RR-74-12

# CREDIBLE INTERVAL TEMPERATURE FORECASTING:

# SOME EXPERIMENTAL RESULTS

Allan H. Murphy
Robert L. Winkler

August 1974

# Credible Interval Temperature Forecasting: Some Experimental Results*

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

## Abstract

This paper describes the results of an experiment involving credible interval temperature forecasts. A credible interval is an interval of values of the variable of concern, in this case maximum or minimum temperature, accompanied by a probability which expresses a forecaster's "degree of belief" that the temperature will fall in the given interval. The experiment was designed to investigate the ability of forecasters to express the uncertainty inherent in their temperature forecasts in probabilistic terms and to compare two approaches (variable-width and fixed-width intervals) to credible interval temperature forecasting.

Four experienced weather forecasters participated in the experiment, which was conducted at the National Weather Service Forecast Office in Denver, Colorado. Two forecasters made variable-width, fixed-probability forecasts using 50% and 75% intervals, while the other two

---

*Supported in part by the U.S. National Science Foundation under grants GA-31735 and GA-41232.

We gratefully acknowledge the cooperation of the NWS forecasters at the Denver WSFO who participated in this experiment: Messrs. Henry W. Chidley, Jack A. Frost, John A. Schwab, and Kenneth C. Tillotson. We would also like to express our appreciation to Messrs. Marshall F. Grace and Norman E. Prosser, Meteorologist in Charge and Principal Assistant, respectively, at the Denver WSFO and to Mr. Lawrence A. Hughes, Regional Meteorologist at the Central Region Headquarters in Kansas City, Missouri, whose assistance greatly facilitated the conduct of this experiment. Finally, we would like to thank Ms. Vasu Deshpande and Mr. Lalgudi Ramnarayan for their computational assistance.

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

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

forecasters made fixed-width, variable-probability forecasts using 5°F and 9°F intervals. On each occasion the forecasters first determined a median, and the variable-width and fixed-width intervals were then centered at the median in terms of probability and width, respectively.

The results indicate that, overall, the medians determined by the forecasters were good point forecasts of maximum and minimum temperatures. Further, a comparison of the average errors for the forecasters' medians with the average errors for the medians derived from climatology reveals that the forecasters were able to improve greatly upon climatology. The variable-width credible intervals were very reliable in the sense that the observed relative frequencies corresponded very closely to the forecast probabilities. Moreover, the variable-width intervals were more reliable and much more precise than the corresponding forecasts derived from climatology. The fixed-width intervals, on the other hand, were assigned probabilities that were, on the average, considerably larger than the corresponding relative frequencies.

In summary, the results indicate that weather forecasters can use credible intervals to describe the uncertainty contained in their temperature forecasts. The implications of these experimental results for probability forecasting in general and temperature forecasting in particular are discussed.

## 1. Introduction

Probability forecasts in meteorology serve two basic purposes: 1) they provide forecasters with a means of expressing the uncertainty inherent in their forecasts and 2) they provide potential users of such forecasts with information needed to make rational decisions in uncertain situations. For these reasons, the National Weather Service (NWS), in 1965, initiated a nationwide program in which probability of precipitation (POP) forecasts were formulated and issued to the general public. This program has now been in existence for almost a decade, and the evidence presently available suggests that the POP forecasts are considered, by both fore-

casters and the general public, to be an important and integral part of the NWS's public weather forecasts (e.g. American Telephone and Telegraph Company, [2], Bickert, [3], and Murphy and Winkler, [6]).

Precipitation occurrence has received the greatest attention in terms of probability forecasting. On the other hand, subjective probability forecasts of other meteorological variables have been prepared on an experimental basis (see, for example, Sanders, [8], Stael von Holstein, [10]). Moreover, probability forecasts of a variety of meteorological variables are currently being prepared on an operational basis using the model output statistics approach and these "objective" forecasts are routinely provided to NWS forecasters as guidance (see Klein and Glahn, [5]). However, the forecasts of these variables disseminated to the general public are still expressed in categorical terms, a situation which is due in part to the lack of suitable modes of expression for the uncertainty contained in these forecasts. In this regard, the ranges of continuous variables such as temperature have generally been divided into several (often five or more) categories. As a result, a single forecast consists of several probabilities, one for each category. Clearly, this mode of expression makes effective communication of the uncertainty inherent in forecasts of such variables very difficult.

One possible (and promising) format for probability forecasts of continuous variables such as temperature involves

the concept of a credible interval, which is an interval of potential values of the variable together with a probability that the actual value of the variable will fall in the interval. Peterson, Snapper, and Murphy [7, p.969] recently conducted an experiment to investigate the feasibility of credible interval temperature forecasting and concluded that "weather forecasters can use credible intervals to describe the uncertainty inherent in their temperature forecasts."

In this paper the results of an experiment involving credible interval temperature forecasts are presented. The experiment was designed to investigate further the ability of forecasters to express the uncertainty in their temperature forecasts in probabilistic terms and to compare two approaches (variable-width credible intervals and fixed-width credible intervals) to credible interval temperature forecasting. In Section 2 the concept of credible interval forecasts is defined and discussed. The experiment itself is described in Section 3, and the results of the experiment are presented in Section 4. Section 5 contains a discussion of some implications of the experimental results for temperature forecasting in particular and for probability forecasting in general.

## 2. Credible Interval Temperature Forecasts

As indicated in the Section 1, uncertainty exists in the forecasts of all the variables presently included in public weather forecasts. Yet these forecasts, with the exception of those relating to precipitation occurrence, are still ex-

pressed in categorical terms. Precipitation occurrence, of course, 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. On the other hand, a continuous variable such as temperature requires either a forecast consisting of several probabilities (see Section 1) or a completely different type of probability forecast. Ideally, an entire probability distribution would be assessed, but assessing such a distribution is not practical either in terms of the time required of the forecaster or in terms of reporting to the general public. Credible intervals represent a mode of expression that provides some probabilistic information without necessitating the assessment of an entire distribution. We will introduce this discussion of the concept of credible intervals and their use in forecasting maximum (high) and minimum (low) temperatures by first considering the mode of expression currently used to describe temperature forecasts operationally.

Weather forecasters usually give point forecasts when forecasting high and low temperatures. The point forecast may, on occasion, be replaced by an interval forecast that specifies a range of temperatures, but the usefulness of the interval forecast is severely limited by the fact that the probability that the forecaster associates with the interval is not given. For example, an interval forecast such as "the high tomorrow will be between $70^{o}$ and $76^{o}$" (all tempera-

tures referred to in this paper are in $^{o}$F) may mean different things on different occasions. If meteorological conditions are relatively stable, the forecaster may feel almost certain that the maximum temperature will be between 70$^{o}$ and 76$^{o}$. On the other hand, if conditions are highly variable, the forecaster may feel that the probability is only, say, about one-half that the maximum temperature will fall in this interval.

Of course, the forecaster can vary the width of the interval forecast. On some occasions, an interval forecast such as "the high temperature will be between 70$^{o}$ and 76$^{o}$" may seem reasonable to the forecaster, while on other occasions a forecast such as "the high temperature will be between 72$^{o}$ and 74$^{o}$" may seem more appropriate. The user would no doubt feel that the former forecast suggests more uncertainty about the high temperature tomorrow than does the latter forecast, and in this sense such interval forecasts may be of some value. Nevertheless, although users may attempt to make inferences concerning the relative uncertainty expressed by different interval forecasts, they cannot, from the information given above, "measure" the uncertainty expressed by a particular interval forecast.

Probability can be thought of as the language of uncertainty. Therefore, in order to convey the amount of uncertainty in an interval forecast, the forecaster must report a probability together with the interval. This probability represents the forecaster's subjective "degree of belief" that the high or low temperature, as the case may be, will

fall in the given interval. When accompanied by a probability, an interval forecast is called a credible interval. For example, a forecaster might say that "the probability is 0.50 that the high tomorrow will be between $72^{\circ}$ and $74^{\circ}$." Peterson, Snapper, and Murphy [7, p.966] state that "the advantage of credible interval forecasts is, then, that they enable forecasters to quantify the uncertainty inherent in their temperature forecasts and to communicate information which may be important to potential users of these forecasts."

