the guidance (PEATMOS) forecasts upon the forecasters' precipitation probability forecasts.

*b. Design of the experiment*

The subjects in the experiment were nine experienced weather forecasters from the WSFO at Great Falls, Mont., and six experienced weather forecasters from the WSFO at Seattle, Wash. Each time they were on public weather forecasting duty, the forecasters made precipitation probability forecasts both before and after examining the guidance forecasts prepared by the NWS using the PEATMOS technique. The forecasters were instructed to examine the PEATMOS forecasts last on each occasion. That is, the pre-PEATMOS forecasts were made after the forecasters had examined all of the available information except PEATMOS. Then, the PEATMOS forecasts were observed and the post-PEATMOS forecasts were made (the latter were assumed to correspond to the forecasters' official PoP forecasts).[11]

At Great Falls forecasts were made for five locations (Billings, Glasgow, Great Falls, Helena, and Missoula), and at Seattle forecasts were made for two locations (Seattle and Yakima). On each occasion, the forecasts were made for three different periods in the future (on the day shift, for "tonight," "tomorrow," and "tomorrow night"; on the midnight shift, for "today," "tonight," and "tomorrow"). The experiment was conducted from December 1972 to March 1973.

*c. Some results of the experiment*

The three probability forecasts of interest are the pre-PEATMOS forecast (denoted by $F_1$), the PEATMOS forecast (denoted by $F_2$), and the post-PEATMOS forecast (denoted by $F_3$). The overall relationships between the probability and the observed relative frequency of precipitation for the three types of forecasts are presented in Table 4. Although firm conclusions are difficult to draw from these data, the forecasters (i.e., $F_1$ and $F_3$) at Seattle appear to be closer than PEATMOS ($F_2$) to the ideal diagonal 45-deg line for which the observed

[11] The design of this experiment was such that any differences between the forecasters' pre-PEATMOS and post-PEATMOS (or official) forecasts could reasonably be attributed to the PEATMOS forecasts. However, the fact that the PEATMOS forecasts were examined after all of the other items of information suggests that the results of the experiment will tend to underestimate the actual effect of the PEATMOS forecasts. In this regard, we plan to undertake more realistic experiments involving the aggregation of information in the probability forecasting process.

TABLE 4. Relative frequency of precipitation as a function of precipitation probability forecast (number of forecasts in parentheses).

| Probability forecast | Great Falls | | | Seattle | | |
|---|---|---|---|---|---|---|
| | Pre-PEATMOS $F_1$ | PEATMOS $F_2$ | Post-PEATMOS $F_3$ | Pre-PEATMOS $F_1$ | PEATMOS $F_2$ | Post-PEATMOS $F_3$ |
| 0.00 | 0.000 (416) | 0.012 (261) | 0.005 (398) | 0.000 (78) | 0.000 (33) | 0.000 (73) |
| 0.02 | 0.000 (5) | 0.000 (99) | 0.000 (3) | 0.000 (28) | 0.000 (25) | 0.000 (36) |
| 0.05 | 0.067 (15) | 0.008 (256) | 0.000 (26) | 0.058 (52) | 0.000 (28) | 0.019 (53) |
| 0.10 | 0.028 (957) | 0.045 (603) | 0.019 (952) | 0.095 (137) | 0.063 (79) | 0.101 (139) |
| 0.20 | 0.096 (521) | 0.093 (699) | 0.097 (526) | 0.091 (132) | 0.118 (152) | 0.083 (121) |
| 0.30 | 0.237 (262) | 0.217 (364) | 0.233 (271) | 0.264 (125) | 0.222 (158) | 0.240 (129) |
| 0.40 | 0.265 (136) | 0.365 (233) | 0.248 (149) | 0.368 (87) | 0.223 (94) | 0.396 (91) |
| 0.50 | 0.439 (171) | 0.443 (106) | 0.488 (172) | 0.373 (75) | 0.377 (61) | 0.400 (70) |
| 0.60 | 0.405 (111) | 0.478 (23) | 0.409 (110) | 0.470 (66) | 0.566 (53) | 0.443 (61) |
| 0.70 | 0.424 (33) | 0.000 (2) | 0.476 (21) | 0.511 (43) | 0.434 (53) | 0.565 (46) |
| 0.80 | 0.412 (17) | — (0) | 0.438 (16) | 0.667 (57) | 0.525 (99) | 0.673 (55) |
| 0.90 | 1.000 (2) | — (0) | 1.000 (2) | 0.793 (53) | 0.544 (90) | 0.772 (57) |
| 1.00 | — (0) | — (0) | — (0) | 0.933 (15) | 0.571 (21) | 0.824 (17) |

relative frequency over the entire sample for each forecast probability exactly equals that probability. At Great Falls, the situation was reversed. Overall, the fit to the ideal line looks quite poor, but we must remember that, although the overall sample size is large, the number of observations upon which many of the relative frequencies in Table 4 are based is relatively small.

