International Institute for
Applied Systems Analysis
IIASA www.iiasa.ac.at

Subjective Probability Forecasting in the Real World: Some Experimental Results

Murphy, A.H. and Winkler, R.L.

IIASA Research Report
December 1973

Murphy, A.H. and Winkler, R.L. (1973) Subjective Probability Forecasting in the Real World: Some Experimental Results. IIASA Research Report. Copyright © December 1973 by the author(s). http://pure.iiasa.ac.at/10/

All rights reserved. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage. All copies must bear this notice and the full citation on the first page. For other purposes, to republish, to post on servers or to redistribute to lists, permission must be sought by contacting repository@iiasa.ac.at

RR-73-16

SUBJECTIVE PROBABILITY FORECASTING IN THE REAL WORLD: SOME EXPERIMENTAL RESULTS

Allan H. Murphy

Robert L. Winkler

December 1973

Research Reports are publications reporting on the work of the author. Any views or conclusions are those of the author, and do not necessarily reflect those of IIASA.

Subjective Probability Forecasting in the Real World: Some Experimental Results*

Allan H. Murphy** and Robert L. Winkler***

Abstract

Three experiments in subjective probability forecasting were designed, and these experiments were conducted in four forecast offices of the U.S. National Weather Service. The first experiment involved credible interval temperature forecasts, the second experiment involved point and area precipitation probability forecasts, and the third experiment involved the effect of guidance forecasts on precipitation probability forecasts. In each case, some background material is presented; the design of the experiment discussed; some preliminary results of the experiment are presented; and some implications of the experiment and the results for probability forecasting in general and probability forecasting in meteorology in particular are discussed.

1. Introduction

Considerable interest exists among statisticians, decision theorists, psychologists, and others in human behavior in

---

*This paper will appear in the *Proceedings of the Fourth Research Conference on Subjective Probability, Utility, and Decision Making* (Rome, 1973). This research was supported in part by the U.S. National Science Foundation Grant GA-31735.

We gratefully acknowledge the cooperation of the NWS forecasters from the WSFOs in Denver, Colorado, Great Falls, Montana, St. Louis, Missouri, and Seattle, Washington, who participated in the experiments described in this paper. We would also like to express out 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.

**Advanced Study Program, National Center for Atmospheric Research, Boulder, Colorado, U.S.A.

***Graduate School of Business, Indiana University, U.S.A., research scholar at the International Institute for Applied Systems Analysis, Laxenburg, Austria.

inferential and decision-making situations. This interest is evidenced by the amount of research conducted in this area in the past decade. The research has consisted largely of experimental work, most of which has involved purposely simple, artificial, laboratory situations. Such simple situations are easy to deal with in terms of isolating certain factors of interest and they are easy to explain to typical subjects, who do not need to possess any particular expertise. However, their very simplicity and artificiality makes the justification for generalizing the results of these experiments to more realistic inferential and decision-making situations questionable. In Winkler and Murphy [10] we discuss some of the difficulties inherent in experimentation in realistic settings, suggest possible procedures for avoiding or at least alleviating such difficulties, and make a plea for more realistic experiments.

Meteorology is apparently still the only field in which probability forecasts are issued to the public on a regular, operational basis, and the forecasters of the U.S. National Weather Service (NWS) represent a large group of experts who make probability forecasts daily. (For a recent review of probability forecasting in meteorology, refer to Murphy [4] or Julian and Murphy [3].) Therefore, meteorology is an ideal setting for realistic experiments, and a large pool of potential subjects is available. We designed three experiments in subjective probability forecasting and conducted these experiments in four Weather Service Forecast Offices (WSFOs) of the NWS. The first experiment involved credible interval temperature forecasts, the second experiment involved point and area precipitation probability forecasts, and the third experiment involved the effect of guidance forecasts on precipitation probability forecasts.

The three experiments are discussed in Sections 2, 3, and 4 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 general and probability forecasting in meteorology in particular are discussed.[1] Section 5 contains a brief summary.

2. An Experiment Regarding Credible Interval Temperature Forecasts

a) Credible Interval Temperature Forecasts

In probability forecasting in meteorology, forecasts of precipitation occurrence have received the greatest attention. The use of probabilities in forecasting precipitation occurrence has been an operational procedure in the NWS since 1965. Of course, precipitation occurrence lends itself quite well to the use of probabilities, since this variable is a simple dichotomy. As a result, only a single probability is needed to express a forecaster's uncertainty about the occurrence of precipitation.

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. However, a probability is <u>not</u> assessed for the interval, so that users of the forecasts are unable to "measure" the uncertainty inherent in any particular interval forecast.