Just as a precipitation probability serves as a measure of a forecaster's uncertainty concerning the occurrence of precipitation, a credible interval serves as a measure of a forecaster's uncertainty concerning maximum or minimum temperature. The two situations differ, however, in that a single precipitation probability completely describes a forecaster's uncertainty, whereas, at least in theory, an infinite number of potential credible intervals for high or low temperature exist, each of which only partially describes a forecaster's uncertainty. To completely represent a forecaster's uncertainty concerning high or low temperature, an entire probability distribution is needed. Unfortunately, an entire distribution is not only difficult to assess (e.g. see Winkler, [11]), but such a distribution is inconvenient for reporting purposes, both because it cannot, in general, be expressed in a short, simple, nontechnical fashion and because most users would be unable to understand or to properly utilize such a forecast. In this regard, a credible interval can be thought

of as a summary measure of a probability distribution.

Given that a credible interval is to be used in forecasting high or low temperature, the next question concerns the selection of a particular interval. In order to make such forecasts at least somewhat comparable (as well as to increase their usefulness), certain restrictions must be placed upon the interval, instead of giving the forecaster complete freedom in the selection of a credible interval on each occasion. One possible restriction that seems reasonable is to limit the forecaster to reporting central credible intervals, which are intervals taken from the "center" of the forecaster's distribution in terms of probability. For instance, the interval from $72^{\circ}$ to $74^{\circ}$ is a 50% central credible interval if the probability that the high will be between $72^{\circ}$ and $74^{\circ}$ is 0.50, the probability that the high will be above $74^{\circ}$ is 0.25, and the probability that the high will be below $72^{\circ}$ is 0.25. Another interval with probability 0.50 may be found such that the probability that the high will be above the upper limit of the interval is, say, 0.30 and the probability that the high will be below the lower limit of the interval is 0.20. Such an interval would be a 50% credible interval but not a 50% central credible interval.

Restricting the forecaster to credible intervals that are central in terms of probability seems particularly reasonable when a restriction on the probability of the interval is added. For instance, the forecaster might be asked to

always report a 50% central credible interval or a 75% central credible interval. In this case, the probability of the interval is fixed but the width of the interval will vary from situation to situation. Sometimes a 50% credible interval for high or low temperature will be only $3^{o}$ wide, while at other times such an interval may be $7^{o}$ wide. For obvious reasons, we will call a forecast of this nature a variable-width credible interval.

An obvious alternative to variable-width forecasts is a restriction that fixes the width of the interval but 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 cases the probability of such an interval might be 0.50, whereas in other cases it might be 0.90. A forecast of this nature will be called a fixed-width credible interval. In the experiment described in this paper, both variable-width credible intervals and fixed-width credible intervals were considered. In the Peterson, Snapper, and Murphy [7] experiment, only variable-width credible intervals were investigated.

## 3. Design of the Experiment

The subjects in the experiment were four experienced weather forecasters from the NWS's Weather Service Forecast Office (WSFO) at Stapleton International Airport, Denver, Colorado. The forecasters, all of whom possessed Bachelor's

degrees in meteorology, ranged in age from 47 to 57 years. They averaged 26 years of weather forecasting experience (range: 18 to 31), 17.5 years of experience at the Denver WSFO (range: 12 to 27), and 5.75 years of probability forecasting experience (range: 4 to 8).

Each time they were on public weather forecasting duty during the period of the experiment, the forecasters made credible interval forecasts of high and low temperatures for Denver. 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, approximately five months were required to obtain at least 30 sets of forecasts from each participant. The forecasts analyzed in this paper were collected over a period from August 1972 to March 1973, and the four participants formulated 32, 34, 30, and 31 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. To obtain these intervals, each forecaster was asked to make a total of five "indifference judgments" at equal odds. The first indifference judgment determines the median of the forecaster's probability distribution (i.e., the temperature that the forecaster feels is equally likely to be exceeded or not exceeded by the actual

high or low temperature). The second indifference judgement determines the 25th percentile of the forecaster's distribution by asking the forecaster to specify a temperature value that divides the interval below the median into two equally likely subintervals, just as the median divided the entire range of temperatures into two equally likely intervals. The third indifference judgment determines the 12-1/2th percentile of the forecaster's distribution by asking the forecaster to specify a value that divides the interval below the 25th percentile into two equally likely subintervals. The fourth and fifth indifference judgments are analogous to the second and third indifference judgments, except that they are concerned with the interval above the median and with specifying the 75th and 87-1/2th percentiles rather than the 25th and 12-1/2th percentiles. For a more detailed explanation of the indifference judgments involved in formulating the variable-width interval forecasts, see the Appendix.

Once the five indifference judgments are made, the 50% central credible interval is the interval from the 25th percentile to the 75th percentile, and the 75% central credible interval is the interval from the 12-1/2th percentile to the 87-1/2th percentile. Thus, the 50% and 75% central credible intervals are convenient to determine in the sense that they require only five simple, equal-odds indifference judgments, whereas, for example, a 95% credible interval would require many additional equal-odds indifference judgments or one or more indifference judgments involving unequal odds. As noted

by Peterson, Snapper, and Murphy [7], unequal-odds judgments are more difficult to make than equal-odds judgments, particularly for forecasters who are inexperienced in making credible interval forecasts. Alternatively, of course, the forecaster could simply be asked directly to give a 50% central credible interval or any other credible interval. The indirect procedure used here provides the forecaster with a systematic procedure for determining credible intervals. If this procedure is followed, then the forecaster can examine the resulting intervals and determine whether they seem reasonable. In the experiment, the forecasters were asked to check their responses as follows:

> Looking at your responses, do you feel that it is equally likely that the maximum temperature will be in the interval from your 25th to 75th percentiles or outside this interval? Also, do you feel that it is three times as likely that the maximum temperature will be in the interval from your 12-1/2th to 87-1/2th percentiles as that it will be outside this interval? If not, you should reconsider your responses and make any changes that seem necessary.

As noted above, only two of the four forecasters made variable-width interval forecasts. The other two forecasters worked within the framework of fixed-width, variable-probability forecasts, using intervals of width $5^{\circ}$ and $9^{\circ}$. 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 determine probabilities for intervals of width $5^{\circ}$ and $9^{\circ}$ centered at the median. All inter-

vals were assumed to include their end points, and all temperatures were recorded to the nearest degree (e.g. the interval from $63^{\circ}$ to $67^{\circ}$ is of width $5^{\circ}$, since it includes all temperatures from $62.5^{\circ}$ to $67.5^{\circ}$). Thus, if the median is $70^{\circ}$, for example, the two intervals of concern in the fixed-width situation would be the interval from $68^{\circ}$ to $72^{\circ}$ and the interval from $66^{\circ}$ to $74^{\circ}$.

Although the fixed-width intervals are symmetric about the median in terms of width, they will not always be symmetric about the median in terms of probability. For example, when the forecaster's probability distribution is asymmetric, fixed-width intervals will not, in general, be symmetric about the median in terms of probability. In such cases, the intervals will not be central credible intervals, and the degree to which these intervals deviate from central credible intervals will be a function of the degree of asymmetry of the forecaster's distribution. Of course, the forecaster could be asked directly for a central credible interval of width $5^{\circ}$ or $9^{\circ}$ instead of using the more indirect approach that involves centering all of the intervals at the median of the forecaster's distribution. Once the median is determined, however, a fixed-width credible interval centered at the median only requires the forecaster to report a probability for that interval. If the interval is not centered at the median, then the forecaster must report both a probability and at least one end point of the interval.

Prior to the start of the experiment, the authors met with the forecasters from the Denver WSFO (including some forecasters who did not take part in the experiment) 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. Since the instruction sets were quite important in this experiment and since a brief description fails to capture the essence of such instruction sets, the instructions for the variable-width approach are presented in the Appendix of this paper. No difficulties arose after the instruction sets were distributed, and we believe that the participants understood the experimental procedures.

## 4. Results of the Experiment

### a) Medians

Whether they were concerned with variable-width or fixed-width intervals, the first task on each forecasting occasion

for all of the participants in the experiment was to determine a median. A comparison of these median temperatures (MTs) with the corresponding observed temperatures (OTs) is presented in Table 1. For the entire sample (n=254), MT equalled OT 12.6% of the time, MT was greater than OT 39.4% of the time, and MT was less than OT 48.0% of the time. Thus, a slight tendency existed for the MTs to underestimate the OTs. A careful examination of Table 1 reveals that this result is due largely to Forecaster 4, who underestimated more than twice as often (59.7% to 25.8%) as he overestimated. The other three forecasters exhibited little systematic bias of this nature, for their frequency of underestimation approximately equalled their frequency of overestimation. Forecaster 4's tendency to underestimate also explains the differences (in terms of underestimation versus overestimation) between variable-width forecasts (Forecasters 1 and 2) and fixed-width forecasts (Forecasters 3 and 4). Since all of the participants formulated forecasts of both maximum and minimum temperature, however, a tendency to underestimate minimum temperatures (56.7% underestimates, 31.5% overestimates) and a lesser tendency to overestimate maximum temperatures (47.2% overestimates, 39.4% underestimates) cannot be explained in terms of any individual forecaster.