One clear result that does emerge from Table 4 is that large differences existed in the frequencies with which various probabilities were used. In general, $F_1$ and $F_3$ tended to have quite similar frequency distributions, whereas $F_2$ was quite different. At Great Falls, the average forecasts were similar (0.198 for $F_1$ and $F_3$, 0.184 for $F_2$), but the standard deviation of the forecasts was much smaller for $F_2$ (0.139) than for $F_1$ and $F_3$ (0.177 and 0.174, respectively). At Seattle, on the other hand, the standard deviations were similar (0.282 for $F_1$, 0.291 for $F_2$, and 0.284 for $F_3$), but the average forecast was much larger for $F_2$ (0.428) than for $F_1$ and $F_3$ (0.349 and 0.351, respectively).

In terms of scoring rules, PEATMOS performed slightly better than the forecasters at Great Falls, but the reverse was true at Seattle, as indicated in Table 5. The scoring rules used were the quadratic rule ($Q$) and logarithmic rule (L), where

$$Q = \begin{cases} 100[1 - (1 - F_i)^2], & \text{if precipitation} \\ 100(1 - F_i^2), & \text{if no precipitation} \end{cases}$$

and

$$L = \begin{cases} \log F_i, & \text{if precipitation} \\ \log(1 - F_i), & \text{if no precipitation} \end{cases}$$

(see Winkler and Murphy, 1968; Stael von Holstein, 1971). The quadratic scoring rule is equivalent to the Brier, or probability, score (Brier, 1950), and, as defined, a higher score indicates better performance according to both rules. Note that at both Great Falls and Seattle, the differences between the average scores for $F_1$ and $F_3$ are quite small. On the other hand, since the scores for $F_3$ are greater than those for $F_1$ at Great Falls and the scores for $F_3$ are less than those for $F_1$ at Seattle, and the scores for $F_2$ are greater than those for $F_1$ and $F_3$ at Great Falls and less than those for $F_1$ and $F_3$ at Seattle, these results suggest that the PEATMOS forecasts may have had "positive" and "negative" effects upon the forecasts prepared at Great Falls and Seattle, respectively. However, a detailed analysis of individual forecasts must be completed before any definitive statements can be made regarding the existence of these effects. Note also that because Seattle and Great Falls experience different weather situations, the scores for Seattle and Great Falls are not directly comparable (that is, the scores in Table 5 do not necessarily imply, for example, that the forecasters at Great Falls were "better" than those at Seattle).

TABLE 5. Average scores for pre-PEATMOS ($F_1$), PEATMOS ($F_2$), and post-PEATMOS ($F_3$) forecasts.

| Type of forecasts | Great Falls | | Seattle | |
|---|---|---|---|---|
| | Quadratic score $Q$ | Logarithmic score $L$ | Quadratic score $Q$ | Logarithmic score $L$ |
| $F_1$ | 92.36 | −0.278 | 80.84 | −0.565 |
| $F_2$ | 94.35 | −0.243 | 73.53 | −0.782 |
| $F_3$ | 92.55 | −0.277 | 80.22 | −0.578 |

Since we are concerned with the aggregation of information, the change in the forecasters' assessed probabilities as a result of examining the PEATMOS forecasts is of interest. To investigate this change, we consider a ratio $T$, where

$$T = (F_3 - F_1)/(F_2 - F_1).$$

Note that $T$ is only defined for cases in which $F_2 \neq F_1$, so that the analysis is confined to those cases. The average value of $T$ was 0.18 at Great Falls and 0.20 at Seattle (standard errors 0.012 and 0.016, respectively). Thus, on the average, the forecasters shifted their forecasts about 20% of the distance from their original forecast to the PEATMOS forecast. Of course, we must keep in mind that the forecasters presumably had already observed all of the other available information sources before examining PEATMOS, so that $F_1$ was made after considering a great deal of information. PEATMOS might have had a greater impact on the forecasts if $F_1$ had been made very early in the process of examining information, and PEATMOS had then been observed.

