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

The Use of Credible Intervals in Temperature Forecasting: Some Experimental Results

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

IIASA Research Memorandum
December 1975

RM-75-70

THE USE OF CREDIBLE INTERVALS IN TEMPERATURE FORECASTING: SOME EXPERIMENTAL RESULTS

Allan H. Murphy
Robert L. Winkler

December 1975

Abstract

Probability can be thought of as the language of uncertainty, and, as such, it provides forecasters with a means of describing the uncertainty inherent in their forecasts in a formal, quantitative manner. Probability forecasts, in turn, provide potential users of forecasts with information required to make rational decisions in uncertain situations.

Since 1965, the National Weather Service (NWS) in the United States has routinely issued precipitation probability forecasts to the general public. Forecasts of maximum and minimum temperature, however, are still expressed in categorical terms (i.e., in terms of a specific temperature or a range of temperatures). In this paper we describe and compare the results of two recent experiments in which NWS forecasters used credible intervals to describe the uncertainty inherent in their temperature forecasts. A credible interval temperature forecast is simply an interval forecast accompanied by the forecaster's subjective probability that the temperature of concern will fall in the interval.

The experiments were conducted in the NWS forecast offices in Denver, Colorado and Milwaukee, Wisconsin and involved four and five forecasters, respectively. In each experiment, one group of forecasters made variable-width interval forecasts and the other group made fixed-width interval forecasts. In the variable-width approach, the forecasters determined 50% and 75% central credible intervals using the "method of successive subdivisions" (i.e., they assessed the median, 25th percentile, $12\frac{1}{2}$th percentile, 75th percentile, $87\frac{1}{2}$th percentile of their probability distributions, in that order). In the fixed-width approach, the forecasters first determined a median and then assessed probabilities for 5^{0}F and 9^{0}F intervals centered at the median.

In evaluating the results of the experiments, several properties of the forecasts (i.e., the medians and the intervals) were of interest, including their reliability and precision. Reliability refers to the degree of correspondence between the probabilities associated with the forecasts and the sample relative frequencies (i.e., the distribution of observed temperatures), while precision relates to the degree of correspondence between the forecasts and the observed temperatures on an individual basis. The latter can be measured by computing the average absolute error (and similar quantities) in the case of the medians and average scores based upon one or more "proper"

scoring rules in the case of the (fixed-width) intervals. The forecasts formulated by the forecasters were also compared with forecasts based solely upon climatological data. The results of the experiments indicate that NWS forecasters can formulate reliable and precise credible interval temperature forecasts and that these forecasts are generally better (in the sense of these two properties) than forecasts based upon climatological data.

The influences of a number of variables were considered in the process of evaluating the forecasts. These variables included: 1) location of experiment (Denver, Milwaukee); 2) type of interval (variable-width, fixed-width); 3) type of temperature (maximum, minimum); 4) forecast length (12 hours, 24 hours, 36 hours); and 5) forecaster. In addition, factors such as forecasting experience, training, and learning effects were investigated within the constraints imposed by the number of forecasters and the length of the experiments.

Finally, we briefly discuss the implications of these experiments and the results for probability forecasting in general and probability forecasting in meteorology in particular.

The Use of Credible Intervals in Temperature Forecasting: Some Experimental Results[1]

Allan H. Murphy[2] and Robert L. Winkler[3]

1. Introduction

Since 1965, the National Weather Service (NWS) in the United States has routinely issued precipitation probability forecasts to the general public. Forecasts of maximum and minimum temperature, however, are still expressed in categorical terms. NWS forecasters usually give point forecasts when forecasting temperature (e.g., "the high temperature tomorrow will be 75°F"), and such forecasts do not provide any information about the uncertainty inherent in the forecasts. Point forecasts are sometimes replaced by interval forecasts (e.g., "the high temperature tomorrow will be between 73°F and 77°F"), but such forecasts only provide a very informal representation of the forecaster's uncertainty. The potential user of the forecast does not know whether the forecaster is almost certain that the high temperature will fall in the forecast interval or whether the forecaster feels, say, that there is only a 50-50 chance that the high temperature will fall in the forecast interval.