The discussion in the preceding paragraph is further supported by the average difference between MT and OT (see Table 1). Forecasters 1, 2, and 3 had average differences ranging from $-0.3^{\circ}$ to $0.0^{\circ}$, whereas for Forecaster 4 the

Table 1. A comparison of the median temperature (MT), the observed temperature (OT), and the forecast temperature (FT).

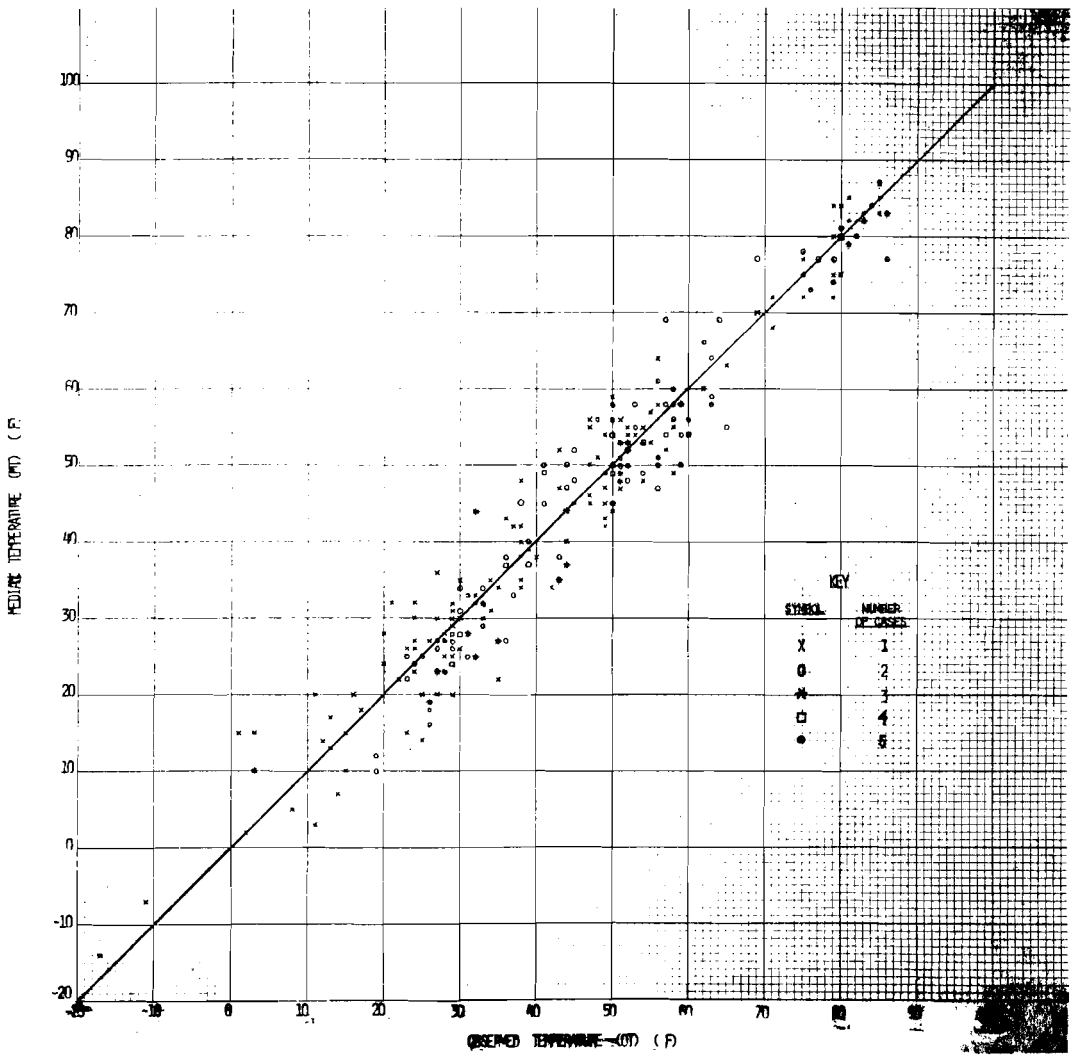
| Set of Forecasts | Number of Forecasts | Percentages | | | Average (Standard Deviation) | | | |
|---|---|---|---|---|---|---|---|---|
| | | MT > OT | MT = OT | MT < OT | MT-OT (°F) | \|MT-OT\| (°F) | $(MT-OT)^2$ $(°F)^2$ | \|FT-OT\| (°F) |
| All | 254 | 39.4 | 12.6 | 48.0 | -0.5 (4.9) | 3.8 (3.1) | 24.0 (33.0) | 3.9 (3.1) |
| Variable-Width | 132 | 44.7 | 12.1 | 43.2 | -0.1 (5.2) | 4.0 (3.3) | 26.5 (36.4) | 4.0 (3.2) |
| Fixed-Width | 122 | 33.6 | 13.1 | 53.3 | -0.8 (4.6) | 3.6 (2.9) | 21.4 (28.9) | 3.8 (3.0) |
| Maximum | 127 | 47.2 | 13.4 | 39.4 | 0.6 (4.8) | 3.8 (3.0) | 23.5 (30.4) | 3.9 (3.0) |
| Minimum | 127 | 31.5 | 11.8 | 56.7 | -1.5 (4.8) | 3.8 (3.2) | 24.6 (35.6) | 3.9 (3.2) |
| 12-Hour | 127 | 40.9 | 13.4 | 45.7 | -0.2 (4.7) | 3.5 (3.0) | 21.4 (31.6) | 3.7 (3.0) |
| 24-Hour | 127 | 37.8 | 11.8 | 50.4 | -0.7 (5.1) | 4.1 (3.2) | 26.7 (34.3) | 4.2 (3.2) |
| Forecaster 1 | 64 | 45.3 | 10.9 | 43.8 | 0.0 (5.3) | 4.1 (3.3) | 27.3 (40.2) | 4.1 (3.2) |
| Forecaster 2 | 68 | 44.1 | 13.2 | 42.6 | -0.3 (5.1) | 3.9 (3.3) | 25.8 (32.7) | 4.0 (3.2) |
| Forecaster 3 | 60 | 41.7 | 11.7 | 46.7 | -0.1 (4.6) | 3.6 (2.9) | 21.2 (28.7) | 3.7 (2.9) |
| Forecaster 4 | 62 | 25.8 | 14.5 | 59.7 | -1.5 (4.4) | 3.6 (3.0) | 21.6 (29.3) | 3.8 (3.1) |

average difference was $-1.5^{\circ}$. Once again, Forecaster 4's forecasts explain both the overall tendency for MT to underestimate OT and the difference between variable-width and fixed-width forecasts. Also, the above comments comparing forecasts of maximum and minimum temperatures are supported by the average differences between the median and observed temperatures.

Of course, even if the average difference is close to zero, the actual differences may tend to be quite large in both directions. However, most of the points in Figure 1, a scatter diagram MT versus OT, are close to the diagonal $45^{\circ}$ line for which MT equals OT. Furthermore, the average absolute difference between MT and OT was $3.8^{\circ}$ (standard error = $0.2^{\circ}$) and the averaged squared difference was $24.0(^{\circ})^2$ [standard error = $2.1(^{\circ})^2$]. These results are remarkably consistent across forecasters and different types of forecasts (see Table 1), although some differences do exist [e.g. as expected, the average values of $|MT-OT|$ and $(MT-OT)^2$ were slightly smaller for the 12-hour forecasts than for the 24-hour forecasts]. Scatter diagrams (not presented here) suggest that the average absolute error was not a function of the observed temperature. In general, then, the medians seem to be good point forecasts.

For comparative purposes, the official temperature forecast (FT) issued to the public was recorded on each occasion. The average difference between FT and OT was $-0.2^{\circ}$, and the corresponding average absolute difference was $3.9^{\circ}$ (see Table 1).

Figure 1. A scatter diagram of median temperature (MT) versus observed temperature (OT) for entire sample of forecasts (n = 254).



Thus, the medians determined by the forecasters for the purposes of the experiment were, on the average, comparable to the official forecasts as point forecasts of maximum and minimum temperatures. Of course, we would not expect the medians and official forecasts to differ a great deal, since both were determined by the same forecaster on almost all occasions.

