An Experiment in Probabilistic Weather Forecasting[1]

CARL-AXEL S. STAËL VON HOLSTEIN

University of Stockholm and The Economic Research Institute at the Stockholm School of Economics, Sweden

(Manuscript received 15 January 1971, in revised form 30 March 1971)

ABSTRACT

Probabilistic forecasts were given for temperature and precipitation for 2 and 5-day periods. They were evaluated by a scoring rule and the scores were given as feedback to the participants together with the actual outcomes. On the average the forecasts did not improve on forecasts based on climatology, and the study indicates some explanations for this. It also shows how the meteorologists' knowledge could be used to produce probabilistic forecasts that are better than climatology. The study also covers empirical evaluation of different scoring rules and of techniques for aggregating sets of forecasts.

1. Introduction

Most weather forecasts today are given in a deterministic way, although precipitation forecasts in the United States are generally stated in probabilistic terms. This means, for instance, that a forecast of the average temperature for the next day is given either in the form of a point or interval estimate, e.g., 27C or 68–72F, or in qualitative terms, e.g., "warmer than today." A forecast such as this can perhaps be interpreted as the temperature value the forecaster judges as most likely to occur.

A user of a weather forecast is likely to be interested in additional information about the degree of certainty the forecaster assigns to his estimate. That is, he would like the forecast to be expressed in terms of a probability distribution. A precipitation forecast could then be expressed as a set of probabilities for different classes representing different amounts of precipitation. These classes could be defined either by their limits expressed in millimeters (or inches) or by qualitative expressions like "no rain," "little rain," and so on, where it is assumed that these qualitative terms can be given quantitative interpretations.

There exist several probability forecasting systems which are based on historic data and which sometimes are called "objective" (e.g., Thompson, 1950; Russo *et al.*, 1966). They are often used without modification by the forecaster. This study, however, is based on the personalistic view of probability. A probability forecast then represents a quantification of the forecaster's judgment of the situation at hand. This means that the probabilities are subjective since they are based on the forecaster's personal beliefs. A forecaster may often base his forecasts on "objective" schemes but his forecast will still be subjective because he will use his judgment to decide how to combine that particular source of information with other kinds of information.

The personalistic school of probability maintains that probabilities are to be regarded as measures of degrees of belief. The mathematical foundations of personal probability have been laid down by de Finetti (1937) and Savage (1954), among others. If an individual assesses probabilities that are consistent with certain reasonable behavioristic postulates, then it can be shown that these probabilities correspond mathematically to a probability measure. The personalistic school does not specify a "correct" assessment. It only requires that assessments be consistent with certain postulates and that they correspond to the assessor's true judgments.

One aim of the present experiment was to study how weather forecasts could be expressed in probabilistic terms. The predictands (described in more detail in the next section) concerned temperature and precipitation and were chosen because they are major parts of the weather reports issued to the public in Sweden. Numerous experiments have been made with probabilistic forecasting in meteorological contexts. The present experiment differs from most of the others in that it contains feedback to the participants with evaluation of their forecasts. [Enger *et al.* (1964) and Allen (1969) report on experiments with feedback concerning forecasts of ceiling and visibility. It is also true that evaluations of official probability forecasts include feedback since the forecasters are being evaluated continuously. That feedback, however, is not controlled by an experimenter and it is difficult to evaluate its impact.] It was also intended to study whether this feedback helped the forecasters improve their forecasts. The values of the predictands were divided into a number of classes which varied from three to eight, whereas

[1] This study was supported by a grant from The Bank of Sweden Tercentenary Fund.

most past experiments in probabilistic weather forecasting seem to have been restricted to dichotomous situations, e.g., with the two states "rain" and "no rain" [e.g., Sanders (1963) and Winkler and Murphy (1968a); for a non-dichotomous example, see, e.g., Enger *et al.* (1964)]. A decision maker is often best helped if he can have a more detailed forecast. Such a situation will be more difficult in some sense, however, than a dichotomous situation.

A second aspect of the experiment is that it allows a study of probability assessments within a practical problem area. Most non-meteorological experiments dealing with the assessment of subjective probability distributions have been of a laboratory nature with artificial problems. It is understandable that there are only a few experiments that relate to practical problems. Such experiments require that the problem be of a repetitive nature and that the true event be known after not too long a time. This is essential so as to give participants the possibility of learning about the problem situation. Other examples of suitable problems have been found in the area of football (de Finetti, 1962; Winkler, 1967b). Staël von Holstein (1969) used the development of buying prices on the stock exchange.

2. Design of the experiment

a. Problem

The assessment task concerned the following five quantities (day 0 represents the day of the forecast):

1) The average temperature for days 1–2.
2) The average temperature for days 4–5.
3) The total amount of rain[2] for days 1–5.
4) The number of days out of days 1–5 with the amount of rain exceeding 0.05 mm.
5) The number of days out of days 1–2 with the amount of rain exceeding 0.05 mm.

The range of possible temperatures was divided into eight classes. The class limits were chosen so as to assign the classes the same climatological probabilities, i.e., make them equally frequent when historical records of average temperatures for days 1–5 were considered. The amount of rain was divided into four classes. The class limits were chosen so that the classes had the climatological probabilities $\frac{3}{8}$, $\frac{1}{4}$, $\frac{1}{4}$ and $\frac{1}{8}$ (going from no rain to an infinite amount), respectively. The classes used for the first three quantities coincided with those used for the official forecasts issued by the SMHI (Swedish Meteorological and Hydrological Institute). Their limits changed, of course, as the experiment progressed. The last two quantities were divided naturally into six and three classes, respectively.

The first three quantities were chosen because they are closely connected with the 5-day forecasts that are issued twice weekly by SMHI. Temperature and precipitation are also two weather characteristics that non-meteorologists have some ideas about. A user of precipitation forecasts could often be more interested in the number of days with rain than in the total amount of rain. Forecasting the number of days with rain was therefore thought to be an interesting (and probably difficult) problem for the participants.