Probability can be thought of as the language of uncertainty, and ideally a temperature forecast would consist of the forecaster's entire probability distribution for the temperature of concern. However, assessing an entire distribution for a continuous variable such as temperature may not be practical either in terms of the time required of the forecaster or in terms of reporting the forecast to the general public. A compromise that seems reasonable is to express temperature forecasts in terms of credible intervals. A credible interval temperature forecast is simply an interval forecast accompanied by the forecaster's subjective probability that the temperature of concern will fall in the interval (e.g., "the probability is 0.60 that the high temperature tomorrow will be between 73°F and 77°F").

---

[1]Supported in part by the National Science Foundation under Grants GA-31735 and GA-41232.

[2]The National Center for Atmospheric Research (NCAR), Boulder, Colorado. NCAR is sponsored by the National Science Foundation.

[3]Graduate School of Business, Indiana University, Bloomington, Indiana.

Thus, credible intervals represent a straightforward extension of the interval forecasts that are sometimes used in current temperature forecasting practice, as noted above, with the probability providing a formal representation of the forecaster's uncertainty regarding the temperature.

Although credible intervals appear to be a "natural" way to convey a weather forecaster's uncertainty about maximum and minimum temperatures, they have received very little attention in the context of temperature forecasting. The earliest work concerning credible interval temperature forecasting was reported in Peterson, Snapper, and Murphy (1972). In this paper the results of two recent experiments in which NWS forecasters used credible intervals to describe the uncertainty inherent in their temperature forecasts are described and compared. The experiments are described in Section 2, some results of the experiments are presented in Section 3, and Section 4 contains a summary and a brief discussion of the implications of the experimental results.

2. Design of the Experiments

The experiments were conducted in the NWS forecast offices in Denver, Colorado and Milwaukee, Wisconsin, and the subjects were experienced weather forecasters. The four forecasters participating in the Denver experiment averaged 26.0 years of weather forecasting experience and 5.8 years of probability forecasting experience, and the five forecasters participating in the Milwaukee experiment averaged 10.5 years of weather forecasting experience and 5.1 years of probability forecasting experience.

The Denver experiment was conducted from August 1972 to March 1973 and the results were analyzed and reported in Murphy and Winker (1974). The Milwaukee experiment was undertaken in order (1) to obtain a larger sample of forecasters and forecasts from which to make inferences regarding credible interval temperature forecasting and (2) to investigate the use of credible intervals in temperature forecasting in different meteorological and climatological regimes. The Milwaukee experiment was conducted from October 1974 to July 1975.

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. At Denver, forecasts were made for periods 12 and 24 hours in the future, and at Milwaukee, forecasts were made for periods 12, 24, and 36 hours in the future. The Denver forecasters formulated 32, 34, 30, and 31 sets of forecasts, and the Milwaukee forecasters formulated 42, 57, 45, and 44 sets of forecasts.

Two of the forecasters at Denver and three of the forecasters at Milwaukee worked within the framework of variable-width, fixed-probability forecasts, using 50% and 75% credible intervals. To obtain these intervals, the method of "successive subdivisions" (e.g., see Peterson, Snapper, and Murphy, 1972)

was used, requiring the forecaster to assess the median, the 25th percentile, the 12½th percentile, the 75th percentile, and the 87½th percentile. Each percentile involved an equal-odds indifference judgment (i.e., the division of an interval into two equally likely subintervals). The 50% credible interval is the interval from the 25th percentile to the 75th percentile, and the 75% credible interval is the interval from the 12½th percentile to the 87½th percentile.

The remaining two forecasters at each location worked within the framework of fixed-width, variable-probability forecasts, using intervals of width 5^0F and 9^0F. First, the median was assessed, just as in the case of the variable-width forecasts. Then, the forecaster assessed probabilities for intervals of width 5^0F and 9^0F centered at the median. All intervals in the experiments were assumed to include their end points, and all temperatures were recorded to the nearest degree.