The climatological median temperature (CT) provides a convenient standard with which to compare MT as a point forecast.[1] The climatological data for the Denver WSFO for the five years preceding the experiment were used to define CT. Thus, for example, for the 150 September days in the five-year period from 1967 to 1971, the median maximum temperature was $79^{\circ}$, and this value was used as CT for all of the September forecasts of maximum temperature. The results for CT are presented in Table 2, and they appear to be similar in many respects to the MT results. For example, on those occasions on which Forecaster 4 was on public weather forecasting duty, CT tended to underestimate OT in a manner similar to that exhibited by MT. In addition, a tendency to underestimate minimum temperatures and a lesser tendency to overestimate maximum temperatures were exhibited by the CT "forecasts."

The similarities between the results for CT and the results for MT suggest that tendencies such as a tendency to underestimate or overestimate may be due in part to unusual temperatures during the experimental period. For instance, the results for the climatological forecasts indicate

Table 2. A comparison of the climatological temperature (CT) and the observed temperature (OT).

| Set of Forecasts | Number of Forecasts | Percentages CT > OT | Percentages CT = OT | Percentages CT < OT | Average (Standard Deviation) CT-OT (°F) | Average (Standard Deviation) \|CT-OT\| (°F) | Average (Standard Deviation) $(CT-OT)^2$ $(°F)^2$ |
|---|---|---|---|---|---|---|---|
| All | 254 | 39.4 | 3.1 | 57.5 | 0.6 (12.0) | 8.9 (8.0) | 143.8 (273.9) |
| Variable-Width | 132 | 44.7 | 2.3 | 53.0 | 3.1 (13.8) | 10.7 (9.3) | 199.3 (351.7) |
| Fixed-Width | 122 | 33.6 | 4.1 | 62.3 | -2.0 (9.0) | 7.1 (5.8) | 83.7 (130.7) |
| Maximum | 127 | 53.5 | 6.3 | 40.2 | 4.7 (12.9) | 9.8 (9.6) | 186.9 (353.7) |
| Minimum | 127 | 25.2 | 0.0 | 74.8 | -3.5 (9.5) | 8.1 (6.0) | 100.7 (151.5) |
| 12-Hour | 127 | 45.7 | 3.9 | 50.4 | 1.5 (12.2) | 9.0 (8.4) | 149.6 (303.2) |
| 24-Hour | 127 | 33.1 | 2.4 | 64.6 | -0.2 (11.8) | 8.9 (7.6) | 137.1 (241.5) |
| Forecaster 1 | 64 | 43.8 | 0.0 | 56.3 | 3.6 (15.7) | 12.3(10.3) | 254.4 (415.2) |
| Forecaster 2 | 68 | 45.6 | 4.4 | 50.0 | 2.5 (12.0) | 9.2 (8.0) | 147.5 (272.2) |
| Forecaster 3 | 60 | 38.3 | 3.3 | 58.3 | -0.6 (9.8) | 7.5 (6.2) | 94.0 (146.0) |
| Forecaster 4 | 62 | 29.0 | 4.8 | 66.1 | -3.3 (8.0) | 6.7 (5.4) | 73.8 (114.3) |

that minimum temperatures were unusually high, on the average, during the period of the experiment, and the forecasters did not "correct" for this situation in formulating their MTs. On the other hand, the MTs were clearly much better point forecasts than the CTs. The average absolute difference between CT and OT was $8.9^{o}$, as compared with an average $|MT-OT|$ of $3.8^{o}$, and the average squared difference between CT and OT was $143.8(^{o})^{2}$, as compared with an average $(MT-OT)^{2}$ of $24.0(^{o})^{2}$. In formulating point forecasts, therefore, the forecasters were able to improve greatly upon climatology.

b) Variable-Width Credible Intervals

The results presented in Table 3 indicate that the variable-width forecasts were very reliable, in the sense that the observed relative frequencies below, in, and above the variable-width intervals were extremely close to the probabilities of the intervals. For the 50% intervals, the relative frequencies were 0.258, 0.455, and 0.288, respectively, as compared with probabilities of 0.25, 0.50, and 0.25. For the 75% intervals, the relative frequencies were 0.106, 0.735, and 0.159, and the probabilities were 0.125, 0.750, and 0.125. Thus, despite the reasonably large sample size (n=132), goodness-of-fit tests yield very small chi-square values for both the 50% and the 75% intervals. Moreover, the observed relative frequencies below, in, and above the intervals do not appear to be functions of the width of the credible intervals (see Table 4).

Table 3. Relative frequency of occurrence of observed temperature below interval (BI), in interval (II), and above interval (AI), and average interval width for (a) variable-width forecasts and (b) climatology.

| Set of Forecasts | Number of Forecasts | Percentages of Observed Temperatures: 50% Intervals BI | 50% Intervals II | 50% Intervals AI | 75% Intervals BI | 75% Intervals II | 75% Intervals AI | Average Width (Standard Deviation of Width)(°F): 50% Intervals | 75% Intervals |
|---|---|---|---|---|---|---|---|---|---|
| (a) Variable-width forecasts | | | | | | | | | |
| All | 132 | 25.8 | 45.5 | 28.8 | 10.6 | 73.5 | 15.9 | 6.23 (1.28) | 11.67 (2.23) |
| Maximum | 66 | 28.8 | 51.5 | 19.7 | 15.2 | 75.8 | 9.1 | 6.29 (1.24) | 11.74 (2.14) |
| Minimum | 66 | 22.7 | 39.4 | 37.9 | 6.1 | 71.2 | 22.7 | 6.18 (1.32) | 11.59 (2.33) |
| 12-Hour | 66 | 22.7 | 51.5 | 25.8 | 9.1 | 80.3 | 10.6 | 6.11 (1.22) | 11.44 (2.02) |
| 24-Hour | 66 | 28.8 | 39.4 | 31.8 | 12.1 | 66.7 | 21.2 | 6.36 (1.33) | 11.89 (2.41) |
| Forecaster 1 | 64 | 29.7 | 37.5 | 32.8 | 9.4 | 76.6 | 14.1 | 5.75 (1.27) | 11.34 (2.65) |
| Forecaster 2 | 68 | 22.1 | 52.9 | 25.0 | 11.8 | 70.6 | 17.6 | 6.69 (1.11) | 11.97 (1.70) |
| (b) Climatology | | | | | | | | | |
| All | 132 | 31.1 | 44.7 | 24.2 | 18.9 | 65.2 | 15.9 | 14.83 (4.22) | 24.15 (5.70) |
| Maximum | 66 | 39.4 | 50.0 | 10.6 | 24.2 | 69.7 | 6.1 | 18.17 (3.09) | 28.62 (4.16) |
| Minimum | 66 | 22.7 | 39.4 | 37.9 | 13.6 | 60.6 | 25.8 | 11.50 (1.96) | 19.68 (2.76) |
| 12-Hour | 66 | 28.8 | 50.0 | 21.2 | 18.2 | 69.7 | 12.1 | 15.36 (4.72) | 24.88 (6.13) |
| 24-Hour | 66 | 33.3 | 39.4 | 27.3 | 19.7 | 60.6 | 19.7 | 14.30 (3.62) | 23.42 (5.18) |
| Forecaster 1 | 64 | 32.8 | 39.1 | 28.1 | 21.9 | 59.4 | 18.8 | 15.52 (4.13) | 25.14 (5.89) |
| Forecaster 2 | 68 | 29.4 | 50.0 | 20.6 | 16.2 | 70.6 | 13.2 | 14.19 (4.24) | 23.22 (5.39) |

Table 4. Relative frequency of occurrence of observed temperature below interval (BI), in interval (II), and above interval (AI), and average error (|MT-OT|), all as a function of interval width for variable-width forecasts.