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 [7] 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, 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 (all temperatures in this paper are expressed in $^{\circ}$F).

An obvious alternative to variable-width forecasts is a forecast for which the width of the interval is fixed but which allows the forecaster to vary the probability associated with the interval. For instance, the forecaster might be asked to report a credible interval that is exactly 5^o 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.

Peterson, Snapper, and Murphy [7, 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, Colorado. 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 thirty or more sets of forecasts from each participant. The data analyzed here 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 provides 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 and 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. All intervals in the experiment were assumed to include their end points, and all temperatures were expressed to the nearest degree (e.g. the 5° interval from 48° to 52° includes all of the temperatures from 47.5° to 52.5°).

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 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 temperatures. Of course, we would not expect the medians and the official forecasts to differ a great deal, since both were 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 χ^2 (1.333, with 2 d.f.) 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 χ^2 (1.646, with 2 d.f.). 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 1 do not indicate a strong relationship, although a positive relationship seems to hold for the range of widths for which a reasonable number of cases exists. The average widths were 6.2° (standard error = 0.11°) and 11.7° (standard error = 0.19°) for the 50% and 75% credible intervals, respectively.

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)

Table 1. Average error as a function of interval width.

50% Credible Intervals

Width (°F)	Number of Forecasts n	Average Error (°F)
3	2	3.00
4	9	2.56
5	22	3.09
6	44	4.55
7	42	3.98
8	6	2.83
9	6	6.00
10	0	----
11	1	11.00
Total/ Average	132	4.00

75% Credible Intervals

Width (°F)	Number of Forecasts n	Average Error (°F)
6	1	1.00
7	2	2.50
8	7	3.00
9	11	3.64
10	12	2.75
11	29	3.83
12	25	3.92
13	29	4.86
14	6	3.83
15	4	2.00
16	3	10.33
17	0	----
18	2	2.50
19	0	----
20	0	----
21	1	11.00
Total/ Average	132	4.00

(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 were 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.[2] In Table 2 the relative frequency of inclusion of the observed temperature 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° line for which the observed relative frequency for each probability exactly equals that probability.

Table 2. Average error and relative frequency of inclusion of observed temperature in interval as a function of probability of interval.

5°F Intervals

Probability of Interval	Number of Forecasts n	Average Error ($^{\circ}$F)	Relative Frequency in Interval
0.30	2	3.50	0.00
0.35	1	8.00	0.00
0.40	22	4.82	0.23
0.50	22	3.86	0.46
0.60	31	3.94	0.35
0.70	24	3.25	0.50
0.75	3	2.33	0.67
0.80	8	1.12	1.00
0.90	7	2.14	0.86
1.00	2	1.00	1.00
Total/Average	122	3.60	0.46

9°F Intervals

Probability of Interval	Number of Forecasts n	Average Error ($^{\circ}$F)	Relative Frequency in Interval
0.50	2	8.00	0.00
0.60	6	4.17	0.50
0.70	29	3.97	0.62
0.75	5	3.60	0.60
0.80	39	4.18	0.62
0.85	4	3.25	0.75
0.90	20	2.70	0.75
0.95	4	3.25	0.50
1.00	13	1.69	0.92
Total/Average	122	3.60	0.66

The average error was expected to be a decreasing function of the probabilities assigned to the 5^{o} and 9^{o} central credible intervals. Although the amount of data is limited for some probabilities, Table 2 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 the 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 differences between forecasts of high and low temperature, among forecasts formulated by different forecasters, and between forecasts prepared on day and midnight shifts. 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) and to forecasts of other variables of interest in other fields (e.g. economic indicators).

3. An Experiment Regarding Point and Area Precipitation Probability Forecasts

a) Point and Area Precipitation Forecasts

Precipitation probability forecasts are 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 a precipitation probability issued to the public is an average point probability of measurable precipitation for an entire forecast area (generally 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. On the other hand, the observation of precipitation is taken at only one point (the official rain gauge). Occasionally, when the probability of precipitation varies considerably over the forecast area, forecasters may issue diffferent forecasts for different parts of the area. When such variations exist, the use of an average point probability for the entire area would, in general, 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, [6]), 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.[3]

The relationship between point and area precipitation probabilities has been studied theoretically (e.g. Epstein, [2]) but not empirically. The experiment reported in this section was designed to investigate the relative ability of forecasters to make point and area (including areal coverage) probability forecasts and the ability of forecasters to differentiate between 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 fourteen experienced weather forecasters from the WSFO at St. Louis, Missouri. 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 (rain gauges) 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 twelve-hour periods in the future (e.g. today, tonight, tomorrow). The experiment was conducted from November 1972 to March 1973.