The experiment covered a period of eight weeks with two sessions being held each week. Each session started on either a Monday or a Thursday, which are the days when the official five-day forecasts are issued. Every forecast thus had to be made in the course of a session, i.e., the results from one session could not influence the next session. The participants were asked to submit their forecasts before 1730 local time when the official forecasts were broadcasted.

b. Participants

The participants can be divided into four groups. The first group included four meteorologists from SMHI, of whom only two participated each time. These were the two persons who prepared the official 5-day forecasts. The second group consisted of six research assistants from the Department of Meteorology at the University of Stockholm. The third group consisted of eight students of meteorology. The fourth group included twelve people associated with the Institute of Mathematical Statistics at the University of Stockholm. Nine of these had participated in a similar experiment concerning stock market prices (Staël von Holstein, 1969). Prior to the experiment it was assumed that the first three groups of participants would represent a high standard of substantive goodness while the statisticians might have a high standard of normative goodness.[3]

Thirteen out of 26 non-meteorologists failed to complete all sixteen sessions; they missed a total of only 24 sessions.

c. Evaluation of assessments

Each forecast was evaluated by a scoring rule which related the forecast to the actual outcome. More explicitly, let $\mathbf{r}=(r_1,\cdots,r_n)$ be a forecast for a quantity with n classes. The quadratic scoring rule, which was used in the feedback, assigns the following score to the forecast when the kth event turns out to be true:

$$Q_k(\mathbf{r})=0.5(1+2r_k-\sum r_i^2).$$

The highest possible score is 1 (when $r_k=1$) and the lowest 0 (when some other r_i is 1). The quadratic scoring rule is equivalent to the probability score, first defined by Brier (1950), which is the scoring rule most often

[2] We shall use "rain" as a synonym for "precipitation" since rain was the only form of precipitation during the actual period (15 April–10 June 1969).

[3] "In essence, then, the normative standard of 'goodness' concerns expertise in probability assessment, while the substantive standard of 'goodness' concerns expertise in the domain in which the assessments are made." (Winkler and Murphy, 1968b, p. 753)

TABLE 1. Results after the final session.

Meteorologists			Assistants			Students			Statisticians		
No.	Score	Rank*	No.	Score	Rank	No.	Score	Rank	No.	Score	Rank
1	2.768	30	5	3.162	3	11	2.915	22	19	3.018	13
2	3.201	1	6	3.070	9	12	2.795	29	20	2.753	31
3	2.976	17	7	3.123	4	13	3.075	7	21	2.631	34
4	2.207	36	8	3.072	8	14	2.914	23	22	2.976	18
			9	3.028	11	15	2.594	35	23	2.889	24
31	2.965	20	10	3.078	6	16	2.708	32	24	2.970	19
						17	2.798	28	25	2.845	25
						18	2.658	33	26	2.838	26
									27	3.025	12
									28	2.978	16
									29	2.934	21
									30	2.803	27
Aver.	2.915		Aver.	3.090		Aver.	2.814		Aver.	2.895	
32	2.984	15	33	3.197	2	34	3.006	14	35	3.030	10
36	3.102	5									
Total average 2.916											

* The consensus forecast (31) and the average forecasts (32–36) are included in the ranking.

in use by meteorologists today. The probability score would assign the score

$$PS_k(\mathbf{r}) = (1-r_k)^2 + \sum_{i \neq k} r_i^2$$

to the forecast $\mathbf{r}$ and thus we would have

$$Q_k(\mathbf{r}) = 1 - 0.5\, PS_k(\mathbf{r}).$$

The quadratic scoring rule has the important property that it encourages the assessor to make his assessed distribution correspond to his true beliefs, i.e., if he wants to maximize his subjective expected score. The assessor's true judgment is expressed by $\mathbf{p} = (p_1, \cdots, p_n)$, and $\mathbf{r}$ need not necessarily be equal to $\mathbf{p}$. The expected score is then $\sum p_k Q_k(\mathbf{r})$ and this expression is maximized when $\mathbf{r} = \mathbf{p}$ (see, e.g., Winkler and Murphy, 1968b). Scoring rules with this property are said to be strictly proper. Some other examples of strictly proper scoring rules will be mentioned in Section 4.

d. Instruction

The participants received written instructions about the experiment and the assessment task. The instructions included a detailed description of the rule by which the assessments would be scored. It was stressed that there was nothing to be gained from assessing distributions that deviated from true judgments. This point was repeated a few times in letters that were sent out along with the results. There was no training session.

e. Feedback

The feedback was presented individually in the form of a sheet with four tables containing the following information.

1) The assessed probability distribution and the corresponding score were given for each of the five quantities, as was the average score over all sessions. It also showed the ranking, based on the average score, for each quantity.

2) Another table presented the average probability distribution and its score for each quantity. The average was taken over all participants.

3) The average score for each quantity was given in a third table, as well as the average over all sessions.

4) The fourth table presented the total score for each participant, his average score and his rankings after the preceding and after the present sessions. It also showed, for each participant group and for all participants, the results of the average forecast of the group. Finally, it included the average score for the actual session and over all sessions for each group and for all participants.

3. Main results

a. Introduction

The final results, presented in Table 1, indicate that it was difficult to make good forecasts of the kind required in the experiment. The scores can be compared with those that an uninitiated person would have obtained. He might have chosen to use climatological probabilities[4] for the five quantities and would then have received the average score 3.053. Only 7 out of 30 participants did better.[5] One of the main reasons for the low scores is the length of the forecasting period, which covered almost 6 days. It is well known that it is

[4] The climatological probabilities for the first three quantities were given in Section 2. They were estimated for the last two quantities from data for the same period as covered by the experiment during the 20 years 1949–68. This gave the distributions (0.216, 0.231, 0.247, 0.162, 0.109, 0.034) and (0.447, 0.381, 0.172), respectively.

[5] If he had assigned equal probabilities to each class he would always receive the score 3.000. Only 10 participants had a higher score.

TABLE 2. Number of times the true value fell into the different classes.