| Interval Width (°F) | Number of Forecasts | | Percentages of Observed Temperatures | | | | | | Average Error (Standard Deviation of Error)(°F) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 50% Intervals | | | 75% Intervals | | | | |
| | 50% Intervals | 75% Intervals | BI | II | AI | BI | II | AI | 50% Intervals | 75% Intervals |
| 3 | 2 | 0 | 50.0 | 50.0 | 0.0 | ---- | ---- | ---- | 3.00 (2.83) | ---- |
| 4 | 9 | 0 | 11.1 | 55.6 | 33.3 | ---- | ---- | ---- | 2.56 (2.30) | ---- |
| 5 | 22 | 0 | 31.8 | 36.4 | 31.8 | ---- | ---- | ---- | 3.09 (2.16) | ---- |
| 6 | 44 | 1 | 18.2 | 36.4 | 45.5 | 0.0 | 100.0 | 0.0 | 4.55 (3.36) | 1.00 (----) |
| 7 | 42 | 2 | 28.6 | 52.4 | 19.0 | 0.0 | 50.0 | 50.0 | 3.98 (3.30) | 2.50 (3.54) |
| 8 | 6 | 7 | 16.7 | 83.3 | 0.0 | 14.3 | 71.4 | 14.3 | 2.83 (2.40) | 3.00 (2.16) |
| 9 | 6 | 11 | 50.0 | 50.0 | 0.0 | 9.1 | 63.6 | 27.3 | 6.00 (5.33) | 3.64 (2.46) |
| 10 | 0 | 12 | ---- | ---- | ---- | 8.3 | 83.3 | 8.3 | ---- | 2.75 (2.86) |
| 11 | 1 | 29 | 100.0 | 0.0 | 0.0 | 6.9 | 82.8 | 10.3 | 11.00 (----) | 3.83 (3.05) |
| 12 | 0 | 25 | ---- | ---- | ---- | 8.0 | 72.0 | 20.0 | ---- | 3.92 (3.05) |
| 13 | 0 | 29 | ---- | ---- | ---- | 17.2 | 65.5 | 17.2 | ---- | 4.86 (3.54) |
| 14 | 0 | 6 | ---- | ---- | ---- | 0.0 | 66.7 | 33.3 | ---- | 3.83 (2.99) |
| 15 | 0 | 4 | ---- | ---- | ---- | 0.0 | 100.0 | 0.0 | ---- | 2.00 (3.37) |
| 16 | 0 | 3 | ---- | ---- | ---- | 66.7 | 33.3 | 0.0 | ---- | 10.33 (3.21) |
| 17 | 0 | 0 | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| 18 | 0 | 2 | ---- | ---- | ---- | 0.0 | 100.0 | 0.0 | ---- | 2.50 (2.12) |
| 19 | 0 | 0 | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| 20 | 0 | 0 | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| 21 | 0 | 1 | ---- | ---- | ---- | 0.0 | 100.0 | 0.0 | ---- | 11.00 (----) |
| **Total/Average** | 132 | 132 | 25.8 | 45.5 | 28.8 | 10.6 | 73.5 | 15.9 | 4.00 (3.25) | 4.00 (3.25) |

As in the case of the point forecasts, climatology can be used as a standard with which to compare the forecasters' variable-width interval forecasts. To generate 50% and 75% central credible intervals from the climatological data, the appropriate percentiles were determined for each month from the maximum and minimum temperatures during that month for the five-year period preceding the experiment. Table 3 includes the results for the climatological forecasts, and these intervals do not appear to be as reliable as the intervals determined by the forecasters. Furthermore, the average widths of the intervals were much greater for climatology (14.83$^{o}$ and 24.15$^{o}$ for the 50% and 75% intervals, respectively) than for the forecasters (6.23$^{o}$ and 11.67$^{o}$). Thus, the forecasters were able to use the information available to them to formulate interval forecasts that were very reliable and were much more precise than the interval forecasts derived from climatology.

Table 3 also indicates that for forecasts of minimum temperature, OT was above the interval more often than would be expected (37.9% of the time for the 50% intervals and 22.7% of the time for the 75% intervals). Of course, this result is consistent with the tendency for the forecasters' MTs to be underestimates of the OTs for minimum temperature forecasts. Similarly, for forecasts of maximum temperature, OT was below the interval slightly more often than expected (28.8% of the time for the 50% intervals and 15.2% of the time for the 75% intervals). Furthermore, the results in

Table 3 indicate that these tendencies were shared by the climatological intervals. As in the case of the point forecasts, tendencies such as a tendency to underestimate or overestimate in assessing credible intervals may be due in part to the nature of the weather during the experimental period.

Another result of interest is the occurrence of more temperatures outside of the intervals for the 24-hour forecasts than would be expected from the probabilities. Only 39.4% of the observations fell within the 50% intervals and only 66.7% of the observations fell within the 75% intervals (see Table 3). In the terminology of previous subjective probability forecasting experiments conducted in different contexts (e.g. Alpert and Raiffa, [1], Stael von Holstein, [9]), too many "surprises" occurred with respect to the 24-hour credible interval forecasts. This result suggests that the intervals were too narrow and that the forecasters failed to allow for the additional uncertainty in 24-hour forecasts as compared with 12-hour forecasts. Note, in Table 3, that the average widths of the 50% forecasts were only slightly greater for the 24-hour forecasts ($6.36^{\circ}$, as compared with $6.11^{\circ}$ for the 12-hour forecasts), and the same is true for the 75% forecasts ($11.89^{\circ}$, as compared with $11.44^{\circ}$).[2]

Finally, some small differences between the forecasters can be observed (see Table 3). Only 37.5% of the observations fell within Forecaster 1's 50% intervals, while 52.9% of the observations fell within Forecasters 2's 50% intervals. Note

also that the average width of the 50% intervals was almost one degree less for Forecaster 1 than for Forecaster 2 ($5.75^{\circ}$, as compared with $6.69^{\circ}$). For the 75% intervals, the two forecasters were much closer to each other and Forecaster 1 was much closer to the expected percentage of observations in the interval (even slightly above this expected percentage) than was the case with the 50% intervals.

The average absolute difference, or error, |MT-OT|, was expected to be an increasing function of the width of the 50% intervals and the width of the 75% intervals (see Peterson, Snapper, and Murphy, [7, p.969]). While the results presented in Table 4 do not indicate a strong relationship, a weak positive relationship seems to hold for the range of widths for which a reasonable number of cases exists (e.g., widths of $5^{\circ}$ and $7^{\circ}$ for the 50% intervals and of $11^{\circ}$ and $13^{\circ}$ for the 75% intervals). In addition, the relationship between the 50% intervals and the corresponding 75% intervals is of some interest. We would expect the width of the 75% intervals to be an increasing function of the width of the 50% intervals, and the results in Table 5 indicate that, on the average, such a relationship does indeed exist.

Although the variable-width credible intervals were constrained to be symmetric about the median in terms of probability, they need not be symmetric about the median in terms of width. That is, the difference between the 75th (87½th) percentile and the median need not equal the difference between the median and the 25th (12½) percentile. For the 50%

Table 5. Width of 75% intervals as a function of width of 50% intervals for variable-width forecasts.

| Width of 50% Intervals (°F) | Number of Forecasts | Average Width of 75% Intervals (Standard Deviation of Width) (°F) |
|---|---|---|
| 3 | 2 | 7.00 (1.41) |
| 4 | 9 | 8.00 (0.71) |
| 5 | 22 | 10.14 (1.32) |
| 6 | 44 | 11.34 (1.06) |
| 7 | 42 | 12.76 (1.19) |
| 8 | 6 | 13.67 (1.51) |
| 9 | 6 | 15.67 (1.63) |
| 10 | 0 | --- |
| 11 | 1 | 21.00 (----) |
| Total/Average | 132 | 11.67 (2.23) |

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 of 36 (67) (29) occasions. For the 75% intervals, the difference between the 87½th percentile and the median was less than (equal to) (greater than) the difference between the median and the 12½th 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 in this sense, and the asymmetries appeared to be approximately equally likely to be in either direction. Furthermore, when each interval was divided at the median (such a division yields two equally likely subintervals), the average absolute difference in width between the two subintervals was $0.57^{o}$ for the 50% intervals and $1.23^{o}$ for the 75% intervals (see Table 6). This absolute difference would be $0^{o}$ for an interval that is symmetric in terms of width. The preponderance of asymmetries among the central credible intervals suggests that fixed-width credible intervals that are constrained to be symmetric in width are not likely to be central credible intervals. Moreover, Table 6 indicates that the climatological forecasts were also asymmetric, which suggests that an underlying meteorological basis for asymmetric intervals may exist on many occasions in Denver.

**Table 6. A comparison of the difference between the upper limit of the** interval and the median ($D_U$) and the difference between the median and the lower limit of the interval ($D_L$) for variable-width forecasts (the figures in parentheses are for climatology).

| Set of Forecasts | Number of Forecasts | $\lvert D_U - D_L \rvert$ (°F) 50% Intervals | 75% Intervals |
|---|---|---|---|
| All | 132 | 0.57 (2.09) | 1.23 (3.77) |
| Maximum | 66 | 0.50 (3.01) | 1.32 (5.92) |
| Minimum | 66 | 0.64 (1.16) | 1.13 (1.62) |
| 12-Hour | 66 | 0.53 (2.30) | 1.26 (4.00) |
| 24-Hour | 66 | 0.60 (1.88) | 1.20 (3.54) |
| Forecaster 1 | 64 | 0.71 (2.24) | 1.66 (3.98) |
| Forecaster 2 | 68 | 0.39 (1.94) | 0.76 (3.56) |

c) Fixed-Width Credible Intervals

The results presented in Table 7 indicate that the average probabilities assigned by the forecasters to the $5^{\circ}$ and $9^{\circ}$ fixed-width credible intervals differed considerably from the relative frequencies with which OT fell in these intervals. The average probabilities were 0.60 and 0.80 for the $5^{\circ}$ and $9^{\circ}$ intervals, respectively, and the corresponding relative frequencies were 0.46 and 0.66. In both cases, the average probability was 0.14 higher than the relative frequency. These results suggest that the probabilities assigned by the forecasters to the fixed-width intervals were, on the average, larger than the observations indicate that they should have been. This situation also existed for the climatological fixed-width intervals, but to a lesser degree; climatology yielded forecasts that were more reliable than those of the forecasters in the fixed-width situation. In contrast, the results presented in 4.b reveal that the correspondence between the probabilities and the relative frequencies was quite close for the variable-width credible intervals and that the intervals determined by the forecasters were more reliable than those generated by climatology. Thus, the variable-width forecasts were much more reliable than the fixed-width forecasts.