At Denver, the authors met with the forecasters and discussed the concept of credible interval temperature forecasts, but we were not able to hold such a meeting at Milwaukee. At both locations, lengthy sets of written instructions were given to the participants. The forecasters then formulated their credible interval forecasts without any assistance from (or contact with) the authors.

3. Results of the Experiments

a) Reliability

The first task for each forecaster on each forecasting occasion was to determine a median, and a comparison of these median temperatures (MTs) with the corresponding observed temperatures (OTs) is presented in Table 1. In both experiments, the MTs slightly underestimated the OTs on the average. The percentage of OTs above the MTs (48.0% at Denver, 53.4% at Milwaukee) exceeds the percentage of OTs below the MTs (39.4% at Denver, 37.2% at Milwaukee), and the average values of MT-OT are slightly negative (-0.5^0F at Denver, -0.8^0F at Milwaukee). In both experiments, this tendency for MT to be a slight underestimate of OT was most pronounced for forecasts of minimum temperature. In fact, for forecasts of maximum temperature, MT slightly overestimated OT on the average. With respect to individual forecasters, Forecaster 4 at Denver and Forecaster 2 at Milwaukee exhibited the strongest tendency toward underestimation.

Climatological median temperatures (CTs) provide a convenient standard with which to compare MT as a point forecast. The climatological forecasts considered here are median maximum and minimum temperatures based upon historical data for the five-year periods immediately preceding the respective experiments, and they were computed on a monthly basis. These forecasts were analyzed in the same manner as the forecasters' assessed medians, and the results are presented in Table 2.

Table 1. A comparison of median temperature (MT) and observed temperature (OT).

Location	Set of Forecasts	Number of Forecasts	Percentage			Average (standard deviation)	
			OT<MT	OT=MT	OT>MT	MT-OT (^{0}F)	|MT-OT| (^{0}F)
Denver	All	254	39.4	12.6	48.0	-0.5(4.9)	3.8(3.1)
	Variable-width	132	44.7	12.1	43.2	-0.1(5.2)	4.0(3.3)
	Fixed-width	122	33.6	13.1	53.3	-0.8(4.6)	3.6(2.9)
	Maximum	127	47.2	13.4	39.4	0.6(4.8)	3.8(3.0)
	Minimum	127	31.5	11.8	56.7	-1.5(4.8)	3.8(3.2)
	12-hour	127	40.9	13.4	45.7	-0.2(4.7)	3.5(3.0)
	24-hour	127	37.8	11.8	50.4	-0.7(5.1)	4.1(3.2)
	Forecaster 1	64	45.3	10.9	43.8	0.0(5.3)	4.1(3.3)
	Forecaster 2	68	44.1	13.2	42.6	-0.3(5.1)	3.9(3.3)
	Forecaster 3	60	41.7	11.7	46.7	-0.1(4.6)	3.6(2.9)
	Forecaster 4	62	25.8	14.5	59.7	-1.5(4.4)	3.6(3.0)
Milwaukee	All	699	37.2	9.4	53.4	-0.8(4.8)	3.7(3.2)
	Variable-width	432	36.8	9.7	53.5	-0.6(4.4)	3.4(2.9)
	Fixed-width	267	37.8	9.0	53.2	-1.0(5.3)	4.1(3.5)
	Maximum	361	47.4	10.5	42.1	0.4(4.8)	3.5(3.2)
	Minimum	338	26.3	8.3	65.4	-2.0(4.5)	3.8(3.1)
	12-hour	233	39.5	7.7	52.8	-0.7(4.6)	3.4(3.1)
	24-hour	233	34.3	10.3	55.4	-1.2(4.2)	3.3(2.8)
	36-hour	233	37.8	10.3	51.9	-0.5(5.5)	4.2(3.5)
	Forecaster 1	126	40.5	12.7	46.8	-0.2(4.1)	3.0(2.8)
	Forecaster 2	171	30.4	8.2	61.4	-1.4(4.4)	3.4(3.0)
	Forecaster 3	135	41.5	8.9	49.6	-0.1(4.6)	3.7(2.8)
	Forecaster 4	135	37.0	8.1	54.8	-1.3(5.2)	4.1(3.4)
	Forecaster 5	132	38.6	9.8	51.5	-0.8(5.4)	4.1(3.6)

