On Subjective Probability Forecasting[1]

FREDERICK SANDERS

Massachusetts Institute of Technology

(Manuscript received 7 November 1962, in revised form 28 December 1962)

ABSTRACT

The subjective process of probability forecasting is analyzed. It is found to contain a *sorting* aspect, in which the forecaster distributes all instances into an ordered set of categories of likelihood of occurrence, and a *labeling* aspect, in which the forecaster assigns an anticipated relative frequency, or probability, of occurrence for each category. These two aspects are identified with the concepts of sharpness and validity, which have been introduced by other writers. The verification score proposed by Brier is shown to consist of the sum of measures of these two qualities. A satisfactory measure of synoptic skill is obtained by applying the Brier score to the synoptic probability forecast and to a control forecast of the climatological probability, and by expressing the difference as a percentage of the control score.

In an analysis of a large number of short-range probability forecasts made by instructors and students in the synoptic laboratory of the Massachusetts Institute of Technology it is found that even inexperienced forecasters are capable of displaying validity and skill except when dealing with events which occur very rarely or nearly always. Skill for average or net conditions over 24-hr periods is found to be roughly twice the skill in forecasts for a particular instant and is found to vary with the directness with which the weather element can be inferred from prognostic charts. The average of the judgment of two or more forecasters with comparable experience is found to be a more skillful statement than the forecast of the most skilled individual.

1. Introduction

The growing interest in probability forecasting has been tempered by doubts whether forecasters are capable of providing realistic statements of the likelihood of occurrence of meteorological events (e.g., Dexter, 1962). The purpose of this paper is to consider the nature of the subjective forecasting process within a probabilistic frame of reference, to provide evidence that skillful probability statements can indeed be formulated subjectively, and to describe a highly effective verification procedure. The results to be presented are derived from forecasts made as part of the synoptic laboratory program at M.I.T. from 1955 to 1962.

2. The subjective process of probability forecasting

Let us consider some aspects of the prediction process from the point of view of the forecaster who acknowledges the stochastic nature of his occupation. It should be recognized that a categorical forecast which is sometimes in error is in fact a restricted form of probability forecast. Suppose for example that "no rain" forecasts are correct 95 per cent of the time while 60 per cent of rain forecasts are hits. In effect the "categorical" forecaster has sorted all instances into two categories of likelihood of rain, five per cent and 60 per cent, respectively. But why restrict the forecaster to implicit use of these two probabilities? Perhaps he can subdivide these into a larger ordered set of probability categories, thus enhancing the usefulness of his advice.

Consider now the elusive concept of forecast skill. Skill must be measured in relation to something. From the viewpoint of the synoptic meteorologist a logical control is the climatological expectancy. Then the forecasters' skill lies in his ability to recognize factors in his array of synoptic information which, to him, make the likelihood of occurrence of meteorological events in a particular instance different from the climatological likelihood. If such skill cannot be demonstrated then the synoptic information is of no avail, the user of meteorological advice may reasonably inquire whether the maintenance of synoptic data networks, and of forecasters, is justifiable.

It has been pointed out (Gringorten, 1958) that such a measure of skill may be unrelated to the effectiveness of the forecasts in making a particular operational decision. In principle, however, a forecast statement may be applied to a variety of operations with different probability thresholds for decision. Then the skill in relation to climatology tends to become a reliable measure of the overall economic utility of the synoptic forecasts.

[1] This research was supported in part by Air Force Cambridge Research Laboratories under Contract Nos. AF 19(604)-5491 and AF 19(604)-8373.

What is the relationship between subjective and objective forecasts? Too often there is a grimly isolated competition to prove which is superior, a situation which the author considers detrimental to the effectiveness of forecasting. If the objective technique wins it may be regarded as a "forecast method" and used first grudgingly and finally uncritically to supplant human judgment. If the objective technique is inferior it may be denoted a "forecast aid" and largely forgotten.

A healthier state of affairs, it would seem, is based on the premise that a forecast is a fallible judgment which can use all the objectively processed help it can get. 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. Because of the flexibility and versatility of the human mind this situation seems likely to continue in the forseeable future. The important point is that the subjectively modified objective result is the best product.

The amount of subjective improvement obtainable varies with the individual forecaster and with the objective technique. In a particular instance the small amount of improvement which the human can provide may not be worth the expense of obtaining it, but a judgment on this matter can hardly be formed without knowledge of the sensitivity of the operational decision to the quality of the forecast.

3. The verification procedure and its effect on forecast strategy

Brier and Allen (1951) have suggested that the verification score introduced by the former (1950) for probability forecasts could not be "played." That is, there was no strategy harmful to the utility of the forecasts which could be employed to improve the score. Experience in the Massachusetts Institute of Technology synoptic program fully supports this contention and indicates that use of this score encourages the forecaster to display the maximum skill which he possesses.

The verification score, expressed as an average over N forecasts, is

$$F=1/N\sum_{i=1}^{N}(f_i-O_i)^2, \tag{1}$$

where f is the forecast probability, and O, the "observed" probability, is assigned the value one if the event occurs and zero if it does not. Perfectly correct and completely confident forecasting would score zero; the nadir of unskillful performance would score one. (We have chosen to consider each individual probability statement as a separate forecast. Other authors have referred to the set of probability statements associated with a particular element and instance as a single forecast.)

We have required that all probabilities be expressed in tenths. Each forecast thus falls in one of eleven categories ranging from zero to ten tenths. Now the above set of N forecasts can be partitioned into eleven subsets in each of which there is a particular value of forecast probability, f_k. The score is then expressed by

$$F=1/N\sum_{k=1}^{11}\sum_{i=1}^{M}(f_k-O_{ki})^2, \tag{2}$$

where i now refers to summation over the M forecasts in the subset for the forecast probability f_k.