Note that the average width of the 50% variable-width intervals was $6.23^{\circ}$, whereas the $5^{\circ}$ fixed-width intervals were assigned an average probability of 0.60; the 75% variable-width intervals averaged $11.67^{\circ}$ in width, whereas

Table 7. Average probability, observed relative frequency, and average Brier score for (a) fixed-width forecasts and (b) climatology.

| Set of Forecasts | Number of Forecasts | Average Probability Assigned to Intervals (Standard Deviation of Probability) 5°F Intervals | Average Probability Assigned to Intervals (Standard Deviation of Probability) 9°F Intervals | Relative Frequency of Observations in Intervals 5°F Intervals | Relative Frequency of Observations in Intervals 9°F Intervals | Average Brier Score (Standard Deviation of Score) 5°F Intervals | Average Brier Score (Standard Deviation of Score) 9°F Intervals |
|---|---|---|---|---|---|---|---|
| (a) Fixed-width forecasts | | | | | | | |
| All | 122 | 0.60 (0.16) | 0.80 (0.11) | 0.46 | 0.66 | 0.46 (0.31) | 0.47 (0.57) |
| Maximum | 61 | 0.61 (0.16) | 0.82 (0.11) | 0.39 | 0.59 | 0.47 (0.33) | 0.55 (0.62) |
| Minimum | 61 | 0.59 (0.15) | 0.79 (0.12) | 0.52 | 0.72 | 0.46 (0.28) | 0.39 (0.50) |
| 12-Hour | 61 | 0.60 (0.15) | 0.81 (0.11) | 0.44 | 0.67 | 0.43 (0.29) | 0.47 (0.58) |
| 24-Hour | 61 | 0.60 (0.16) | 0.79 (0.11) | 0.48 | 0.64 | 0.49 (0.32) | 0.47 (0.56) |
| Forecaster 3 | 60 | 0.62 (0.09) | 0.76 (0.11) | 0.47 | 0.67 | 0.50 (0.29) | 0.42 (0.48) |
| Forecaster 4 | 62 | 0.58 (0.20) | 0.84 (0.10) | 0.45 | 0.64 | 0.42 (0.32) | 0.52 (0.65) |
| (b) Climatology | | | | | | | |
| All | 122 | 0.23 (0.08) | 0.37 (0.12) | 0.19 | 0.43 | 0.30 (0.43) | 0.49 (0.35) |
| Maximum | 61 | 0.18 (0.06) | 0.29 (0.10) | 0.23 | 0.46 | 0.35 (0.54) | 0.54 (0.45) |
| Minimum | 61 | 0.28 (0.07) | 0.44 (0.08) | 0.15 | 0.39 | 0.25 (0.28) | 0.45 (0.17) |
| 12-Hour | 61 | 0.22 (0.08) | 0.35 (0.12) | 0.20 | 0.41 | 0.32 (0.46) | 0.50 (0.38) |
| 24-Hour | 61 | 0.24 (0.08) | 0.39 (0.11) | 0.18 | 0.44 | 0.28 (0.40) | 0.48 (0.30) |
| Forecaster 3 | 60 | 0.24 (0.09) | 0.38 (0.13) | 0.20 | 0.42 | 0.30 (0.42) | 0.46 (0.33) |
| Forecaster 4 | 62 | 0.22 (0.07) | 0.36 (0.10) | 0.18 | 0.44 | 0.30 (0.45) | 0.52 (0.36) |

the $9^{o}$ fixed-width intervals averaged 0.80 in probability. On the average, then, the fixed-width intervals have higher probabilities for narrower intervals when compared with the variable-width intervals. Narrow intervals are desirable provided that they are reliable, but the fixed-width intervals were not very reliable because they were too narrow on the average.

The discrepancies between the probabilities and relative frequencies were much larger for forecasts of maximum temperature (average probabilities 0.61 and 0.82; relative frequencies 0.39 and 0.59) than for forecasts of minimum temperature (average probabilities 0.59 and 0.79; relative frequencies 0.52 and 0.72). On the other hand, the 12-hour and 24-hour forecasts were very close in this respect, and noticeable differences between Forecasters 3 and 4 occurred only for the $9^{o}$ intervals.

In order to evaluate the fixed-width interval forecasts, the Brier score (B) (Brier, [4]) was computed for each forecast, and the average scores are presented in Table 7. The score on each occasion is given by

$$B = \begin{cases} 2(1 - r)^2 & \text{if OT in interval,} \\ 2r^2 & \text{if OT not in interval,} \end{cases}$$

where r is the probability assigned to the interval. Since a lower score is "better," the forecasts of minimum temperature were better in terms of average scores than the forecasts of maximum temperature. This result is consistent with

the discrepancy between average probabilities and relative frequencies for forecasts of maximum temperature. With regard to length of forecast, the 12-hour and 24-hour forecasts had identical average scores for the $9^{\circ}$ intervals, but the 12-hour forecasts were somewhat better than the 24-hour forecasts for the $5^{\circ}$ intervals. The difference between the forecasters was large and somewhat surprising in the sense that Forecaster 3 performed much better than Forecaster 4 for the $9^{\circ}$ intervals, but the reverse was true for the $5^{\circ}$ intervals.

Further, note that, according to the average Brier scores, the fixed-width forecasts were slightly better than the climatological forecasts for $9^{\circ}$ intervals (0.47 versus 0.49), while the climatological forecasts were considerably better than the fixed-width forecasts for the $5^{\circ}$ intervals (0.30 versus 0.46). This latter, apparently negative result can be explained in terms of the characteristics of the Brier score. In this regard, the expected Brier score is equal to $2r(1 - r)^2 + 2(1 - r)r^2$, and the expected scores corresponding to the average probabilities assigned to the $5^{\circ}$ intervals by the forecasters and climatology are 0.48 (for $r = 0.60$) and 0.35 (for $r = 0.23$), respectively. Thus, the climatological intervals would be expected to receive a considerably better average score than the forecasters' intervals because the proabilities assigned to the $5^{\circ}$ intervals were, on the average, further from 0.50 for climatology than for the forecasters.[3] Moreover, on the occasions of concern in this experiment, a large number of observed temperatures fell just

outside of the climatological intervals. Only slight changes in some of these observed temperatures would have increased the average score for climatology considerably. The difference between the average scores for the forecasters' $5^{o}$ intervals and the climatological $5^{o}$ intervals, then, does not necessarily reflect unfavorably upon the fixed-width interval forecasts.

In Table 8 the relative frequency of observations in the intervals is given as a function of the proability assigned to the intervals. A weak positive relationship appears to exist, but if these values were graphed, many of the points would lie far from the "ideal" diagonal $45^{o}$ line for which the observed relative frequency for each probability exactly equals that probability. In addition, the average absolute difference between MT and OT is given in Table 8 as a function of the probability assigned to the interval. The average error was expected to be a decreasing function of the probability, and, although the number of forecasts was limited for some probabilities, the results in Table 8 indicate that the average error did tend to decrease as the probability increased.