Observations from the Illinois State Water Survey network of rain gauges in the St. Louis area were used to verify the forecasts. This network included rain gauges at the five points for which point probabilities of precipitation were determined by the forecasters. A larger set of twenty rain gauges was chosen to verify the forecasts of area probability and expected areal coverage. Within the constraints imposed by the location of available rain gauges, the smaller set of five points and the larger set of twenty points were chosen to provide a reasonable coverage of the St. Louis metropolitan area.

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 a magnitude of at least 0.10.[4] The largest sample variance for a set of point probabilities in the entire experiment was 0.068, which yields a sample standard deviation of 0.26.

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 twenty rain gauges rather than just five rain gauges, this probability (denoted by A) 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 cases (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.

Another comparison of interest is that of the 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 rain gauges, p_i is the probability of precipitation at rain gauge i, and δ_i is an indicator variable that equals one if precipitation occurs at rain gauge i and zero otherwise. From the forecasts, 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 = 0.0001), 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 larger than any point probability, since precipitation at any point implies precipitation in the 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 a function of the location (i.e. St. Louis) and/or the particular weather situations which occurred during the period. On the other hand, perhaps the variability should have been greater, but the forecasters were simply unable to differentiate among the points more often (or as often as they should). 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. 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 "lead time" of the forecast, and the forecast shift (i.e. day, midnight). For example, 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 (this analysis was delayed because the recorded observations were not immediately available). 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. In this regard, the differences among the relative frequencies of precipitation at the five points for which point probability forecasts were made are of particular interest.

These experimental results 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.

4. An Experiment Regarding the Effect 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 [5]) indicated that the relative importance and the order of examination of information sources vary among forecasters and among weather situations, and a more recent and more extensive questionnaire (Murphy and Winkler [6]) that we administered to NWS forecasters has 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 [9]). Ideally, controls concerning all information sources would be useful, but this ideal situation is very difficult to attain in an operational setting.

Guidance forecasts prepared by the NWS using a procedure called PEATMOS represent an information source of particular interest because the guidance forecasts themselves are expressed in probabilistic terms. PEATMOS, which stands for Primitive Equation and Trajectory Model Output Statistics, 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 questionnaires mentioned above have provided some information relative to the importance and the order of examination of different information sources by weather forecasters in the process of arriving at precipitation probabilities, no experimental investigations of this process have been conducted. The experiment reported in this section was designed to investigate the effect of the guidance (PEATMOS) forecasts on 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, Montana, and six experienced weather forecasters from the WSFO at

Seattle, Washington. 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 the post-PEATMOS forecasts were made.

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 (e.g. today, tonight, tomorrrow). 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 relationships between the probability and the observed relative frequency of precipitation for the three types of forecasts are presented in Table 3. While 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° line for which the observed relative frequency over the entire sample for each forecast probability exactly equals that probability. At Great Falls, the situation was reversed. Overall, the probabilities and the observed relative frequencies are quite close in many cases, but considerable room for improvement exists in other cases.

One clear result that does emerge from Table 3 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 4. The scoring rules used were the quadratic rule (Q) and logarithmic rule (L):

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

Table 3. Relative frequency of precipitation as a function of precipitation probability forecast. (Number of forecasts in parentheses)

	Great Falls			Seattle		
Probability Forecast	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)

Table 4. Average scores for pre-PEATMOS (F_1), PEATMOS (F_2), and post PEATMOS (F_3) forecasts.

	Great Falls		Seattle	
Type of Forecasts	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

and

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

($i = 1, 2, 3$). In each case, a higher score indicates better performance. Note that at both Great Falls and Seattle, the differences between the average scores for F_1 and F_3 were quite small. Note also that because Seattle and Great Falls experience different weather situations, the scores for Seattle and Great Falls are not comparable (that is, the results do not necessarily imply, for example, that the forecasters at Great Falls were "better" than those at Seattle).

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):

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

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 (the standard errors were 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 were made very early in the process of examining information, and PEATMOS was then observed.

d) Discussion

The results of the experiment indicate that the forecasters did not shift their probabilities much in response to PEATMOS. First, the results, in terms of scores, for the pre-PEATMOS forecasts and the post-PEATMOS forecasts were virtually identical, while the results for PEATMOS were quite different. Second, the computations involving the ratio T indicated that the shift in the forecasts (from F_1 to F_3) 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 after all 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, within the limits imposed by the sample size, we are investigating the effects of factors such as the individual forecaster, the lead time of the forecast, 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 PEATMOS, using data relative to the performance of PEATMOS to obtain likelihoods for the formal application of Bayes' theorem.[5]

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 last, they appear to have little impact upon the forecasters' precipitation probabilities and even less impact upon their performance, as measured by scoring rules.[6] The results should also have implications for the intuitive revision of probabilities on the basis of additional information, although further analysis is needed to fully investigate these implications.