Now we shall examine the strategic problem facing the forecaster as he attempts to minimize his score. Consider one subset and let $O_{ki}=\bar{O}_k+O_{ki}'$ where the bar is an average over the M forecasts in the subset. Then the score for the k-th subset is given by

$$F_k=1/M\sum_{i=1}^{M}(f_k-O_{ki})^2=(f_k-\bar{O}_k)^2+\overline{O_k'^2} \tag{3}$$

or

$$F_k=(f_k-\bar{O}_k)^2+\bar{O}_k(1-\bar{O}_k)$$

since O_{ki} can have the values one and zero only. The first term on the right-hand side of (3) is a measure of the "validity" of the forecasts, to use Miller's (1962) term. To minimize his score, the forecaster must put a realistic label on this subset. That is, the forecast probability, f_k, must correspond as closely as possible to the relative frequency of occurrence of the event, $\bar{O}_k$. Validity, as pointed out by numerous writers (e.g., Brier, 1957), is important for the operational effectiveness of the forecasts.

The second term on the right-hand side of (3) is a measure of the "sharpness" of the forecasts, to use a term introduced by Bross (1953). Its value is zero only when $\bar{O}$ is one or zero, that is, when all instances have been sorted into two categories in one of which the event always occurs and in the other of which the event never occurs. The maximum value occurs when $\bar{O}$ is 0.5. The value of this term depends only on $\bar{O}_k$. But the score for the whole set of N forecasts is the weighted average of the scores obtained in the various subsets:

$$F=1/N\sum_{k=1}^{11}M_kF_k=1/N\sum_{k=1}^{11}M_k(f_k-\bar{O}_k)^2 +1/N\sum_{k=1}^{11}M_k\bar{O}_k(1-\bar{O}_k)^2. \tag{4}$$

Therefore the forecaster can minimize the sharpness contribution to his overall score only by recognizing nearly certain instances as often as possible, so that many cases fall in categories in which $\bar{O}_k$ is near one

or zero and few in categories in which $\bar{O}_k$ is near 0.5. He cannot accomplish his aim, however, merely by assigning near-certain values of forecast probability indiscriminately in defiance of his own better judgment, for it is then likely that $f_k - \bar{O}_k$ will have large positive values when he asserts high confidence that the event will occur and large negative values when he claims that the event will almost certainly not occur. That is, he will be hurt by overconfidence. We shall see that the state of the art largely determines the distribution of cases among the various categories of f (or $\bar{O}$), though some forecasters are more perceptive than others in recognizing instances of relative certainty.

The Brier score can be readily interpreted when the climatological expectancy is used as a control. Let r be the climatological relative frequency of occurrence and define predicted departure of probability and observed departure of relative frequency as

$$d \equiv f - r, \qquad E \equiv O - r.$$

Then the score, expressed as an average over N forecasts, is

$$F = 1/N \sum_{i=1}^{N} (d_i - E_i)^2. \tag{5}$$

When the predicted and climatological probabilities are expressed in tenths each forecast falls in one of 21 possible categories ranging from $d = -1.0$ to $d = +1.0$. Only eleven of these categories will occur for any single forecast event since the extreme values of d range from $1.0 - r$ to $0.0 - r$. The extreme categories of d will rarely be used since, for example, $d = +1.0$ is a statement that a very rare event will almost certainly occur.

In terms of probability categories, the score can be written

$$F = 1/N \sum_{k=1}^{21} \sum_{i=1}^{M} (d_k - E_{ki})^2, \tag{6}$$

where index i now refers to summation over the M members of the subset in which the forecast probability departure is d_k. The score for the forecasts in the k-th subset is

$$F_k = 1/M \sum_{i=1}^{M} (d_k - E_{ki})^2. \tag{7}$$

The climatological control score for these forecasts, obtained by setting $d_k = 0$, is

$$C_k = 1/M \sum_{i=1}^{M} E_{ki}^2. \tag{8}$$

If we let $E_{ki} = \bar{E}_k + E_{ki}'$, then the amount of improvement over climatology shown by the forecasts in this subset is

$$C_k - F_k = \bar{E}_k^2 - (d_k - \bar{E}_k)^2. \tag{9}$$

Here $d_k - \bar{E}_k$ is a measure of validity and $\bar{E}_k^2$ a measure of sharpness of the probability forecasts.

The amount of improvement over climatology shown by the entire set of N forecasts is

$$C - F = 1/N \sum_{k=1}^{21} M_k (C_k - F_k) = 1/N \sum_{k=1}^{21} M_k \bar{E}_k^2 - 1/N \sum_{k=1}^{21} M_k (d_k - \bar{E}_k)^2.$$

TABLE 1. Values of $C_k - F_k$, the improvement over climatology, in hundredths, for a subset in which the forecast departure from climatological probability is d_k and the observed departure of relative frequency is $\bar{E}_k$.