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

Location	Set of Forecasts	Number of Forecasts	Percentage			Average (standard deviation)	
			OT<CT	OT=CT	OT>CT	CT-OT(°F)	|CT-OT| (°F)
Denver	All	254	39.4	3.1	57.5	0.6(12.0)	8.9(8.0)
	Variable-width	132	44.7	2.3	53.0	3.1(13.8)	10.7(9.3)
	Fixed-width	122	33.6	4.1	62.3	-2.0(9.0)	7.1(5.8)
	Maximum	127	53.5	6.3	40.2	4.7(12.9)	9.8(9.6)
	Minimum	127	25.2	0.0	74.8	-3.5(9.5)	8.1(6.0)
	12-hour	127	45.7	3.9	50.4	1.5(12.2)	9.0(8.4)
	24-hour	127	33.1	2.4	64.6	-0.2(11.8)	8.9(7.6)
	Forecaster 1	64	43.8	0.0	56.3	3.6(15.7)	12.3(10.3)
	Forecaster 2	68	45.6	4.4	50.0	2.5(12.0)	9.2(8.0)
	Forecaster 3	60	38.3	3.3	58.3	-0.6(9.8)	7.5(6.2)
	Forecaster 4	62	29.0	4.8	66.1	-3.3(8.0)	6.7(5.4)
Milwaukee	All	699	40.6	4.6	54.8	-1.3(8.9)	7.1(5.5)
	Variable-width	432	39.6	6.5	53.9	-0.7(8.7)	6.8(5.5)
	Fixed-width	267	42.3	1.5	56.2	-2.4(9.2)	7.7(5.6)
	Maximum	361	51.0	4.2	44.9	0.6(8.9)	7.0(5.6)
	Minimum	338	29.6	5.0	65.4	-3.4(8.4)	7.2(5.5)
	12-hour	233	43.8	4.3	51.9	-1.0(9.3)	7.4(5.8)
	24-hour	233	36.9	4.3	58.8	-1.8(8.5)	6.9(5.3)
	36-hour	233	41.2	5.2	53.6	-1.2(8.9)	7.1(5.5)
	Forecaster 1	126	39.7	5.6	54.8	-1.1(6.3)	5.3(3.6)
	Forecaster 2	171	42.7	7.0	50.3	-0.5(9.2)	7.0(5.9)
	Forecaster 3	135	35.6	6.7	57.8	-0.6(9.9)	7.8(6.1)
	Forecaster 4	135	44.4	1.5	54.1	-1.3(9.9)	8.2(5.6)
	Forecaster 5	132	40.2	1.5	58.3	-3.5(8.3)	7.1(5.5)

As in the case of MT, CT exhibits a slight tendency to underestimate OT. Also, on the average, CT underestimates OT strongly for forecasts of minimum temperature and overestimates OT for forecasts of maximum temperature. Thus, the forecasters' tendency to underestimate or overestimate may be due in part to unusual (i.e., above or below normal, respectively) temperatures during the experimental periods. In any event, these tendencies are not strong, and the medians appear to be quite reliable point forecasts.

For probability forecasts, reliability refers to the degree of correspondence between the probabilities and the sample relative frequencies. The results presented in Table 3 indicate that the variable-width intervals were extremely reliable at both locations. For the 50% intervals, the relative frequencies were 0.258, 0.455, and 0.288 at Denver and 0.181, 0.539, and 0.280 at Milwaukee, 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 at Denver and 0.081, 0.794, and 0.125 at Milwaukee, as compared with probabilities of 0.125, 0.75, and 0.125. The tendency to underestimate noted in the discussion of the MTs is reflected in the higher relative frequency of occurrence of OTs above the intervals than below the intervals.