## 5. Summary and Discussion

In this paper we have described the results of an experiment involving credible interval temperature forecasts. The results indicate that, overall, the medians determined by the forecasters were good point forecasts of maximum and minimum temperature. Moreover, the forecasters were able to

Table 8. Observed relative frequency and average error ($|MT-OT|$) as a function of probability for fixed-width forecasts.

| Interval Probability | Number of Forecasts 5°F Intervals | Number of Forecasts 9°F Intervals | Relative Frequency of Observations in Interval 5°F Intervals | Relative Frequency of Observations in Interval 9°F Intervals | Average Error (Standard Deviation of Error)(°F) 5°F Intervals | Average Error (Standard Deviation of Error)(°F) 9°F Intervals |
|---|---|---|---|---|---|---|
| 0.30 | 2 | 0 | 0.00 | ---- | 3.50 (0.71) | ---- |
| 0.35 | 1 | 0 | 0.00 | ---- | 8.00 (----) | ---- |
| 0.40 | 22 | 0 | 0.23 | ---- | 4.82 (2.97) | ---- |
| 0.50 | 22 | 2 | 0.46 | 0.00 | 3.86 (3.04) | 8.00 (1.41) |
| 0.60 | 31 | 6 | 0.35 | 0.50 | 3.94 (2.69) | 4.17 (3.06) |
| 0.70 | 24 | 29 | 0.50 | 0.62 | 3.25 (2.85) | 3.97 (2.44) |
| 0.75 | 3 | 5 | 0.67 | 0.60 | 2.33 (4.04) | 3.60 (2.61) |
| 0.80 | 8 | 39 | 1.00 | 0.62 | 1.13 (0.99) | 4.18 (3.14) |
| 0.85 | 0 | 4 | ---- | 0.75 | ---- | 3.25 (2.50) |
| 0.90 | 7 | 20 | 0.86 | 0.75 | 2.14 (3.18) | 2.70 (2.99) |
| 0.95 | 0 | 4 | ---- | 0.50 | ---- | 3.25 (3.30) |
| 1.00 | 2 | 13 | 1.00 | 0.92 | 1.00 (1.41) | 1.69 (2.39) |
| Total/Average | 122 | 122 | 0.46 | 0.66 | 3.60 (2.92) | 3.60 (2.92) |

improve greatly upon climatology, as evidenced by the much smaller average error for the forecasters' medians than for the medians derived from climatology. Although one forecaster did exhibit a tendency to underestimate the observed temperatures, forecasts based upon climatology exhibited a similar tendency on those occasions, indicating that the underestimation may be explained in part by the temperatures on the particular occasions of concern. With regard to the maximum and minimum temperature forecasts, the forecasters underestimated minimum temperatures and slightly overestimated maximum temperatures on the average.

The variable-width credible intervals were very reliable in the sense that the observed relative frequencies corresponded very closely to the forecast probabilities. Although some specific instances existed where the correspondence was not as good (e.g. too many observations fell outside the intervals for 24-hour forecasts, indicating that the intervals should have been wider), overall the forecasters who formulated variable-width interval forecasts performed admirably. Furthermore, the variable-width interval forecasts were much more precise than the corresponding forecasts derived from climatology. The fixed-width intervals, on the other hand, were assigned probabilities that were, on the average, considerably higher than the corresponding relative frequencies. This lack of reliability of the fixed-width intervals was observed for all of the specific "stratifications" that were studied, with the discrepancy being the

greatest for forecasts of maximum temperature.

We realize that care must be taken when generalizing results based upon only four forecasters. However, despite the small sample, we believe that these experimental results have important implications for temperature forecasting. First, the use of probabilities, via credible intervals, in temperature forecasting allows the forecaster to express his degree of uncertainty concerning the maximum or minimum temperature. Point forecasts do not describe uncertainty, and interval forecasts without probabilities only describe uncertainty in a vague, informal manner. Second, to the extent that these experimental results indicate that credible interval temperature forecasting is feasible and that the procedures investigated here (particularly the variable-width procedure) yield reasonable results, these procedures could be very useful in temperature forecasting in practice. In this regard, further experimentation would be quite valuable in order to provide a larger sample of forecasts from which to make inferences and to investigate the use of these procedures in different meteorological and climatological regimes. In addition, the study of some considerations of interest (e.g. learning effects) requires a more extensive experiment. If possible, experiments in a fully operational setting involving both the formulation and the dissemination of credible interval temperature forecasts would be most desirable.

Although the experiment and the discussion have been oriented toward temperature forecasting, the procedures con-

sidered in this paper are quite general and can be used to determine credible interval forecasts of other continuous variables. As a result, the implications of the experiment extend far beyond temperature forecasting to forecasting of other meteorological variables and variables of interest in other fields (e.g. economic indicators). In this regard, the experiment reported upon in this paper was conducted in an operational setting and the participants were experienced weather forecasters. Thus, this experiment was much more realistic than most experiments that have been conducted in the area of subjective proability forecasting (e.g. see Winkler and Murphy, [12]), and with some instruction and "practice," the participants apparently had little difficulty understanding the task. We also see implications for further experimentation in the area of subjective probability forecasting, since we believe that more experiments should be conducted in realistic (preferably operational) settings with true experts serving as subjects.

APPENDIX

## Instructions for Variable-Width Interval Forecasting of Maximum and Minimum Temperature

In forecasting the maximum (max) and minimum (min) temperature, you undoubtedly are somewhat uncertain about what the actual max and min will be. It is possible to give a point forecast (i.e., a single value) that represents your "best estimate" about the max or min, but point forecast alone does not completely represent your uncertainty. A convenient way to convey this uncertainty is through the use of interval forecasts (i.e., intervals of values, as opposed to the single values used as point forecasts). Specifying an interval and the probability that the max (or min) temperature will be within the interval conveys a considerable amount of information about your uncertainty. On some days, you may feel that the odds are even that the max will be in a particular five degree interval; on other days, you may be much more uncertain, so you feel that the odds are even that the max will be in a particular ten degree interval. In this experiment you will be asked to determine an interval such that the probability is 50% that the max (or min) temperature will be in the interval, and you will be asked to determine an interval such that the probability is 75% that the max (or min) temperature will be in the interval. An interval is assumed to include its end points; for example,

the interval 72-76°F is a five degree interval (it includes 72, 73, 74, 75, and 76). Note that in determining your interval forecasts, you will be working with intervals that are of fixed probability (50% and 75%), and you will have to determine the end points of the intervals; hence, the intervals are of variable width (the width depending on how uncertain you are on a given occasion). Other participants in the experiment will be working with intervals that are of fixed width but variable probability -- you need not concern yourself with this procedure, since all of your forecasts will involve the fixed probability-variable width approach.

The first step in determining the interval forecasts is to determine a median, which will be used as a mid-point for the variable width intervals. A median is a value that you feel is equally likely to be exceeded or not exceeded. For example, if you feel that it is equally likely that the max temperature tomorrow will be above 74 or below 74, then 74 is your median. The following dialogue should illustrate how you might arrive at a median.

Experimenter: What is your best intuitive estimate of tomorrow's max temperature?

Forecaster: About 90 degrees.

Experimenter: My first step will be an attempt to sharpen up that initial estimate. If we were both to wager the same amount of money, would you rather bet that the max temperature will be above 90 degrees or below?

Forecaster: Above 90 degrees.

Experimenter: Would you rather bet that it will be above 94 degrees or below?

Forecaster: Below.

Experimenter: Above or below 91 degrees?

Forecaster: Hmmm...probably above.

Experimenter: Above or below 92 degrees?

Forecaster: It doesn't make much difference there.

Experimenter: Above or below 93 degrees?

Forecaster: Below.

Experimenter: Fine. Then we will select 92 degrees as your indifference judgment. You think that it is just as likely that tomorrow's max temperature will be above 92 degrees as that it will be below 92 degrees. Is that right?

Forecaster; That seems right.

Experimenter: In a sense, 92 degrees, which is a median, is is your best estimate of tomorrow's max temperature; it can be viewed as a point forecast.

The next step is to determine your 25th percentile (the median is sometimes called the 50th percentile). The 25th percentile is the value that divides the interval <u>below the median</u> into two equally likely subintervals. Note that the median divided the entire set of possible values into two equally likely intervals, so the procedure for determining the 25th percentile is very similar to the procedure for determining the median. For example, suppose that your median for the max temperature tomorrow is 74. Then if you feel that it is equally likely that the max temperature tomorrow will be below 71 or between 71 and 74, then 71 is your 25th percentile. The following continuation of the dialogue presented above illustrates the determination of a 25th percentile.

Experimenter: In a sense, 92 degrees, which is a "median," is your best estimate of tomorrow's max temperature. The next series of questions that I'll ask is designed to explore just how certain you are that tomorrow's max temperature will be near 92 degrees. First, assume that all bets are off in case the max temperature is greater than 92 degrees. Do you think that it is more likely that tomorrow's max temperature will fall below 80 degrees or between 80 and 92 degrees? I am after two equally likely intervals below 92 degrees.

Forecaster: It is more likely to be between 80 and 92 degrees.

Experimenter: Below 85 degrees, or between 85 and 92 degrees?

Forecaster: That's pretty difficult. Probably below 85 degrees.