d_k \ $\bar{E}_k$	−1.0	−0.9	−0.8	−0.7	−0.6	−0.5	−0.4	−0.3	−0.2	−0.1	0.0	+0.1	+0.2	+0.3	+0.4	+0.5	+0.6	+0.7	+0.8	+0.9	+1.0
−1.0	100	80	60	40	20	0	−20	−40	−60	−80	−100										
−0.9	99	81	63	45	27	9	−9	−27	−45	−63	−81	−99									
−0.8	96	80	64	48	32	16	0	−16	−32	−48	−64	−80	−96								
−0.7	91	77	63	49	35	21	7	−7	−21	−35	−49	−63	−77	−91							
−0.6	84	72	60	48	36	24	12	0	−12	−24	−36	−48	−60	−72	−84						
−0.5	75	65	55	45	35	25	15	5	−5	−15	−25	−35	−45	−55	−65	−75					
−0.4	64	56	48	40	32	24	16	8	0	−8	−16	−24	−32	−40	−48	−56	−64				
−0.3	51	45	39	33	27	21	15	9	3	−3	−9	−15	−21	−27	−33	−39	−45	−51			
−0.2	36	32	28	24	20	16	12	8	4	0	−4	−8	−12	−16	−20	−24	−28	−32	−36		
−0.1	19	17	15	13	11	9	7	5	3	1	−1	−3	−5	−7	−9	−11	−13	−15	−17	−19	
0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
+0.1		−19	−17	−15	−13	−11	−9	−7	−5	−3	−1	1	3	5	7	9	11	13	15	17	19
+0.2			−36	−32	−28	−24	−20	−16	−12	−8	−4	0	4	8	12	16	20	24	28	32	36
+0.3				−51	−45	−39	−33	−27	−21	−15	−9	−3	3	9	15	21	27	33	39	45	51
+0.4					−64	−56	−48	−40	−32	−24	−16	−8	0	8	16	24	32	40	48	56	64
+0.5						−75	−65	−55	−45	−35	−25	−15	−5	5	15	25	35	45	55	65	75
+0.6							−84	−72	−60	−48	−36	−24	−12	0	12	24	36	48	60	72	84
+0.7								−91	−77	−63	−49	−35	−21	−7	7	21	35	49	63	77	91
+0.8									−96	−80	−64	−48	−32	−16	0	16	32	48	64	80	96
+0.9										−99	−81	−63	−45	−27	−9	9	27	45	63	81	99
+1.0											−100	−80	−60	−40	−20	0	20	40	60	80	100

Values of $C_k - F_k$ as a function of d_k and $\bar{E}_k$ are given in Table 1. The greatest gains are available when the forecaster recognizes categories of instances in which the relative frequencies of occurrence of events are vastly different from their climatological expectancies. This way of rewarding skill in the use of synoptic information seems entirely reasonable. Once the forecaster has distinguished a category which he denotes qualitatively, say, "much more probable than the climatological likelihood," then his gain is maximized when the forecast probability departure coincides in fact with the departure of relative frequency of occurrence. He may, however, be undermined by overconfidence. If he recognizes a category in which the relative frequency in fact exceeds the climatological expectancy by two-tenths but chooses to label it with a forecast departure of five-tenths or more he will show negative skill because of the lack of validity of these predictions. Negative skill, moreover, always results when the signs of the predicted and observed departures are opposed. Notice, however, that the forecaster never shows negative skill when he is underconfident in relation to climatology. For example, suppose he has recognized a category of instances in which $\bar{E}_k$ is actually $+0.5$. He will always show some positive skill so long as he assigns to it a probability which represents a positive departure smaller than this value. If the forecasters' personal strategy is designed to maximize the likelihood of showing some positive skill rather than to maximize the amount of skill shown, then his predictions will display underconfidence. In our experience, however, use of large departures from climatological likelihood is encouraged by competition among individuals and by a natural human tendency to claim a degree of certainty which outstrips what is justified by the present state of knowledge.

Use of climatological control has been criticized on the ground that it discourages prediction of large departures from the norm. In the context of probability, however, it is difficult to see why the forecasting of large departures from climatological expectancy is desirable unless borne out by subsequent events.

4. The M.I.T. forecasts

As a part of the M.I.T. synoptic laboratory program a large number of probability forecasts have been made since 1955, by both students and experienced instructors. An analysis of the results, mostly of the instructors' forecasts, has yielded information on various aspects of subjective probability forecasting. The forecasts were the probabilities of occurrence, to the nearest tenth, of a wide variety of meteorological events. The range of the predictions was mainly 24 hr but it varied from 2 to 72 hr. Forecasts of instantaneous surface conditions were concerned with wind direction and speed, total cloud amount, ceiling, visibility, occurrence of precipitation, type of precipitation, and temperature. Forecasts also dealt with precipitation amounts, occurrence of thunder, and temperature change over specified time intervals. Air Force reconnaissance reports were used to verify flight-level probability forecasts of cloudiness, icing and turbulence. Most forecasts were for locations in the United States, but some referred to the Atlantic and Pacific Oceans, Europe, North Africa and Asia. Forecasts based on current data were made mostly in late winter and early spring but forecasts for other seasons, based on past data, were included.

Whenever possible the forecast probabilities referred to only the affirmative part of a dichotomous question, for example, "Will precipitation be falling 24 hours subsequent to the initial time?" or "Will the westerly component of the wind be positive?" In other instances, the forecasts referred to several categories, for example, of ceiling, visibility and temperature. Conventional synoptic analyses and prognoses both subjective and numerical, were used as guidance along with various available statistical and synoptic climatological summaries.

5. Validity and sharpness of the forecasts

The validity of the instructors' predictions for the 1955–1956 seasons is illustrated by Fig. 1. The close agreement between forecast probability and corresponding relative frequency of occurrence near the extremes is not as impressive as it seems, for these predictions referred to events which rarely, or nearly always, occurred. Similar data, expressed as departures from climatological expectancy, are plotted in Fig. 2. A tendency toward overconfidence is clearly evident, since the magnitudes of the departures of relative frequency of occurrence are not as large as the corresponding forecast departures of probability. The values of the climatological probability were not known at the time these forecasts were made, and this circumstance, coupled with the forecaster's somewhat enthusiastic estimate of his own powers, is likely responsible for the lack of "calibration." In all subsequent forecasts an estimate of the climatological probability was available as guidance.