As in the case of point forecasts, climatology can be used as a standard of comparison. Climatological variable-width interval forecasts were generated by determining the appropriate percentiles from the five years of historical data on a monthly basis, and the performance of the climatological variable-width intervals is summarized in Table 4. An examination of the percentages of observations below, in, and above the intervals indicates that the climatological intervals do not appear to be quite as reliable as the intervals determined by the forecasters.

The results for the forecasters' fixed-width intervals and the corresponding climatological intervals are given in Tables 5 and 6, respectively. At Denver, the relative frequencies of observations in the intervals (0.46 for the 5^0F intervals, 0.66 for the 9^0F intervals) were considerably lower than the average probabilities assigned to the intervals (0.60 and 0.80, respectively). At Milwaukee, on the other hand, the relative frequencies (0.40 for the 5^0F intervals, 0.66 for the 9^0F intervals) were lower than the average probabilities (0.47 and 0.72, respectively) but were closer to the average probabilities than at Denver. Thus, the fixed-width intervals were less reliable than the variable-width intervals, although the fixed-width intervals at Milwaukee were quite reliable. At both Denver and Milwaukee, the climatological fixed-width intervals were more reliable than the forecasters' intervals.

b) Precision

It is possible, of course, for point and interval forecasts to be reliable without being very precise. For instance, point forecasts could differ from the observed temperatures by very

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

Location	Set of Forecasts	Number of Forecasts	Percentage 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
Denver	All	132	25.8	45.5	28.8	10.6	73.5	15.9	6.2(1.3)	11.7(2.2)
	Maximum	66	28.8	51.5	19.7	15.2	75.8	9.1	6.3(1.2)	11.7(2.1)
	Minimum	66	22.7	39.4	37.9	6.1	71.2	22.7	6.2(1.3)	11.6(2.3)
	12-hour	66	22.7	51.5	25.8	9.1	80.3	10.6	6.1(1.2)	11.4(2.0)
	24-hour	66	28.8	39.4	31.8	12.1	66.7	21.2	6.4(1.3)	11.9(2.4)
	Forecaster 1	64	29.7	37.5	32.8	9.4	76.6	14.1	5.8(1.3)	11.3(2.6)
	Forecaster 2	68	22.1	52.0	25.0	11.8	70.6	17.6	6.7(1.1)	12.0(1.7)
Milwaukee	All	432	18.1	53.9	28.0	8.1	79.4	12.5	5.9(2.0)	10.1(3.1)
	Maximum	216	24.1	57.9	18.1	10.2	82.9	6.9	5.9(2.0)	10.1(3.2)
	Minimum	216	12.0	50.0	38.0	6.0	75.9	18.1	5.9(1.9)	10.1(3.0)
	12-hour	144	20.1	50.0	29.9	8.3	81.9	9.7	5.5(1.8)	9.3(2.8)
	24-hour	144	13.2	56.9	29.9	4.9	77.8	17.4	6.0(2.1)	10.3(3.3)
	36-hour	144	20.8	54.9	24.3	11.1	78.5	10.4	6.2(2.0)	10.6(3.0)
	Forecaster 1	126	19.0	53.2	27.8	13.5	72.2	14.3	4.8(0.9)	8.1(1.3)
	Forecaster 2	171	12.9	59.1	28.1	4.1	82.5	13.5	6.5(2.5)	10.5(3.4)
	Forecaster 3	135	23.7	48.1	28.1	8.1	82.2	9.6	6.2(1.6)	11.3(2.9)

Table 4. Relative frequency of observed temperature below interval (BI), in interval (II), and above interval (AI), and average interval width for climatological forecasts corresponding to variable-width forecasts.