	Class							
Quantity	1	2	3	4	5	6	7	8
1	4	2	5	2	1	—	2	—
2	4	—	1	3	2	3	2	1
3	9	2	2	3				
4	5	3	3	2	2	1		
5	9	3	4					

difficult to make good forecasts for periods exceeding 24 hr, let along periods of 48 hr or 5 days.

It must, however, be kept in mind that the experiment was rather short and it is possible that some of the results can be considered short-term effects. It also seems clear that SMHI regarded the period covered by the experiment as rather difficult with more sudden changes in weather than is usual for that time of the year. Table 2 presents the number of times for each quantity that each event was the true one. The period seems to have been cooler and with less precipitation than usual.

b. *Comparison of groups of forecasters*

One purpose of the experiment was to study whether there were any differences in performance among the four groups of forecasters and to see whether their performance improved with time. To begin with the the second question there does not seem to be a clear tendency toward improvement or impairment. It is also difficult to make comparisons between the different sessions because some must be regarded as easier than others. It is true that the scores were high in four of the last five sessions, but this is not convincing evidence of an improvement since the 14th session produced a low score. In this instance, the participants could not forecast a sudden drop in temperature and a large amount of rain. The end of the experiment marked the beginning of a long, hot, dry period, in which it would have been easy to make good predictions. This means that the average scores might have increased rapidly if the experiment had continued for some time.

The research assistants came out best with all six among the eight best participants. The students were on the whole the least successful. How does it happen that the assistants did better than the meteorologists? A more detailed study of the meteorologists' records reveals great fluctuations in their scores. This was the result of rather tight forecasts. In case of favorable weather they sometimes received scores as high as 4.5, but more often the weather changed "unexpectedly," which resulted in poor scores. It is quite clear that the meteorologists would have benefitted (like any probability assessor) from more extensive feedback—as to the effect of the tightness of the distributions on the scores, for example. No group received any feedback beyond the list of results. The research assistants, however, provided this themselves through internal discussions. They compared their forecasts and the resulting scores. In this way they perhaps realized more rapidly than the other groups that it seldom paid off to have strong opinions (in this application).

It seems that the good performance of the research assistants as a group is one positive result of the experiment. The meteorologists would no doubt have received better scores if they had been able to take part in similar discussions.

c. *Comparison of predictands*

There seems to be a clear difference in the difficulty of the five different quantities as represented by the average scores. These are shown in Table 3 along with average scores for forecasts based on uniform distributions and for forecasts based on climatological probabilities. This means that the average score increased when the number of classes decreased, a natural consequence of the scoring rule. It may be more interesting to observe that the number of persons who did better than a forecaster assessing uniform distributions was 9, 3, 14, 13 and 16, respectively, for the five predictands. This number also increased when the number of classes decreased. The last three numbers would have decreased to 6, 10 and 14, respectively, if the comparison had been made with forecasts based on climatological probabilities rather than with uniform distributions.

It was assumed that it would be easier to make forecasts for average temperatures for the first 2 days of the period than for the last 2 days. The results seem to confirm this. There were, however, 10 participants who succeeded better with the latter forecasts.

There was some overlapping of 1–2 days between two adjacent sessions. This implies that the average temperature for "Friday–Saturday" was forecasted twice, first on a Monday as quantity 2 and then on a Thursday as quantity 1. This enables us to study whether additional information helped improve performance. The second forecast, related to the first 2 days of a period, had a higher average score in the first 5 weeks, whereas the contrary was true for the last 3 weeks. The average difference taken over all 8 weeks was 0.066 in favor of the second forecast.

TABLE 3. Average scores for different quantities.

	Quantity				
	1	2	3	4	5
Actual scores	0.546	0.519	0.614	0.570	0.667
Forecasts based on uniform distributions	0.563	0.563	0.625	0.583	0.667
Forecasts based on climatological probabilities	0.563	0.563	0.656	0.593	0.679

TABLE 4. Average scores for each quantity and event taken over those sessions where this event was true.

Events					Event						
Quantity	1–8	2–7	3–6	4–5	Quantity	1	2	3	4	5	6
1	0.625	0.501	0.542	0.518	3	0.660	0.658	0.617	0.462		
2	0.431	0.501	0.545	0.571	4	0.544	0.598	0.663	0.630	0.501	0.384
					5	0.698	0.682	0.602			

d. Comparison of events

Are certain events more difficult to predict? It somehow seems more difficult to predict a large amount of rain than a small amount. It might come as a surprise to find that there was less rain than expected but the surprise would probably be greater when there was a large amount of rain. Table 4 presents the average score for each quantity and event taken over all forecasts for sessions when this event was the true one. The classes for the average temperatures have been taken two by two, from the extreme classes toward the middle classes.

Extreme average temperatures seem to have been easy to predict for the first two days of the period. This could be related to the fact that in the four actual cases the temperature on the day of the forecast was well within the limits of the same extreme class. When average temperatures for the last two days are considered, we find that the more extreme the true event, the lower the score. These forecasts generally had their modes in the central classes and extreme classes therefore received low scores.

The first class for quantity 3 might have been expected to have had a higher score since this class by definition had the highest climatological probability. It seems, however, that in most sessions the second class was given the highest probability. This explains why this class appears to have been equally easy to predict. The fourth quantity generally had the forecasts centered around the third event (i.e., 2 days of rain). The fifth quantity was often centered around the second event. The first event (no wet day) was not always easy to predict and only three good scores at the end of the experiment make it look like the easiest event.

e. A check for consistency

There is one relation between the two different kinds of precipitation forecasts that can be checked for consistency. The first class for quantity 3 includes two cases, i.e., no precipitation at all and a positive but small amount of precipitation. The probability for this event cannot therefore be less than the probability for no wet day in the period (event 1 for quantity 4). This was violated in 16 out of 423 forecasts. The two probabilities were equal in 73 further forecasts, and in 35 of these cases the common value was greater than zero (implying that the probability of only a little precipitation was zero, which probably was not consistent with the forecaster's beliefs).