5. Summary

In this paper we have discussed three experiments involving subjective probability forecasting in meteorology. The three experiments were conducted in operational settings and the participants were experienced weather forecasters. Thus, the experiments were more realistic than most experiments that have been conducted in the area of subjective probability forecasting. Even though the results presented here do not represent a thorough, complete analysis of the data from the three experiments, these results have obvious implications for probability forecasting in meteorology. The Denver experiment indicates that credible interval temperature forecasting is feasible, and that the procedures used in the experiment could be very useful in temperature forecasting in practice. The St. Louis experiment indicates that the variables of concern in probability forecasting in meteorology must be carefully defined and that a better understanding of the relationships among various probabilities (e.g. point and area probabilities of precipitation) may improve the forecaster's ability to determine probability forecasts. The Great Falls-Seattle experiment indicates that a guidance forecast may have little impact of the forecaster's precipitation probability when this (guidance) forecast is the last information source examined (however, see Footnote 5) and that an analysis of the process by which weather forecasters aggregate information to arrive at a probability forecast should be very useful.

In addition to their obvious implications for probability forecasting in meteorology, the three experiments discussed here have potentially important implications for subjective probability forecasting (or more broadly, for human behavior in inferential and decision-making situations) in general. Experiments concerning human behavior in realistic inferential and decision-making situations have important implications for the the determination of inputs for formal models, the training and utilization of experts, the roles of humans and computers, the gathering and summarizing of information, and many other important questions. 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 upon in any manner? The experiments discussed herein represent a modest step in the direction of studying certain aspects of this question. Moreover, we feel that the forecasters' performances in all three of these experiments could be improved and that further work in this regard would be most valuable.

Footnotes

[1] Space prohibits a thorough discussion or a complete analysis of the experiments described in this paper, and in forthcoming papers we intend to discuss each of the experiments in much greater detail.

[2] While these intervals are too "tight," they are not as tight as the distributions obtained in many other experiments involving probability assessment (e.g. Alpert and Raiffa, [1]; Stael von Holstein, [8]). We attribute the forecasters' performance in this experiment to the degree of their (substantive) expertise; for the most part the participants in other experiments were not experts in the areas of concern or the uncertain quantities of interest were of the almanac type.

[3] 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.

[4] 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.

[5] In using Bayes' theorem in this analysis, we would like to work with data regarding PEATMOS that is conditional upon the other information already examined by the forecaster. However, the available data concerning the performance of PEATMOS are unconditional in nature, so that in using these data we are implicitly assuming that PEATMOS is conditionally independent of the other information.

[6] We should note, however, that precipitation in the locations of concern in this experiment is 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.

References

[1] Alpert, M., and H. Raiffa, "A progress report on the training of probability assessors." Cambridge, Mass., Harvard University, unpublished manuscript, 1969.

[2] Epstein, E.S., "Point and area precipitation probabilities," Monthly Weather Review, 94 (1966),595-598.

[3] Julian, P.R., and A.H. Murphy, "Probability and statistics in meteorology: a review of some recent developments," Bulletin of the American Meteorological Society, 53 (1972),957-965.

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

[5] Murphy, A.H., and R.L. Winkler, "Forecasters and probability forecasts: the responses to a questionnaire, "Bulletin of the American Meteorological Society, 52 (1971), 158-165.

[6] Murphy, A.H. and R.L. Winkler, "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 manuscript, 1973.

[7] Peterson, C.R., K.J. Snapper, and A.H. Murphy, "Credible interval temperature forecasts," Bulletin of the American Meteorological Society, 53 (1972),966-970.

[8] Stael von Holstein, C.-A.S., "Probabilistic forecasting: an experiment related to the stock market," Organizational Behavior and Human Performance, 8 (1972),139-158

[9] Winkler, R.L., and A.H. Murphy, "Information aggregation in probabilistic prediction," IEEE Transactions on Systems, Man, and Cybernetics, SMC-3 (1973),154-160.

[10] Winkler, R.L., and A.H. Murphy, "Experiments in the laboratory and the real world," Organizational Behavior and Human Performance, 10 (1973), 252-270.