A similar comparison for some 1961 predictions, shown in Fig. 3, indicates a considerable sobering of viewpoint to the extent that the forecasts, while more nearly valid, are now slightly underconfident. Perhaps the psychological position of the instructor leads him toward a somewhat timid strategy, designed to avoid large errors and to maximize the likelihood of beating the climatological control. The instructor forecasts for Spring 1962, illustrated in Fig. 4, show nearly perfect validity, except for extreme values of probability departure, which were rarely used.

Very well, but is long experience in probability forecasting necessary for approximate validity? Evidently not, to judge from an analysis of a small sample of forecasts made in the fall of 1961 (Fig. 5). These pre-

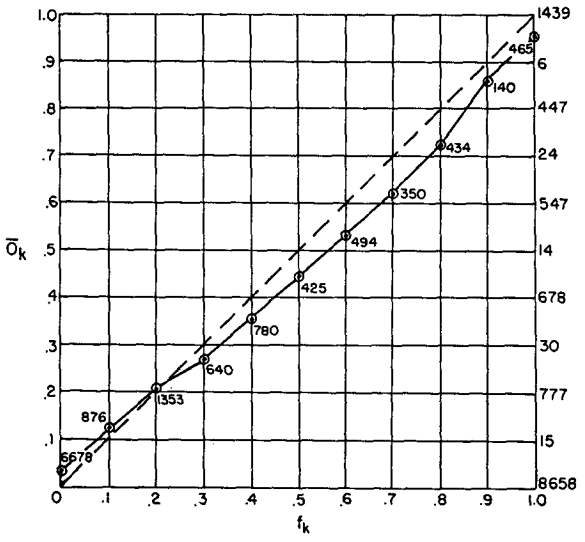

FIG. 1. Forecast probability, f_k, vs. observed relative frequency of occurrence, O_k, for instructors' forecasts in 1955–1956 seasons.

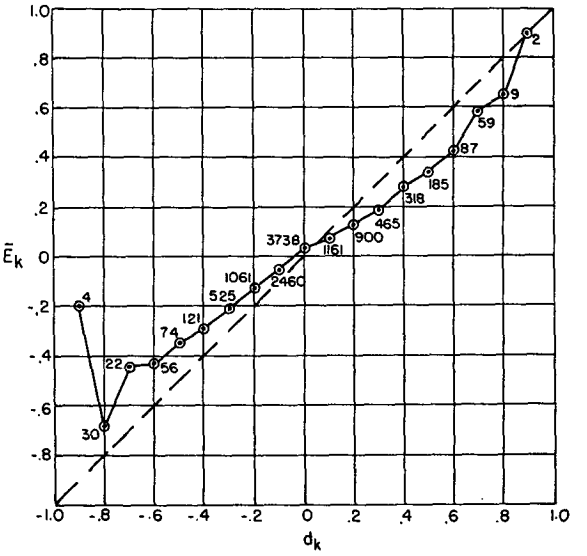

FIG. 2. Departure of forecast probability from climatological expectancy, d_k, vs. observed departure of relative frequency of occurrence, $\bar{E}_k$, for instructors' forecasts in 1955–1956 seasons. Number of forecasts of each departure value is shown next to corresponding data point.

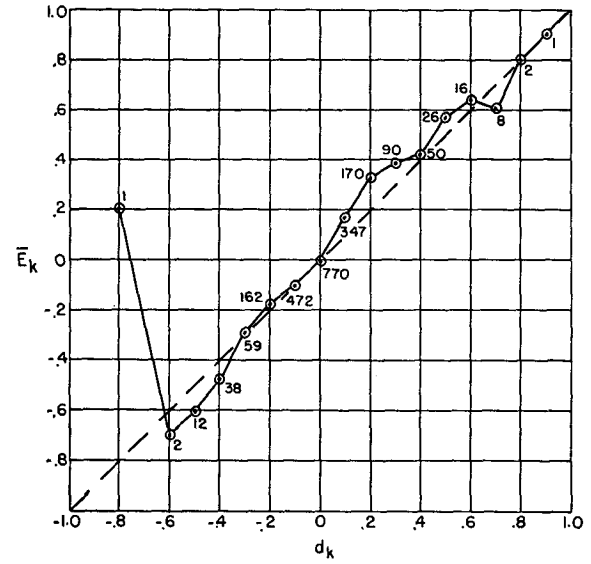

FIG. 3. d_k vs. $\bar{E}_k$, for instructors' forecasts, Spring 1961. Number of forecasts of each departure value is shown next to corresponding data point.

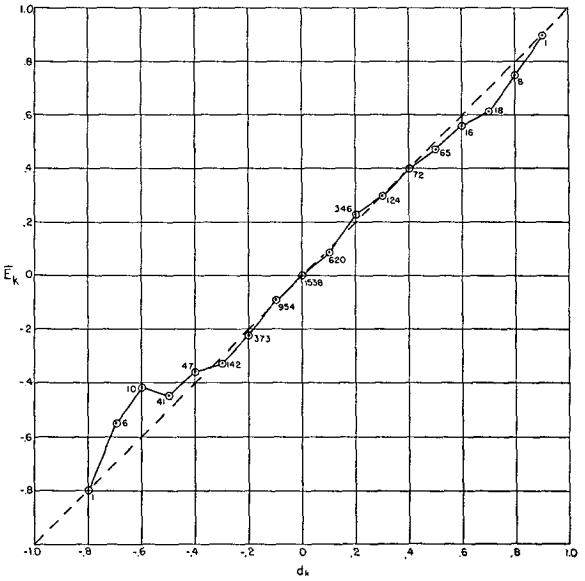

FIG. 4. d_k vs. $\bar{E}_k$, for instructors' forecasts, Spring 1962. Number of forecasts of each departure value is shown next to corresponding data point.