Experimenter: Below 84 degrees or between 84 and 92 degrees?

Forecaster: That's about it. I can't choose between the two intervals.

Experimenter: Fine -- then we will accept 84 degrees as your 25th percentile.

Next, it is necessary to go through this type of procedure once more on the "low" side (the side below the median), in order to determine your 12½th percentile. As you can probably guess by now, the 12½th percentile divides the interval <u>below the 25th percentile</u> into two equally likely subintervals.

The dialogue continues:

Experimenter: Now that you've decided that 84 is your 25th percentile, let's assume that all bets are off if tomorrow's max temperature is above 84 degrees. Do you think that it is more likely that tomorrow's max temperature will fall below 70 degrees or between 70 and 84 degrees?

Forecaster: Between 70 and 84 degrees.

Experimenter: Below 75 degrees or between 75 and 84 degrees?

Forecaster: Between 75 and 84 degrees.

Experimenter: Below 80 degrees or between 80 and 84 degrees?

Forecaster: That's pretty close, but I'd say below 80 degrees.

Experimenter: Below 78 degrees or between 78 and 84 degrees?

Forecaster: Between 78 and 84 degrees, but it's pretty close again.

Experimenter: Below 79 degrees or between 79 and 84 degrees?

Forecaster: I guess those intervals are about equally likely.

Experimenter: Then we will select 79 degrees as your 12½th percentile.

The next step is to determine your 75th percentile, the value that divides the interval <u>above the median</u> into two equally likely subintervals. As you might suspect, the procedure for determining the 75th percentile is like the procedure for determining the 25th percentile. Let's go back to the dialogue.

Experimenter: Now let's move on to the upper range, the range above the median. Assuming that all bets are off if tomorrow's max temperature is below 92 degrees, do you think that it is more likely to be between 92 and 100 or above 100?

Forecaster: Definitely between 92 and 100.

Experimenter: Between 92 and 95 or above 95?

Forecaster: Still between 92 and 95.

Experimenter: Between 92 and 94 or above 94?

Forecaster: Now I am indifferent.

Experimenter: In that case we will take 94 as your 75th percentile.

Finally, it is necessary to determine your 87½th percentile, the value that divides the interval above the 75th percentile into two equally likely subintervals. The procedure is similar to that for determining the 12½th percentile, so the dialogue might be as follows:

Experimenter: If I can "push" you to determine one more indifference point, let's assume that all bets are off if the max temperature tomorrow is less than 94, which we just determined to be your 75th percentile. Do you think that the max temperature is more likely to be between 94 and 96 or above 96?

Forecaster: Between 94 and 96.

Experimenter: Between 94 and 95 or above 95?

Forecaster: That's pretty difficult, but I guess I'm about indifferent.

Experimenter: These are difficult judgments to make. Since you're about indifferent, we'll take 95 as your 87½th percentile.

The median, the 25th percentile, the 12½th percentile, the 75th percentile, and the 87½th percentile have been determined, in that order. These values can be used to determine interval forecasts. The probability is 50% that the max temperature will be between the 25th percentile and the 75th percentile, and the probability is 75% that the max temperature will be between the 12½th percentile and the 87½th percentile. Thus, we have one interval forecast with probability 50% and one with probability 75%. It is useful to

reconsider the values that have been determined to make sure that they coincide with your best judgments. To illustrate this, we return to the dialogue one more time.

Experimenter: Now let's carefully consider the values that you have estimated. First, consider the intervals A, B, C, and D, where A is below 84 degrees, B is between 84 and 92, C is between 92 and 94, and D is above 94. Assume that there is a four-way bet this time and you can pick only one of the intervals. Which one would you prefer?

Forecaster: Hmmm...Clearly not B or C. I guess I like A the best, but D looks pretty good, too.

Experimenter: People occasionally squeeze the outside boundaries in too closely when making judgments like this for the first time.

Forecaster: I must have done that because now I clearly like the outside two intervals better than the middle ones.

Experimenter: Then move the outer boundaries out one degree each so that the boundaries are at 83 degrees, 92 degrees, and 95 degrees. Now which interval would you prefer to bet on?

Forecaster: These estimates are better now. Any one of the intervals looks just as good as any other one to me. Also, I think that the max temperature is just as likely to fall inside the interval between 83 and 95 degrees as it is to fall outside that interval.

Experimenter: Good. Now let's consider the intervals P, Q, R, and S, where P is below 79 degrees, Q is between 79 and 83, R is between 95 and 96, and S is above 96. I have taken the liberty of shifting your 87½th percentile up to 96, since the 75th percentile is now 95. In a four-way bet among these four intervals, which one would you prefer?

Forecaster: The outside intervals look better again, so perhaps I need to move the 12½th and 87½th percentiles. Let's see -- suppose they were 78 and 97. The 97 seems okay, but the 78 might still be a little high. I guess 77 and 97 would make me indifferent.

Experimenter: Fine. Then your interval estimate with probability 50% is from 83 to 95, and your interval estimate with probability 75% is from 77 to 97. It is interesting that the boundaries are spread out asymmetrically around 92 degrees. The lower bound of 83 degrees has been pushed much farther away than the upper boundary of 95 degrees.

Forecaster: I was thinking about that when making my estimates. A weak cold front is moving in from the northwest. It may reach here early tomorrow morning, but it may take until tomorrow night. If it gets here before morning, then it won't get very warm tomorrow. But, it the front is delayed, then the max temperature should be around 92 degrees.

Experimenter: Then that explains why the upper boundary is so much closer to 92 degrees. There is little chance for any change in conditions to produce much of an increase above your median of 92.

Forecaster: That's right. Looked at that way, these intervals display a lot of what I know about tomorrow's max temperature. They don't indicate why the max temperature could drop but they certainly show that it can. I wouldn't expect to always have such asymmetric intervals when compared with the median, but it sure seems reasonable in this particular situation.

For convenience, here is a summary of the procedure. First, consider the maximum temperature in degrees Fahrenheit (on the day shift, this refers to tomorrow's maximum; on the midnight shift, this refers to today's maximum) and complete the following steps:

1. Determine your median.
2. Determine your 25th percentile.
3. Determine your 12½th percentile.
4. Determine your 75th percentile.
5. Determine your 87½th percentile.
6. Look at the resulting intervals to make sure that they agree with your judgments, making any changes you deem necessary.

Next, consider the minimum temperature in degrees Fahrenheit (on both the day and midnight shifts, this refers to tonight's minimum), and repeat the six steps listed above.[4]

## Footnotes

[1]CT represents only one of a set of possible "models," any of which could be used as a standard of comparison. Other models involving climatology and/or persistence can, of course, be formulated.

[2]Of course, in certain weather situations the credible intervals for 24-hour forecasts may be smaller than those for 12-hour forecasts. However, the results of this experiment indicate that, on the average, the widths of the 24-hour forecasts were, as expected, greater than the widths of the 12-hour forecasts.

[3]Ironically, the climatological forecasts are being "rewarded" by the Brier score for being "less certain" about the high and low temperatures, since the expected score decreases as r shifts away from 0.50 in either direction.

[4]The set of instructions provided to the forecasters concluded with response sheets to be used for "practice" forecasts. To conserve space, these response sheets are not included in the Appendix.

## References

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

[2] American Telephone and Telegraph Company, "Weather announcement study." New York, N.Y., American Telephone and Telegraph Company, Market and Service Plans Department, unpublished report, 1971.

[3] Bickert, C. von E. "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, 1967.

[4] Brier, G.W. "Verification of forecasts expressed in terms of probability." Monthly Weather Review, 78 (1950) 1-3.

[5] Klein, W.H., and H.R. Glahn, "Forecasting local weather by means of model output statistics." Bulletin of the American Meteorological Society, 55 (1974), in press.

[6] Murphy, A.H., and R.L. Winkler, "Probability forecasts: a survey of National Weather Service forecasters." Bulletin of the American Meteorological Society, 55 (1974), in press.

[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] Sanders, F., "Skill in forecasting daily temperature and precipitation: some experimental results." Bulletin of the American Meteorological Society, 53 (1973) 1171-1179.

[9] Stael von Holstein, C.-A.S. "Assessment and evaluation of subjective probability distributions." Stockholm, Sweden, Stockholm School of Economics, Economic Research Institute, 225 pp., 1970.

[10] Stael von Holstein, C.-A.S. "An experiment in probabilistic weather forecasting." Journal of Applied Meteorology, 10 (1971) 635-645.

[11] Winkler, R.L. "The assessment of prior distributions in Bayesian analysis." Journal of the American Statistical Association, 62 (1967) 776-800.

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