4. Different scoring rules

The quadratic scoring rule was used in the feedback and the preceding analyses have been based on it. There are many other rules which have a relatively simple structure and which are strictly proper. The following are two examples of such rules:

Spherical scoring rule: $S_k(\mathbf{r}) = r_k / \sqrt{\sum r_i^2}$

Logarithmic scoring rule: $L_k(\mathbf{r}) = \ln r_k$

In order to prevent the last scoring rule from taking infinitely small values, it has been suggested that a truncated version be used (Shuford *et al.*, 1966), i.e.,

$$L_k^*(\mathbf{r}) = \begin{cases} 1 - \ln r_k / \ln a, & r_k \geq a \\ 0, & r_k < a \end{cases}$$

It should be noted, however, that L^* is not a strictly proper scoring rule.

It was often remarked by the participants that the actual scoring rule did not always value assessments that were nearly right more than assessments that were far from right. Consider, for instance, the following two precipitation forecasts (0.5, 0.3, 0.1, 0.1) and (0.1, 0.3, 0.5, 0.1). The three scoring rules defined so far all give the two forecasts the same score if the fourth event turns out to be true. Most people, however, would probably regard the second forecast as better than the first one. Loosely speaking, the first forecast is more distant than the second forecast from the true event.

A scoring rule is said to be *sensitive to distance* if $S_k(\mathbf{r}) > S_k(\mathbf{r}')$, where $\mathbf{r}'$ is more distant than $\mathbf{r}$ from the true event. This requires a formal definition of what should be meant by the relation "more distant than." One definition, given by Staël von Holstein (1970a), says that $\mathbf{r}'$ is more distant than $\mathbf{r}$ if $\mathbf{r}' \neq \mathbf{r}$, and $\mathbf{r}'$ has at least the same probability mass as $\mathbf{r}$ in both tails (on either side of the true event) of the distribution. [A different definition is suggested by Murphy (1970).] The following strictly proper scoring rule, the ranked probability score (Epstein, 1969), was then shown to be sensitive to distance:

$$RPS_k(\mathbf{r}) = \frac{3}{2} - \frac{1}{2(n-1)} \sum_{i=1}^{n-1} \left[\left(\sum_{j=1}^{i} r_k \right)^2 + \left(\sum_{j=i+1}^{n} r_k \right)^2 \right] - \frac{1}{n-1} \sum_{i=1}^{n} |i-k| r_i.$$

TABLE 5. Average scores and rankings with four different scoring rules.

Person	Scoring rules: Quadratic Score	Quadratic Rank	Spherical Score	Spherical Rank	Logarithmic Score	Logarithmic Rank	RPS Score	RPS Rank
1	2.768	24	2.016	19	2.765	25	3.860	21
2	3.201	1	2.628	1	3.316	6	4.163	1
3	2.976	12	2.276	8	2.825	23	3.860	20
4	2.207	30	1.267	30	1.965	30	3.268	30
5	3.162	2	2.478	2	3.495	1	4.148	3
6	3.070	7	2.315	5	3.277	9	4.104	4
7	3.123	3	2.359	4	3.294	7	4.039	7
8	3.072	6	2.309	7	3.410	3	4.044	6
9	3.028	8	2.207	11	3.098	13	4.025	8
10	3.078	4	2.425	3	3.319	5	4.158	2
11	2.915	16	2.077	16	2.923	21	3.919	15
12	2.795	23	1.866	24	3.035	19	3.799	26
13	3.075	5	2.310	6	3.439	2	4.072	5
14	2.914	17	2.103	15	3.038	18	3.938	13
15	2.594	29	1.641	29	2.478	28	3.738	29
16	2.708	26	1.810	26	2.829	22	3.843	23
17	2.798	22	1.897	23	3.093	14	3.911	17
18	2.658	27	1.789	27	2.428	29	3.774	28
19	3.018	10	2.226	9	3.271	10	3.988	10
20	2.753	25	1.829	25	2.968	20	3.866	19
21	2.631	28	1.714	28	2.726	27	3.849	22
22	2.976	13	2.148	13	3.375	4	3.954	11
23	2.889	18	2.073	17	3.041	17	3.928	14
24	2.970	14	2.135	14	3.065	15	3.897	18
25	2.845	19	1.932	21	3.107	12	3.775	27
26	2.838	20	1.969	20	2.759	26	3.805	25
27	3.025	9	2.224	10	3.225	11	4.002	9
28	2.978	11	2.176	12	3.280	8	3.947	12
29	2.934	15	2.055	18	3.063	16	3.916	16
30	2.803	21	1.915	22	2.821	24	3.817	24

These three scoring rules were applied to the experimental data. The average scores for the 16 sessions are shown in Table 5 for each participant together with his ranking. The different scoring rules give essentially the same ranking of the assessors. The quadratic and spherical rules seem to be closest. The rank correlation coefficient varies between 0.818 (spherical and logarithmic scoring rules) and 0.959 (quadratic and spherical scoring rules).

Different scoring rules can, of course, produce different rankings of assessors when it comes to single assessments. There is, for instance, often a great difference between the quadratic score and the ranked probability score. These differences seem to cancel out, however, when average scores are compared over a number of assessments [for further evidence see Winkler and Murphy (1968a) and Staël von Holstein (1969)].

5. Some simple mechanical forecasting models

We shall now turn to some simple mechanical forecasting models. These all rely on some form of persistence, e.g., that the probability of rain for a certain day is dependent on whether it rained the previous day. The parameters of these models were estimated from data for the 20 years 1949-68 for the same period covered by the experiment.

a. Average temperatures

1) The limits of the eight classes have been defined to make the classes equally likely when average temperatures for 5 days were considered. The averages used in the experiment covered periods of 2 days which should give the extreme classes more weight. We can nevertheless study what happens when equal probabilities are assigned.

2) The relative frequencies for the eight classes as obtained from 320 observed periods ought to improve on the preceding model. The distribution thus obtained for the first 2 days was (0.175, 0.106, 0.122, 0.100, 0.084, 0.122, 0.103, 0.188) and for the last 2 days (0.100, 0.131, 0.103, 0.112, 0.075, 0.141, 0.100, 0.237). The goodness of fit in relation to the uniform distribution is poor. We can say that these two distributions represent the climatological probabilities.