dictions referred to precipitation amounts and temperature changes over 24-hr periods extending 72 hr after the initial time. Forecaster A was the instructor, B had lengthy experience in categorical prediction but not in probabilistic forecasting, C had limited categorical experience, and the rest had no forecasting experience whatever and only a few week's exposure to synoptic data and analyses. Forecasters C through I conferred among themselves but generally not with A and B. All individuals show an ability to make meaningful probability statements. No pronounced tendency toward overconfidence or underconfidence is apparent except for the latter characteristic in the instructor's forecasts. The students tended systematically to overforecast the probability of occurrence of the event, possibly because of previous academic experience in which the result of a laboratory experiment is usually positive.

Validity in itself does not indicate skill in the synoptic forecasts, since a forecast of the climatological probability would be highly valid in the long run. The effec-

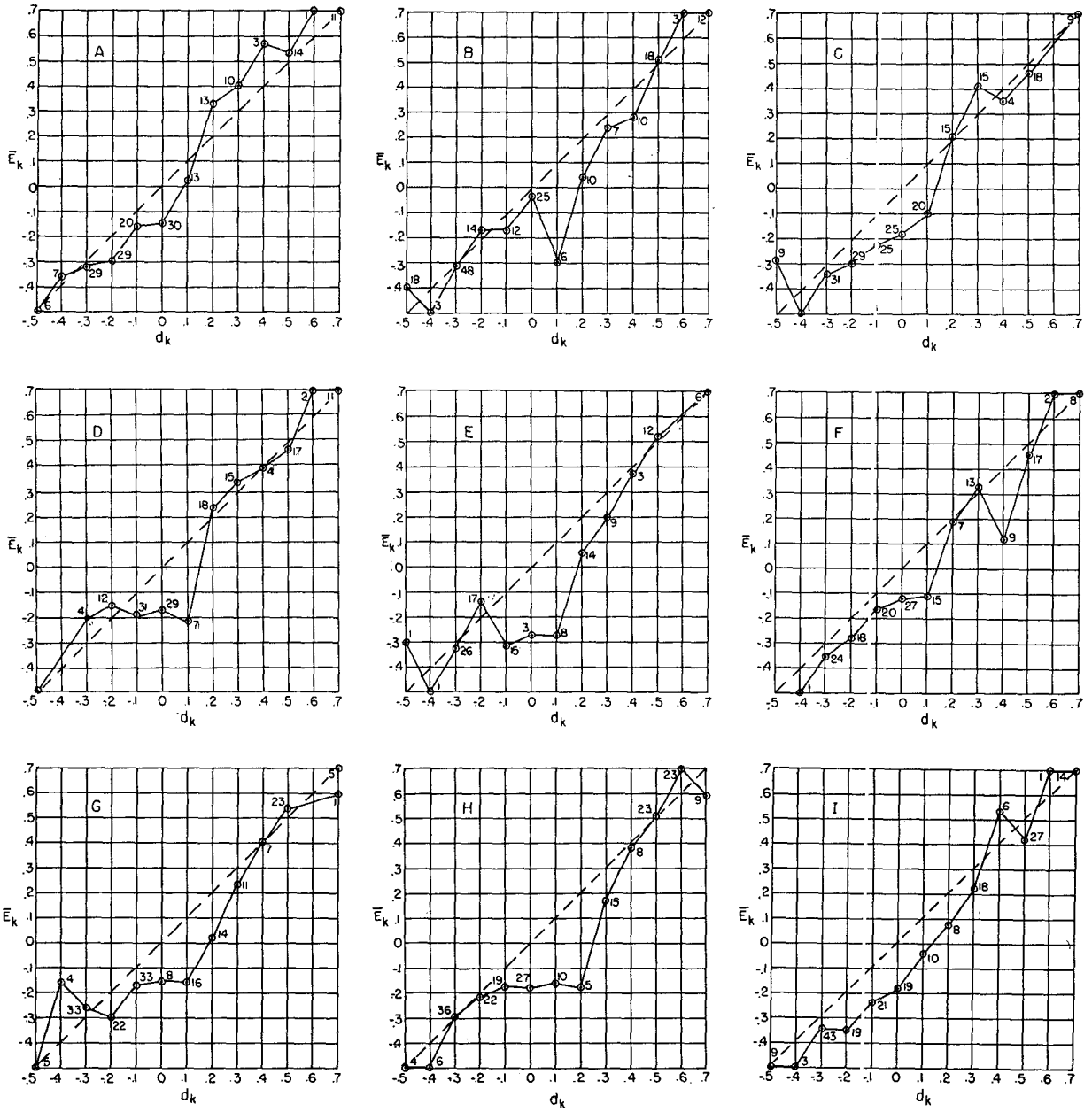

FIG. 5. d_k vs. $\bar{E}_k$, for nine forecasters, Fall 1961.

tiveness of the forecasts in contrast to the climatological control can be assessed by calculating the gain due to sharpness and the penalty due to lack of validity. These quantities are listed in Table 2.