### d. Discussion

The results of the experiment indicate that the PEATMOS forecasts have had relatively little effect upon the forecasts formulated by the forecasters.[12] First, the results, in terms of scores, for the pre-PEATMOS forecasts and the post-PEATMOS forecasts were virtually identical, while the results for the PEATMOS forecasts were quite different. Second, the computations involving the ratio $T$ indicated that the shift in the forecasts (from $F_1$ to $F_3$) in response to PEATMOS was only about 20% of the distance from the pre-PEATMOS forecast to the PEATMOS forecast. However, this result may be partially due to the restriction, imposed by the experiment, that PEATMOS be examined last by the forecaster, after all of the other information sources had been observed and the pre-PEATMOS forecast had been made.

In further analyses of the Great Falls-Seattle data, we are conducting a more detailed analysis of the relationships among the pre-PEATMOS forecasts, the PEATMOS forecasts, and the post-PEATMOS forecasts and we are

[12] We should note, however, that precipitation in the locations of concern in this experiment tends to be strongly influenced by local (e.g., topographical) effects, and that the PEATMOS technique does not, as yet, satisfactorily take account of these effects. Thus, the implications of the results of this experiment are more likely to be applicable in areas of the country in which such effects are prominent.

investigating the effects of factors such as the individual forecaster, the forecast shift, the forecast length, and the location for which the forecast was prepared. A particular line of analysis that seems promising is to use Bayes' theorem to revise the pre-PEATMOS forecasts on the basis of the PEATMOS forecasts, using data relative to the performance of PEATMOS to obtain the likelihoods required for the formal application of Bayes' theorem.

The results of this experiment have implications with regard to the relative importance of guidance forecasts in the subjective precipitation probability forecasting process. When such forecasts are examined, they appear to have little impact upon the forecasters' precipitation probabilities and even less impact upon their performance, as measured by scoring rules. The results should also have implications for the intuitive revision of probabilities on the basis of additional information, although further analysis is required to fully assess these implications.

## 5. Summary and conclusions

In this paper we have discussed three experiments involving subjective probability forecasting in meteorology. These experiments were conducted in four National Weather Service Forecast Offices and the participants in the experiments were experienced weather forecasters. Although the results presented here do not represent a thorough, complete analysis of the data from the three experiments, we believe that even these preliminary results have important implications for probability forecasting in meteorology. In this regard, the St. Louis experiment indicates that the variables of concern in precipitation probability forecasting must be carefully defined and that a better understanding of the relationships among various probabilities (e.g., point and area probabilities of precipitation) may lead to improvements in the forecasters' ability to make probability forecasts. In addition, the results of this experiment suggest that forecasters, at present, may not be able to distinguish among different points in a forecast area (by assigning different probabilities to these points). However, similar experiments should be conducted in other locations, especially in locations in which local effects are important, to determine whether this tentative conclusion is warranted. The Denver experiment indicates that credible interval temperature forecasting is feasible and that the assessment procedures used in the experiment could be very useful in temperature forecasting in practice. Specifically, the results of the experiment reveal that the forecasters who used the variable-width, fixed-probability procedure were particularly successful in expressing the uncertainty inherent in their temperature forecasts in probabilistic terms. The Great Falls-Seattle experiment indicates that guidance (i.e., PEATMOS) forecasts may have little impact upon forecasters' precipitation probability forecasts, at least when these (guidance) forecasts are the last information source examined by the forecasters. In this regard, further analyses of the process by which weather forecasters aggregate information in formulating their probability forecasts should be very useful.

Further analyses of the results of these experiments can be expected to provide additional information relative to the practice of probability forecasting in meteorology and, more specifically, relative to the questions to which these experiments were addressed (e.g., can forecasters distinguish among different points in a forecast area with regard to the likelihood of precipitation, can forecasters make credible interval temperature forecasts). However, additional experiments (and other studies) are required to resolve these and many other questions of a practical nature related to probability forecasting. As indicated above, point and area precipitation probability forecasting experiments should be conducted in other locations, particularly in one or more locations that are subject to significant local effects (such experiments require that a dense network of raingages be available for evaluation purposes). In addition, experiments in credible interval temperature forecasting should be undertaken in several locations to investigate the feasibility of making such forecasts under a variety of meteorological conditions and to compare different credible interval forecasting procedures. Moreover, experiments should also be designed in which forecasters use credible interval procedures to forecast other continuous variables. Finally, we believe that the process by which weather forecasters aggregate information needs to be further investigated experimentally in operational settings, including reasonable constraints on the number of information sources considered and/or the order in which these sources are examined by the forecasters. As indicated in Section 1, the results of such experiments could provide a basis for reducing the amount of information which forecasters have to assimilate in formulating their forecasts.