3) One way of introducing dependence is to assume that the temperature is dependent on the average temperature for the preceding week. More precisely, it was assumed that the probability distribution for the average temperature for the first 2 (last 2) days of a period was dependent on the class of the average temperature during the period one week earlier.

The average scores for the three forecasting rules and the average scores of all participants are shown in Table 6. It is surprising to find that the equal probability model performed better than the other two.

b. Total amount of precipitation

1) The climatological probabilities are given by the definition of the class limits as (0.375, 0.250, 0.250, 0.125). The average score for this model was 0.656.

2) A simple persistence model makes the amount of precipitation for one week depend on the corresponding amount the preceding week. The average score was 0.665.

Both models did considerably better than the average of the participants' scores, 0.614.

c. Number of days with precipitation

It is well known that some dependence exists between the occurrence of precipitation one day and its occurrence the following day. We shall examine some simple models that take this into account. It is not unnatural that there are more examples of such models in the literature than there are models for the amount of

TABLE 6. Average scores for quantities 1 and 2 for three mechanical forecasting rules.

Quantity	Forecasting rule 1	2	3	Average score all participants
1	0.563	0.557	0.539	0.546
2	0.563	0.544	0.531	0.519

TABLE 7. Distribution of the number of wet days in 5-day periods, theoretical and observed.

Number of wet days	Theoretical distribution Probability	Frequency	Observed frequency
0	0.174	55.5	69
1	0.260	83.2	74
2	0.260	83.2	79
3	0.184	59.0	52
4	0.092	29.6	35
5	0.029	9.4	11
Total	0.999	319.9	320

precipitation. The amount represents a continuum of values (with the exception of a mass point at zero), whereas the occurrence of precipitation is a dichotomous variable.

1) The simplest form of dependence is given by a Markov chain of order one. The weather state (i.e., "rain" or "no rain") one day is then dependent on the state the preceding day but not on days earlier than that. We define $X_i=1(0)$ if day i is wet (dry). The transition probabilities are then given by $p_{jk}=P(X_{i+1}=k|X_i=j)$, $j,k=0,1$. The transition matrix $\mathbf{P}=\{p_{jk}\}$ was estimated to be

		Day $i+1$ Dry	Wet
Day i	Dry	0.725	0.275
	Wet	0.469	0.531

We want to forecast the number of days of rain in the next n days ($n=2$, 5) given the state on the day preceding the forecast. This means that we want to determine

$$\pi_{jk}=P(\sum_{i=1}^{n} X_i=k|X_{-1}=j), \quad j=0,1; \quad k=0,\cdots,n.$$

This distribution is naturally well determined when we know $\mathbf{P}$. A simple test was performed to study whether the assumption underlying this model could be justified. The probability of k days of rain can be written as

$$P(\sum X_i=k)=qP(\sum X_i=k|X_{-1}=0)+(1-q)P(\sum X_i=k|X_{-1}=1),$$

where $q=p_{10}/(p_{01}+p_{10})$ is the absolute probability that a day will be dry (Gabriel and Neumann, 1962). We can then compare the actual distribution over 320 five-day periods with the theoretical distribution as obtained from the estimate of $\mathbf{P}$. These distributions are shown in Table 7. The goodness of fit between the two distributions can be measured by the usual χ^2 statistic, whose value is 6.58 based on three degrees of freedom.

2) The probabilities $\{\pi_{jk}\}$ could also be estimated directly from the 20 years' data.

3) One way of introducing a longer dependence is to estimate the probabilities

$$P(\sum_{i=1}^{n} X_i=k|\sum_{i=-6}^{-2} X_i=j), \quad j=0,\cdots,5; \quad k=0,\cdots,n; \quad n=2,5.$$

That is, the number of wet days one week is dependent on the corresponding number the preceding week.

4) It is well known that the binomial distribution is not very good for describing the number of wet days. It was included, however, for a comparison with the Markov model. The distribution was based on $P(\text{wet day})=0.37$.

5) Finally, the probabilities

$$P(\sum_{i=1}^{n} X_i=k)$$

were estimated without including dependence on the states on earlier days. These probabilities were given in footnote 4.

The average score for the five forecasting rules and the average score of all participants are shown in Table 8. The binomial model did poorly as expected since it does not recognize the possibility of a long sequence of wet or dry days. Model 5 did better than expected for 5-day periods. It includes a dependence between days within the period but excludes any dependence on states on days before the period. This may also be true of model 3 because dependence on the number of wet days in the preceding period seems to be rather weak. The number of wet days in 2-day periods seems, on the contrary, to be best described by a model relating to the most recent observation, i.e., X_{-1}.

6. A mechanical forecasting model based on the forecaster's past performance

The meteorologists' results fluctuated a great deal. They performed either remarkably well or rather poorly. Their distributions were often quite tight, leading to extreme scores. The experiment (similar to other experience) shows that weather can change quite rapidly and that it is therefore not easily predictable

TABLE 8. Average scores for quantities 4 and 5 for five mechanical forecasting rules.

	Forecasting rule 1	2	3	4	5	Average score all participants
Quantity 4 (day 1–5)	0.585	0.588	0.596	0.555	0.593	0.570
Quantity 5 (day 1–2)	0.698	0.697	0.687	0.648	0.679	0.667

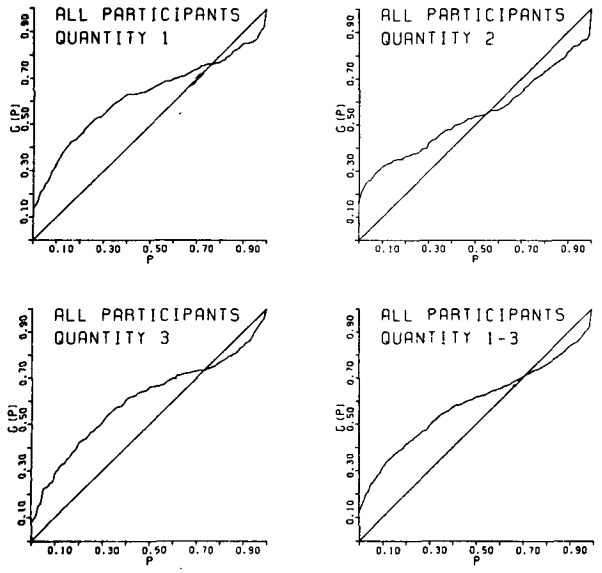

FIG. 1. Empirical distribution of observed fractiles.

for periods as long as 5 days. This means that there is often good cause for spreading the distributions. There are, of course, situations where it is relevant to assign a large probability to, for instance, "no precipitation" or to the extreme temperature classes.