The amount of gain over climatology due to sharpness is highly variable. Its general level is influenced by the climatological likelihood of occurrence of the events to which the forecasts refer, by whether the forecast events refer to a particular instant or to some specific time interval, and by the sensitivity of occurrence of these events to the synoptic patterns which can be brought into focus by existing data networks and techniques of analysis and prognosis. We shall discuss these matters later. It suffices here to observe that the sharpness gain is less variable when expressed as a percentage of the climatological control score. The penalty for lack of validity is small except in the earliest series, for which climatological probabilities were not available at the time the forecasts were made, and in the short series in the fall of 1961, in which the sample size was probably too small to permit a stable estimate of the validity to be made. Validity of forecasts may be somewhat more important, however, in evaluating the differences among forecasters. The performance of the

TABLE 2. Gain over climatology due to sharpness and penalty due to lack of validity, for instructors' forecasts. Values in parentheses are the gains and penalties expressed as percentages of the corresponding climatological control score.

Series	N	$(1/N)\sum_{k=1}^{21} M_k\bar{E}_k^2$	$(1/N)\sum_{k=1}^{21} M_k(d_k-\bar{E}_k)^2$
1955–6	11,277	0.0198 (21.5%)	0.0048 (7.1%)
Spring 1961	2224	0.0453 (26.4%)	0.0031 (1.8%)
Spring 1962	4382	0.0316 (22.7%)	0.0004 (0.3%)
Fall 1961	186	0.1235 (55.8%)	0.0084 (3.8%)

nine forecasters in the fall 1961 series is illustrated in Table 3. The differences among individuals in the amount of gain over the control reside mainly in variations in *sorting* ability (sharpness), but the effect of variations in *labeling* ability (validity) is by no means negligible in this sample. It must be conceded that the sample size is rather small, however, and that in a larger group of forecasts we would expect the validity penalty to decrease. Therefore firm conclusions as to the relative importance of these two aspects of forecasting in the comparison of individual abilities must await analysis of a larger sample of forecasts.

The gain over climatology due to sharpness shown in Table 2 is much smaller for 1955–1956 than for the other series. Note, however, that this gain did not represent a correspondingly small percentage improvement over the climatological score. In these years many of the forecasts referred to events with very small climatological expectancies, for which a near-zero probability was almost always forecast. Thus the overall level of scores was small, as was the absolute amount of skill shown by the synoptic forecasts. It appears that the skill of the forecasts is more adequately measured for our purposes by the percentage improvement over climatology than by the absolute amount of gain.

6. Variations of forecast skill

Unfortunately, the high level of skill shown in the Fall 1961 series given in Table 2 did not represent a leap to a new plateau of forecast effectiveness. These

TABLE 3. Gain over climatology due to sharpness and penalty due to lack of validity, for individual forecasters in the Fall 1961 Series.

Forecaster	$(1/N)\sum_{k=1}^{21} M_k\bar{E}_k^2$	$(1/N)\sum_{k=1}^{21} M_k(d_k-\bar{E}_k)^2$
A	0.1235	0.0084
B	0.1257	0.0104
C	0.1093	0.0150
D	0.1057	0.0120
E	0.1106	0.0278
F	0.0983	0.0129
G	0.0885	0.0144
H	0.0915	0.0155
I	0.1250	0.0116

predictions referred to precipitation amounts accumulated and temperature changes occurring over successive 24-hr periods, while virtually all the others were "spot" forecasts referring to conditions at a specific observation time. Small-scale variability which defies synoptic analysis and prognosis thus appears to reduce by a factor of about two the skill attainable in the prediction of 24-hr mean or net conditions.

The percentage improvement attained over climatology depends, of course, upon the geographical area for which the forecasts are made. Comparable spot forecasts made in 1961 and 1962 from past data are evaluated on this basis in Table 4. Ironically, the best performance is for the Weather Ships, part of whose reason for being is sparsity of data in the North Atlantic. This shortcoming notwithstanding, the strongly defined circulation patterns and the high degree of representativeness of observations in this region permit surprisingly skillful forecasts to be made. The generally low skill shown in the Asian forecasts is attributable to the tendency of the weather to run true to form (i.e., climatological probabilities close to certainty) during the monsoon seasons, and to the low representativeness of many of the observations, particularly of wind. Tropical storm activity affected a number of the forecast stations in the summer predictions made in 1962, in which considerable skill over climatology was demonstrated.

TABLE 4. Percentage improvement over climatology, 100X (C−F)/C, in forecasts for various geographical regions.

Area	Instructor 1961	Instructor 1962	Student range 1961	Student range 1962
North Atlantic Weather Ships (Dec. Jan.)	28.4	31.6	23.7–10.0	33.7– 21.3
Europe, North Africa, Near East (March)	25.9	27.6	18.5– 9.3	27.8– 14.9
Asia (Dec.)	21.3	5.8	18.4– 5.5	6.6––1.1
Asia (July)	11.4	27.5	11.5– 2.0	29.3– 17.8

The variation of skill in predictions for various weather elements is illustrated by Table 5 in which some of the instructors' spot forecasts are analyzed. The results show a logical relationship to the directness with which forecast elements can be obtained from prognostic charts. Thus, considerable skill is shown in prediction of surface wind direction, which can be obtained readily from the forecast horizontal pressure gradient. The lesser skill for wind speeds suggests that the general configuration of the pressure field is predicted more adequately than the spacing of isobars. Forecasts related to temperature display similar skill. These were based mainly on forecasts of thickness of isobaric layers, or vertical pressure gradient. The high skill shown in prediction of snow vs. rain may be unrepresentative because of small sample size, but is

TABLE 5. Percentage improvement over climatology, 100X (C−F)/C, for instructors' forecasts for various weather elements.