In conclusion, the three experiments discussed here also have potentially important implications for subjective probability forecasting (or more broadly, for human behavior in inferential and decision-making situations) in general, including implications for the determination of inputs for formal models, the training and utilization of experts, the roles of humans and computers, and the gathering and summarizing of information. As indicated in Winkler and Murphy (1973b), the ultimate practical question with regard to studies of human behavior in inferential and decision-making situations is: How does a highly-motivated, experienced individual in an operational setting in his area of expertise, given appropriate feedback regarding past predictions and decisions and regarding the decision-making process itself, perform inferential and decision-making tasks, and can his performance be improved in any manner? The experiments discussed herein represent a modest step in the direction of studying certain aspects of this question in a meteorological context. Moreover, we feel that the forecasters' performances in all three of

these experiments could be improved upon and that further work in this regard would be most valuable.

*Acknowledgments.* We gratefully acknowledge the cooperation of the NWS forecasters from the WSFOs in Denver, Colo., Great Falls, Mont., St. Louis, Mo., and Seattle, Wash., who participated in the experiments described in this paper. We would also like to express our appreciation to the personnel in these WSFOs and in the Central and Western Region Headquarters of the NWS whose assistance facilitated the conduct of these experiments.

## References

Alpert, M., and H. Raiffa, 1969: A progress report on the training of probability assessors. Cambridge, Mass., Harvard University. Unpublished manuscript.

American Telephone and Telegraph Company, 1971: Weather announcement study. New York, N.Y., American Telephone and Telegraph Company, Market and Service Plans Department. Unpublished report.

Bickert, C. von E., 1967: A study of the understanding and use of probability of precipitation forecasts in two major cities. Denver, Colo., University of Denver, Denver Research Institute. Unpublished report.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

Epstein, E. S., 1966: Point and area precipitation probabilities. *Mon. Wea. Rev.*, **94**, 595–598.

Julian, P. R., and A. H. Murphy, 1972: Probability and statistics in meteorology: a review of some recent developments. *Bull. Amer. Meteor. Soc.*, **53**, 957–965.

Klein, W. H., and H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. *Bull. Amer. Meteor. Soc.*, **55**, 1217–1227.

Murphy, A. H., 1972: Probability forecasting in meteorology: a review of recent developments. Boulder, Colo., National Center for Atmospheric Research. Unpublished manuscript.

——, and R. L. Winkler, 1971: Forecasters and probability forecasts: the responses to a questionnaire. *Bull. Amer. Meteor. Soc.*, **52**, 158–165.

——, and ——, 1973: National Weather Service forecasters and probability forecasts: preliminary results of a nationwide survey. Boulder, Colo., National Center for Atmospheric Research, and Bloomington, Ind., Indiana University. Unpublished report.

——, and ——, 1974: Subjective probability forecasting in the real world: some experimental results. *Proc. Fourth Research Conference on Subjective Probability, Utility, and Decision Making* (Rome, 1973). In press.

Peterson, C. R., K. J. Snapper, and A. H. Murphy, 1972: Credible interval temperature forecasts. *Bull. Amer. Meteor. Soc.*, **53**, 966–970.

Stael von Holstein, C.-A. S., 1971: An experiment in probabilistic weather forecasting. *J. Appl. Meteor.*, **10**, 635–645.

——, 1972: Probabilistic forecasting: an experiment related to the stock market. *Organizational Behavior and Human Performance*, **8**, 139–158.

Winkler, R. L., and A. H. Murphy, 1968: "Good" probability assessors. *J. Appl. Meteor.*, **7**, 751–758.

——, and ——, 1973a: Information aggregation in probabilistic prediction. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC, **3**, 154–160.

——, and ——, 1973b: Experiments in the laboratory and the real world. *Organizational Behavior and Human Performance*, **10**, 252–270.