Meteorologists' official forecasts are generally given as point estimates, i.e., the most probable class is given as the estimate. This class is presented in verbal form (e.g., "warmer than normal"). It is only natural that their point estimates cannot always be right but records show that they most often catch the tendency in the weather. If it is recognized that meteorologists make good point forecasts but that their distributions are too tight, then the following formal procedure might be a good basis for formulating the forecast.

Let the meteorologist pick the most probable class (the procedure is easily modified if he judges two or more classes equally probable). His forecasting record is then examined for all instances when he has chosen the same class. The relative frequencies for the occurrence of the different classes can then be calculated and used as a first approximation to a probability forecast. Starting from this distribution the meteorologist may then want to adjust the probabilities in some way. But he will be more aware, for instance, that temperatures can change rapidly and that he has been "caught" a number of times before.

Some historical data were obtained from SMHI. There were 391 observations on forecasts of average temperatures for 5-day periods for 1966–69 and 156 observations on forecasts of the total amount of precipitation for 5-day periods for 1968–69. These data are not directly adaptable to the forecasts made in the experiment. For one thing, there should be individual data for each meteorologist. Further, the period of observation includes the experiment. Finally, the data on forecasts of average temperatures relate to 5-day forecasts whereas the experiment concerned 2-day forecasts which can be expected to fluctuate more. Nevertheless, the data available can serve an illustrative purpose.

It seems most relevant to apply these relative frequencies to the meteorologists' consensus in each session since they are related to the meteorologists' past performance in official forecasts. The resulting average scores for the first three quantities turn out to be 0.622, 0.565 and 0.720, respectively. This is a substantial improvement on the consensus forecast which received the scores 0.594, 0.557 and 0.555. In addition, the sum for the three quantities, 1.907, is better than that of any participant.

There is apparently good cause for using these historical data when formulating a forecast. It must, however, be stressed that the resulting distributions should not be regarded as final forecasts. They are intended to serve but one purpose, i.e., to remind the forecaster that past experience has shown that actual outcomes can be far from the class he judges most probable. He could base his forecast on such a distribution but he must be prepared to let all relevant information influence his final decision.

7. External validity of forecasts

One implication of an assessed distribution F is that the forecaster should regard it as equally likely that the true value falls between the fractiles x_k and $x_{k+0.01}$, irrespective of k [x_k is defined by $F(x_k)=k$]. For example, it should be equally likely that the true value falls between the 0.02 and 0.03 fractiles as that it falls between the 0.53 and 0.54 fractiles. Furthermore, if the fractiles corresponding to the outcomes of a long series of assessments are determined, then the observed fractiles would be roughly uniformly distributed over the interval [0, 1].

Given the number of forecasts, an empirical fractile distribution G can be determined, where $G(p)=1/n$ times the number of forecasts with fractile not greater than p, for $0\leq p\leq 1$. This was done for each participant individually and for the total of all participants. Only the forecasts for the first three quantities were used, since the last two represented discrete variables for which it would be more difficult to determine the fractiles. The empirical distributions for all participants are found in Fig. 1, for each of the three quantities and for the aggregate.

In order to determine the fractiles it was assumed that the probability assigned to a class was uniformly distributed over its range of values. In the cases when the true value fell into one of the unbounded classes it was assumed that the forecaster would regard all possible fractiles as equally likely, and a fractile was determined with the help of a random number. That is, if the

average temperature fell into the first class and there was a probability p assigned to that class, then the fractile would be Zp, where Z would be a random number from the interval [0, 1]. Another estimation problem occurs for the first class of the third quantity. The assessed probability for this class can be regarded as consisting of two parts, the probability of no rain at all and a continuous distribution for small amounts of rain. It would have been advantageous to have the two parts separated. Assuming consistent assessments, the first probability was given by the probability of zero wet days (quantity 4). The remaining probability for the first class was then assumed to be uniformly distributed over the range of values of that class. All fractiles were rounded off to the nearest percent.

It is obvious that this way of determining the fractiles, especially the use of random numbers for the extreme classes, could be unfair to some participants. Fig. 1, however, represents the aggregate of all 423 forecasts and it should give a fair picture of the participants' assessments even though the forecasts are not quite independent. It should be noted that 25% of all forecasts fell below the 0.07 fractile and 25% above the 0.79 fractile. It seems evident that, in general, the distributions were much too tight. The forecasts were also biased toward too high values, i.e., the participants generally expected higher average temperatures and more rain than was actually observed. These results seem to hold not only when all forecasts are studied together but also for the individual participants.

Another definite indication that a distribution is too tight is when no probability is assigned to the "true" event. It should be a surprise to the forecaster that the "unbelievable" actually occurred. Nine participants had more than one such surprise per session, seemingly a very high number. Only two persons were completely free from surprises and seven more had an average of less than 0.5 surprise per session. Thirteen percent of all forecasts led to surprises. This number increased from 4% for quantity 5 to 23% for quantity 2.

The forecasts might have been validated better if the experiment had continued for some time and the participants had received feedback concerning the external validity of forecasts in earlier sessions.