Location	Set of Forecasts	Number of Forecasts	Percentage 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
Denver	All	132	31.1	44.7	24.2	18.9	65.2	15.9	14.8(4.2)	24.2(5.7)
	Maximum	66	39.4	50.0	10.6	24.2	69.7	6.1	18.2(3.1)	28.6(4.2)
	Minimum	66	22.7	39.4	37.9	13.6	60.6	25.8	11.5(2.0)	19.7(2.8)
	12-hour	66	28.8	50.0	21.2	18.2	69.7	12.1	15.4(4.7)	24.9(6.1)
	24-hour	66	33.3	39.4	27.3	19.7	60.6	19.7	14.3(3.6)	23.4(5.2)
	Forecaster 1	64	32.8	39.1	28.1	21.9	59.4	18.8	15.5(4.1)	25.1(5.9)
	Forecaster 2	68	29.4	50.0	20.6	16.2	70.6	13.2	14.2(4.2)	23.2(5.4)
Milwaukee	All	432	17.8	56.9	25.2	8.1	81.7	10.2	14.5(3.9)	23.7(4.9)
	Maximum	216	24.1	63.0	13.0	12.0	82.4	5.6	14.7(3.2)	24.2(3.7)
	Minimum	216	11.6	50.9	37.5	4.2	81.0	14.8	14.3(4.5)	23.1(5.8)
	12-hour	144	19.4	57.6	22.9	9.7	79.2	11.1	14.6(4.0)	23.8(4.9)
	24-hour	144	18.1	53.5	28.5	7.6	79.9	12.5	14.4(3.6)	23.5(4.8)
	36-hour	144	16.0	59.7	24.3	6.9	86.1	6.9	14.5(4.1)	23.7(5.0)
	Forecaster 1	126	13.5	63.5	23.0	5.6	87.3	7.1	13.8(4.0)	22.5(5.0)
	Forecaster 2	171	18.7	57.9	23.4	8.2	79.5	12.3	14.7(4.1)	24.1(4.9)
	Forecaster 3	135	20.7	49.6	29.6	10.4	79.3	10.4	15.0(3.6)	24.2(4.6)

Table 5. Average probability and observed relative frequency for fixed-width forecasts.

Location	Set of Forecasts	Number of Fore-casts	Average probability assigned to intervals		Relative frequency of observations in intervals	
			5°F Intervals	9°F Intervals	5°F Intervals	9°F Intervals
Denver	All	122	0.60	0.80	0.46	0.66
	Maximum	61	0.61	0.82	0.39	0.59
	Minimum	61	0.59	0.79	0.52	0.72
	12-hour	61	0.60	0.81	0.44	0.67
	24-hour	61	0.60	0.79	0.48	0.64
	Forecaster 3	60	0.62	0.76	0.47	0.67
	Forecaster 4	62	0.58	0.84	0.45	0.64
Milwaukee	All	267	0.47	0.72	0.40	0.66
	Maximum	145	0.48	0.73	0.43	0.69
	Minimum	122	0.45	0.70	0.37	0.62
	12-hour	89	0.54	0.80	0.45	0.75
	24-hour	89	0.48	0.73	0.49	0.71
	36-hour	89	0.39	0.63	0.27	0.52
	Forecaster 4	135	0.50	0.73	0.40	0.67
	Forecaster 5	132	0.44	0.71	0.41	0.64

Table 6. Average probability and observed relative frequency for climatological forecasts corresponding to fixed-width forecasts.