Weather element		Spring 1961	Spring 1962
Probability of:	Total cloud amount >4/8	18.6	19.8
	Precip. occurring at obs. time or within preceding hour	15.1	15.7
	Precip., if occurring, in frozen form	36.0	31.8
	Westerly wind component >0	33.8	24.7
	Southerly wind component >0	34.7	35.6
	Wind speed>critical value	22.2	15.7
	Temperature change>critical value	21.8	35.4

TABLE 6. Percentage improvement over climatology as a function of r, the climatological likelihood of the event, for the instructors' forecasts.

r	1955–1956 N	1955–1956 100X (C-F)/C	Spring 1961 N	Spring 1961 100X (C-F)/C
0.0	3270	−28.0	201	−4.1
0.1	2988	8.1	563	16.2
0.2	1990	13.1	476	19.8
0.3	1360	17.0	383	22.6
0.4	236	28.0	191	21.8
0.5	154	22.8	235	32.7
0.6	442	5.3	112	44.9
0.7	354	18.4	45	43.9
0.8	347	48.2	38	29.8
0.9	136	−26.0	15	17.9

attributable mainly to the skill of the thickness forecasts. Least skill is shown in forecasts of cloud and precipitation. Evidently the inferring of these elements from prognostic charts, even with the aid of initial moisture fields and initial and prognostic vertical motions, is still a weak link in the prognostic chain.

The forecaster has little or no skill in making probability statements about events which occur very rarely, or about events which occur almost always. This finding is evident in an analysis in which the percentage improvement over climatology is evaluated as a function of the climatological probability of occurrence of the event. The results, for the instructor forecasts for the 1955–1956 and Spring 1961 Series, are given in Table 6. The climatological expectancies, like the forecast probabilities, were expressed to the nearest tenth, so that the events for which $r=0$ actually occurred not more than five per cent of the time. Negative skill is shown in both series for $r=.0$ and for the 1955–1956 series when $r=.9$. Substantial skill is indicated through the middle ranges, for which the synoptic information offers abundant illustration of the conditions for occurrence and non-occurrence of the event. The tendency in both series for optimum performance to occur for events with climatological expectancy above 50 per cent is not readily explained but is probably attributable to the types of weather elements involved in these forecasts.

Some information is available concerning the decay of forecast skill with time. The percentage improvement over climatology for various forecasts is plotted in Fig. 6 as a function of number of hours subsequent to initial time. The values pertaining to the 24-hr period forecasts were entered at the mid-time of the interval. The 100 per cent value plotted at the initial time for the spot forecasts implies, of course, that initial conditions were perfectly known at the forecast locations. The skill shown in the period forecasts for the first 24 hr is spuriously large, since at the time these forecasts were made some information was available as late as 18 hr after the initial time. In any case, reasonable extrapolation of the data suggests that skill in the period forecasts is substantially larger than in the spot forecasts, as previously mentioned, and that a fair degree of skill could be shown at ranges substantially beyond those for which the forecasts were made. It is interesting to note that the skill for forecasts referring to precipitation amount in 12-hr periods in the San Francisco Bay area, reported by Root (1962), tends to fall between the skill shown by our spot forecasts and by our forecasts for 24-hr periods (see his Fig. 5).

7. Differences of opinion and consensus forecasting

How objective are subjective probability forecasts? More precisely, to what extent do different forecasters, given the same array of synoptic information, assign the same probability of occurrence of a certain event? This is a sensitive point because the users' confidence in the forecast is likely to be undermined if widely differing probabilities are quoted in a specific instance, even though in fact each forecaster's advice is valid and as sharp as his colleagues'. To provide some evidence, forecasts made by each of 12 students for two days during the Spring 1961 series were analyzed. Past data for March 1958 for the European area were being used at

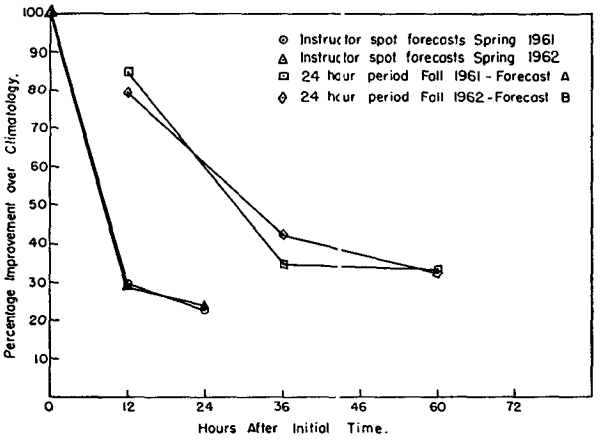

FIG. 6. Percentage improvement over climatology as a function of forecast range.

the time. Sixteen probability forecasts for each of 10 stations were made each day. A group-mean probability was computed for each forecast and the departure of each student's value was determined to the nearest tenth. Histograms of the frequency distributions of student departures for the two days are shown in Fig. 7. On both days the root-mean-square departure from the student mean is 0.12 and the distributions are virtually identical, despite unlimited discussion among the students and instructor during formulation of the forecasts on the first day and complete silence on the second. From the lack of difference in the distributions one can only conclude that there was much more talking than listening on the first day. If the forecasts for which the student mean was equal to the climatological probability (to the nearest tenth) are separated from the rest, however, different distributions result, as illustrated in Fig. 8. The relatively small dispersion for these forecasts suggests that when the synoptic information offers little concrete guidance the forecasters agree on remaining close to the climatological expectancy. On the other hand, when the information affords