8. Aggregation of forecasts

So far we have only been concerned with individual forecasts. In many practical situations forecasts will be made by several people. Five-day weather forecasts in Sweden, for instance, are generally made by two meteorologists. A forecast represents some consensus of their opinions, which, of course, may differ. The consensus can be reached in essentially two ways, one behavioristic and the other formal.

a. Behavioristic approach

The ordinary way of reaching a consensus is to have the assessors discuss the assessment task and bring out any evidence they think relevant. The discussion will hopefully lead to an agreement on a distribution which would represent a consensus. In each session the two acting meteorologists were asked to submit a consensus forecast together with their own forecasts. The consensus is represented by "participant" 31. It received an average score of 2.965 which is just a little lower than the score, 2.984, of the meteorologists' aggregated forecast. The consensus was better in 7 out of the 16 sessions.

It could be difficult to avoid some behavioristic problems with this approach, e.g., the composition of the group and the relative strengths within the group. There are different ways in which the consensus hopefully could be improved. Winkler (1968) presents a thorough discussion of some behavioristic methods as well as some formal methods.

b. Linear aggregation

The second approach requires the assessors to make individual forecasts in the form of subjective probability distributions. Their forecasts can then be aggregated in some fashion in order to form a consensus which can be used as a basis for a decision.

The simplest way of aggregating distributions is to take linear combinations of them. This requires that the assessors be assigned weights $\{w_i\}$ with $\sum w_i=1$. Let $r_i=(r_{i1},\cdots,r_{in})$ be the ith person's forecast. The aggregated distribution is then $r^*=(r_1^*,\cdots,r_n^*)$, where $r_k^*=\sum w_i r_{ik}$.

The difficulty in this method lies in the choice of an appropriate set of weights. This choice must be made by the person who will use the forecasts as a basis for a decision. He will assess weights which represent his subjective judgment of the "goodness" of the different assessors. If there is no information available on the experience or the capability of the assessors, then he would probably choose to assign them equal weights. It seems natural, however, that the various forecasts should carry unequal weights when there is some information available on the past performance of the assessors. The assignment of weights may be facilitated by some rule, provided that the weights thus obtained agree reasonably well with the aggregator's own judgment. Examples of such rules are found in Winkler (1967b, 1968) and Staël von Holstein (1969, 1970b). We shall here only discuss the following four simple rules:

1) Equal weights.
2) Weights proportional to average scores for previous sessions.
3) Weights proportional to $(m+1-R_i)$, where m is the number of assessors and R_i the rank of person i. (It is assumed that this ranking is based on the assessors' scores for previous sessions; the highest average score is given rank 1.)
4) Weights proportional to $1/R_i$.

TABLE 9. Average score for the consensus group.

	Weight system	Quantity 1	2	3	4	5	1–5
				Different weights			
	1	0.550	0.554	0.695	0.609	0.720	3.128
	2	0.552	0.553	0.696	0.609	0.720	3.131
	3	0.559	0.549	0.701	0.615	0.719	3.143
	4	0.562	0.550	0.694	0.613	0.720	3.139
	Q''	0.562	0.546	0.701	0.613	0.721	3.144
				Same weights			
	1	0.550	0.554	0.695	0.609	0.720	3.128
	2	0.551	0.554	0.695	0.609	0.720	3.129
	3	0.558	0.554	0.693	0.616	0.724	3.145
	4	0.555	0.558	0.694	0.616	0.725	3.148
	Q''	0.556	0.553	0.692	0.613	0.723	3.137
Average score consensus group		0.517	0.537	0.672	0.584	0.690	3.001

The list of results included the scores of the average forecasts of five different groups (i.e., Nos. 32–35 for the four groups of participants and No. 36 for the total of all participants). Table 1 shows that the scores for these average forecasts are always better then the average scores for the corresponding groups. This is a general property of the quadratic scoring rule (Staël von Holstein, 1970b) and was therefore known in advance. Another result is of greater interest, i.e., that the average forecasts are always in or near the top in their respective groups. Although all assistants were successful in relation to the other participants, the "average assistant" is found to do even better (on the average).

In this case all forecasts in a group were assigned equal weights. It seems reasonable that the scores could have been improved if unequal weights had been used.

The four aggregation rules mentioned above were studied for a consensus group of the 13 participants who took part in all 16 sessions. The aggregated distributions were only examined for the last ten sessions while the first six were used to give initial weights based on the average scores for those sessions. The group as a whole was a little better than average with an average score of 2.981 for all 16 sessions. In the last ten sessions they had an average of 3.001 as compared with 2.914 for all participants.

Two sets of aggregations were made, one with the same weights for all five quantities, the other with different weights for each of them. The average scores (quadratic scoring rule) for the last ten sessions are shown in Table 9. The aggregation based on weights proportional to average scores improves little on the equal weights consensus. This may seem somewhat surprising but the reason is that the relative differences between the scores are generally small and the scores themselves are therefore not directly suitable as weights. The last two weighting rules assign the greater part of the weight to those who succeeded best in the past.

We have arbitrarily defined the quadratic scoring rule to have the range [0, 1] but we could also have used, for instance, $Q''=Q-0.5$ (a scoring rule remains strictly proper under linear transformations). This would assign more weight to the assessors with the highest scores. The last line in each block in Table 9 shows the scores for aggregations with weights based on Q''. These scores are close to those based on the ranks.

It seems reasonable that separate sets of weights should be used for different quantities since different assessors may find some quantities more difficult than others. This is not confirmed by the experimental results, which may be due to the short period covered by the experiment. Roughly similar results were obtained when the logarithmic scoring rule or the ranked probability score was used instead of the quadratic rule.

9. Conclusions

The experiment has shown that it is feasible to ask people to express forecasts in the form of subjective probability distributions even when the quantity to be forecasted is non-dichotomous. Such forecasts should be more valuable to the consumer of the forecasts than categorical forecasts. Meteorologists (at least in the United States) are used to expressing forecasts for dichotomous events in terms of probabilities and in this way have become acquainted with probabilities and their meaning. It should not therefore present any greater conceptual problem to extend these forecasts to quantities expressed in terms of any number of events, even though very serious practical and operational problems also exist.