Location	Set of Forecasts	Number of Fore-casts	Average probability assigned to intervals		Relative frequency of observations in intervals	
			5^0F Intervals	9^0F Intervals	5^0F Intervals	9^0F Intervals
Denver	All	122	0.23	0.37	0.19	0.43
	Maximum	61	0.18	0.29	0.23	0.46
	Minimum	61	0.28	0.44	0.15	0.39
	12-hour	61	0.22	0.35	0.20	0.41
	24-hour	61	0.24	0.39	0.18	0.44
	Forecaster 3	60	0.24	0.38	0.20	0.42
	Forecaster 4	62	0.22	0.36	0.18	0.44
Milwaukee	All	267	0.22	0.37	0.19	0.36
	Maximum	145	0.23	0.38	0.19	0.39
	Minimum	122	0.21	0.36	0.20	0.33
	12-hour	89	0.21	0.36	0.15	0.34
	24-hour	89	0.23	0.40	0.28	0.42
	36-hour	89	0.21	0.36	0.16	0.33
	Forecaster 4	135	0.22	0.37	0.19	0.30
	Forecaster 5	132	0.22	0.38	0.20	0.42

little on the average because large positive and negative differences cancel. For the medians, a measure of the precision of individual forecasts is provided by the absolute difference between the median temperature and the observed temperature. Average values of |MT - OT| and |CT - OT| are given in Tables 1 and 2, respectively. The average |MT - OT| was 3.8°F at Denver and 3.7°F at Milwaukee, whereas the average |CT - OT| was 8.9°F at Denver and 7.1°F at Milwaukee. Thus, the forecasters' medians were much more precise point forecasts than the climatological medians, and the results in Table 1 show that the precision is remarkably consistent across forecasters and different types of forecasts.

A measure of precision for the variable-width forecasts is provided by the width of the interval, and average widths are given in Tables 3 and 4. For the 50% intervals, the average widths were 6.2°F and 5.9°F at Denver and Milwaukee, respectively. The corresponding average widths for the climatological variable-width intervals were 14.8°F and 14.5°F. The average widths for the 75% intervals were 11.7°F at Denver and 10.1°F at Milwaukee, and the corresponding average widths for the climatological intervals were 24.1°F and 23.7°F, respectively. Thus, the forecasters' variable-width intervals were considerably more precise than the corresponding climatological intervals. The average widths did not vary too much by forecaster or type of forecast, although Forecaster 1 at Milwaukee had particularly narrow intervals and interval widths tended to increase as the lead time increased from 12 hours to 24 hours and 36 hours.

The precision of the fixed-width forecasts can be investigated by examining the average probabilities assigned to the intervals. From Tables 5 and 6, it is clear that the forecasters' fixed-width intervals were more precise than the corresponding climatological intervals. The average probabilities assigned to the 5°F intervals were 0.60 at Denver and 0.47 at Milwaukee, as compared with 0.23 at Denver and 0.22 at Milwaukee for climatology. For the 9°F intervals, the average probabilities were 0.80 at Denver and 0.72 at Milwaukee, as compared with 0.37 at both locations for climatology. The average probabilities did not seem to vary systematically across any of the stratifications that were considered with the exception that they tended to decrease as the lead time associated with the forecasts increased, especially at Milwaukee.

c) Forecasting Experience and Learning Effects

The effect of forecasting experience, as well as learning effects during the course of the experiments, have also been investigated. With regard to experience, the one forecaster whose performance seemed to be somewhat better than that of the other forecasters was Forecaster 1 at Milwaukee. Ironically, Forecaster 1 had the least forecasting experience (1.5 years) and the least probability forecasting experience (1.5 years) of any of the forecasters participating in the experiments.

Of course, 1.5 years still represents a fair amount of experience. Moreover, Forecaster 1 was the youngest forecaster involved in the experiments and was most likely less removed in terms of time from his formal college training in meteorology than were the other forecasters.

In order to investigate learning effects, the set of forecasts for each forecaster was divided into two halves, with one half consisting of the first 50% of the forecasts formulated by that forecaster and the other half consisting of the last 50% of that forecaster's forecasts. Each of these halves was then analyzed separately. Space limitations do not permit inclusion or detailed discussion of the results, but no large, systematic differences appeared to exist between the two sets of data. Thus, in the crude sense of dividing the sample of forecasts into two halves, no learning effects could be discerned from the results of the experiments.

4. Summary and Discussion

In this paper some results of two experiments involving credible interval temperature forecasts have been described. The results indicate that for both experiments, the medians determined by the forecasters were good point forecasts of maximum and minimum temperatures. They were just as reliable as, and much more precise than, the corresponding climatological forecasts. The medians tended to underestimate the observed temperatures slightly, particularly for forecasts of minimum temperature. However, the same tendency was exhibited by climatological forecasts, indicating that the underestimation may be explained in part by the fact that the temperatures during the experimental periods were, on the average, above normal.

The variable-width credible intervals were very reliable at both locations in the sense that the observed relative frequencies corresponded closely with the forecast probabilities. Moreover, the intervals were much narrower on the average than the corresponding climatological intervals. The fixed-width intervals were not as reliable as the variable-width intervals, although it should be noted that the reliability of the fixed-width intervals was much better at Milwaukee than at Denver. Specifically, the probabilities assigned to the fixed-width intervals were larger on the average than the corresponding relative frequencies. The forecasters' fixed-width intervals were more precise than climatological fixed-width intervals, but the former's lack of reliability suggests that this precision may be at least partially spurious.

A brief investigation revealed that, for the particular sample of nine forecasters in the two experiments, experience at weather forecasting in general and probabilistic weather forecasting in particular did not appear to have any effect on the results. In addition, stratification of the experimental

results into two halves did not reveal any systematic learning effects.

The results of the experiments discussed in this paper have implications for weather forecasting and for probability forecasting in general. With respect to weather forecasting, the experiments demonstrate that forecasters can formulate reliable, precise credible interval temperature forecasts. In this regard, it is of interest to note that NWS forecasters routinely receive "objective" guidance forecasts for most of the variables for which they must formulate forecasts operationally and that the guidance forecasts for many of these variables are expressed in probabilistic terms (e.g., see Klein and Glahn, 1974). The "objective" forecasts of maximum and minimum temperature, however, are point forecasts. Evidently, NWS forecasters are able to quantify the uncertainty inherent in their temperature forecasts in a reliable and precise manner in the absence of such information.

Credible interval forecasts could be very useful in temperature forecasting, and consideration should be given to formulating such forecasts on a regular, operational basis. However, prior to the initiation of such a "program," it would be desirable to conduct more extensive credible interval temperature forecasting experiments in several different locations and to investigate the practical problems related to routinely issuing these forecasts to the general public. In addition, the possibility of making credible interval forecasts for other continuous meteorological variables should be investigated.

With respect to probability forecasting in general, the high reliability of the variable-width forecasts is quite encouraging, particularly in view of the results of similar experiments in other (e.g., non-meteorological) contexts which lack such reliability (e.g., see Hogarth, 1975). Perhaps the fact that the forecasts at Denver and Milwaukee were made by experienced weather forecasters in an operational setting (as opposed to inexperienced subjects in a laboratory experiment) contributed to their reliability (see Winkler and Murphy, 1973). This supposition could be investigated further by experimentation involving various types of probability forecasts in other realistic situations. Finally, credible interval forecasting itself deserves additional attention, both in the laboratory and in the real world.

References

Hogarth, R.M. (1975). "Cognitive Processes and the Assessment of Subjective Probability Distributions." Journal of the American Statistical Association, 70, 271-294.

Klein, W.H., and Glahn, H.R. (1974). "Forecasting Local Weather by Means of Model Output Statistics." Bulletin of the American Meteorological Society, 55, 1217-1227.

Murphy, A.H., and Winkler, R.L. (1974). "Credible Interval Temperature Forecasting: Some Experimental Results." Monthly Weather Review, 102, 784-794.

Peterson, C.R., Snapper, K.J., and Murphy, A.H. (1972). "Credible Interval Temperature Forecasts." Bulletin of the American Meteorological Society, 53, 966-970.

Winkler, R.L., and Murphy, A.H. (1973). "Experiments in the Laboratory and the Real World." Organizational Behavior and Human Performance, 10, 252-270.