There seems, however, to be a tendency for forecasters to be overconfident in their assessments. It may be more obvious in situations like those in the present experiment where it apparently was difficult to make good forecasts. It will take some training to make forecasters produce good probabilistic forecasts. The training should consist not only of feedback with the true outcomes of the forecasting situations but also some evaluation of the forecasts. This feedback should also be accompanied by discussion related to the forecasts and the outcomes. The assistants in this experiment showed that discussion increased their understanding of the assessment task. Good training is time-consuming but it is necessary in order to obtain reliable forecasts.

This experiment made use of continuous quantities (the amount of rain included a discrete probability at zero) but the distributions were expressed in the form of probabilities for a set of mutually exclusive and totally exhaustive events. Another way of formulating assessments for such quantities could be to have the assessor assess a set of fractiles (e.g., median, quartiles, octiles, etc.) and then try to fit a distribution to these fractiles. For a discussion of a variety of such techniques see Winkler (1967a) and Staël von Holstein (1970b).

It is clear that much more must be learned about how people behave when confronted with probability assessment tasks. More experiments of this kind are

needed, preferably in immediate connection with the daily routine of making weather forecasts. Of course, forecasts such as these need not be restricted to weather characteristics like temperature and precipitation. Any quantity to be forecasted will serve the purpose of training the forecaster to make better probabilistic forecasts and at the same time provide data for research on assessment techniques.

There seems to be enough evidence to say that any "reasonable" scoring rule will lead to almost the same ranking of assessors when the assessors have made a large number of forecasts. Nevertheless, it may be essential to study the individual properties of the different scoring rules for single assessments, since the score will be used as feedback. If it does not seem reasonable to the forecaster then it will be quite useless for training purposes. Evaluation of probabilistic weather forecasts has most often been based on the quadratic scoring rule (or rather the equivalent probability score). Since this rule has some deficiencies, other rules may have better training effects. The ranked probability score seems to represent an improvement. Murphy (1970) makes a thorough comparison between the two rules and ends up with the same recommendation.

The assessor may sometimes be helped by some formal procedure like the ones discussed in Sections 5 and 6. There are undoubtedly even better "objective" models available than those mentioned in Section 5. But it should be kept in mind that such procedures could only serve as a basis for the formulation of the forecast. One should not dispose of the forecaster's personal judgment. The forecaster might add something, or to quote Sanders (1963, p. 192):

> The objective technique provides a probability reference point which the forecaster "sharpens" by critical appraisal with the use of additional information. In the author's experience there are no objective predictions which cannot be improved upon by the forecaster in this way, even when the objective method produces results which are superior to subjective forecasts made before its introduction.

Furthermore, no situation is identical with those required as input by this type of model. For instance, the simple Markov chain model in Section 5 requires all dry days as well as all wet days to be identical. Obviously this cannot be true, but even more refined models will have similar deficiencies.

REFERENCES

Allen, R. A., 1969: Operational evaluation of a ceiling and visibility prediction technique. Silver Spring, Md., ESSA, Weather Bureau, Rept. FAA-RD-70-17, 19 pp.

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

de Finetti, B., 1937: La prévision: Ses lois logiques, ses sources subjectives. *Ann. Inst. Poincaré*, **7**, 1–68 (see also English translation by Henry E. Kyburg, Jr.: Foresight: Its logical laws, its subjective sources. *Studies in Subjective Probability*, 1964, New York, Wiley, 93–158.)

——, 1962: Does it make sense to speak of "good probability appraisers"? *The Scientist Speculates: An Anthology of Partly-Baked Ideas*, London, Heinemann, 357–364.

Enger, I., J. A. Russo and E. L. Sorenson, 1964: A statistical approach to 2–7-hr prediction of ceiling and visibility, Vol. 1. Hartford, Conn., Travelers Research Center, Inc., Contract Cwb-10704, Tech. Rept. No. 2 (7411-118), 48 pp.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987.

Gabriel, K. R., and J. Neumann, 1962: A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quart. J. Roy. Meteor. Soc.*, **98**, 90–95.

Murphy, A. H., 1970: The ranked probability score and the probability score. *Mon. Wea. Rev.*, **98**, 917–924.

Russo, J. A., I. Enger and G. T. Merriman, 1966: A statistical approach to the 12-48 hr prediction of precipitation probability. Hartford, Conn., Travelers Research Center, Inc., Contract Cwb-11100, Final Rept. (7671–217), 107 pp.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191–201.

Savage, L. J., 1954: *The Foundations of Statistics*. New York, Wiley, 294 pp.

Shuford, E. H., A. Albert and H. E. Massengill, 1966: Admissible probability measurement procedures. *Psychometrika*, **31**, 125–145.

Staël von Holstein, C.-A. S., 1969: The assessment of discrete subjective probability distributions—An experimental study. Stockholm, University of Stockholm, Institute of Mathematical Statistics, Res. Rept. 41, 22 pp.

——, 1970a: A family of strictly proper scoring rules which are sensitive to distance. *J. Appl. Meteor.*, **9**, 360–364.

——, 1970b: *Assessment and Evaluation of Subjective Probability Distributions*. Stockholm, Economic Research Institute, 225 pp.

Thompson, J. C., 1950: A numerical method for forecasting rainfall in the Los Angeles area. *Mon. Wea. Rev.*, **78**, 113–124.

Winkler, R. L., 1967a: The assessment of prior distributions in Bayesian analysis. *J. Amer. Statist. Assoc.*, **62**, 776–800.

——, 1967b: The quantification of judgment: Some experimental results. *Proc. Amer. Statist. Assoc.*, 386–395.

——, 1968: The consensus of subjective probability distributions. *Management Sci.*, **15**, B61-B75.

——, and A. H. Murphy, 1968a: Evaluation of subjective precipitation probability forecasts. *Preprints of Papers, First Statistical Meteorological Conf.*, Hartford, Conn., Amer. Meteor. Soc., 148–157.

——, and ———, 1968b: "Good" probability assessors. *J. Appl. Meteor.*, **7**, 751–758.