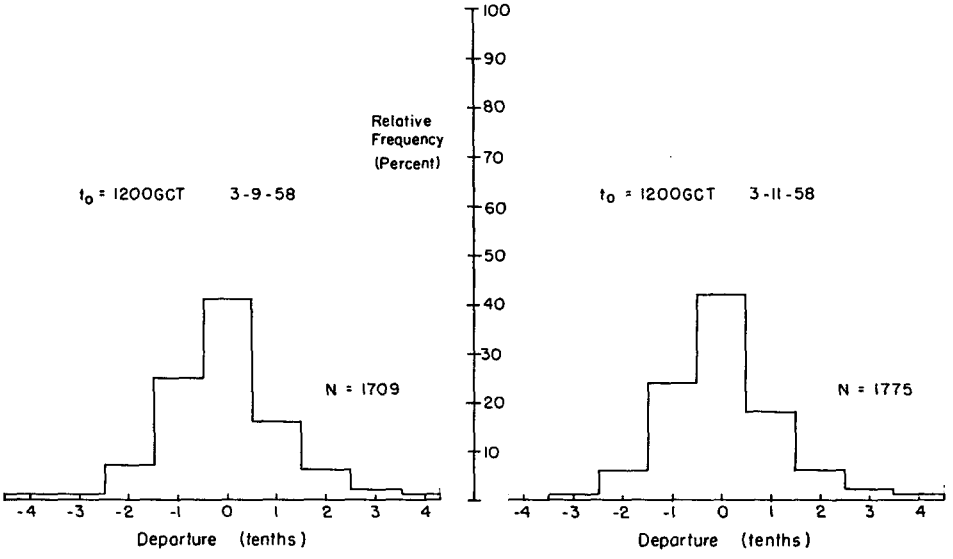

FIG. 7. Distribution of individual student forecast probabilities about the group mean for each forecast, for two days in the Spring 1961 series.

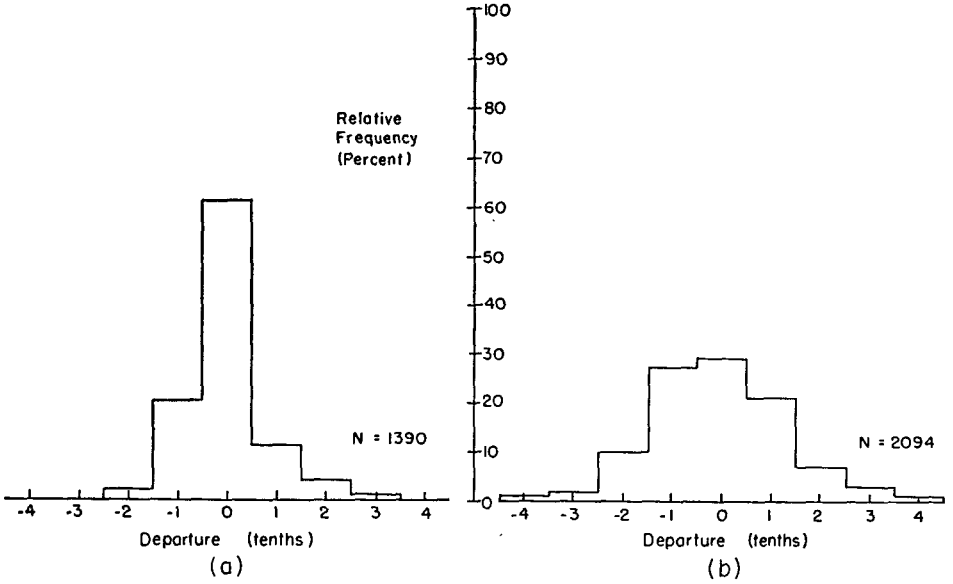

FIG. 8. Distribution of individual student forecast probabilities about the group mean for two days in the Spring 1961 series a) for forecasts in which the group mean was equal to the climatological probability and b) for forecasts in which the group mean was not equal to the climatological probability.

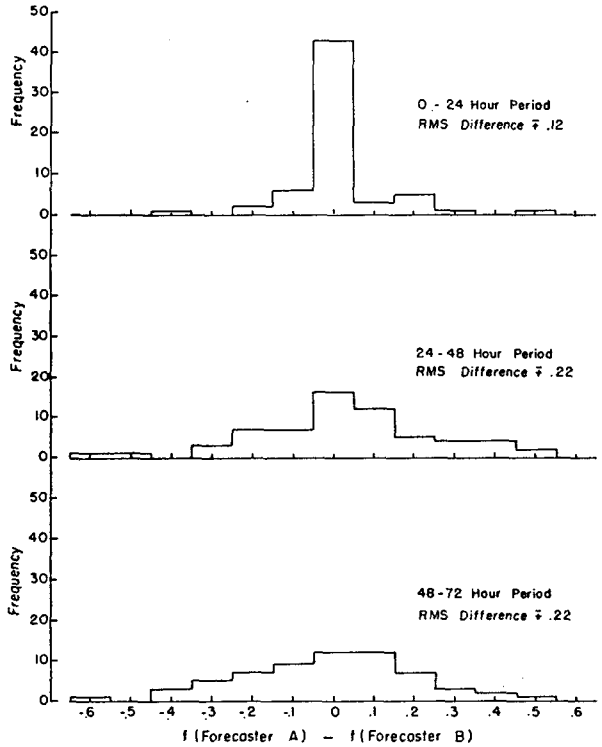

FIG. 9. Distribution of differences between forecast probabilities of Forecasters A and B, for Fall 1961 series.

a prospect of gain over climatology, the opportunity is grasped more firmly by some forecasters than by others.

A most interesting result appeared as a by-product of this analysis. The Brier score was computed for the group-mean probability for each forecast. The improvement over the climatological control score was 22.9 per cent for this mean forecast, while the mean improvement shown by the individual student forecast scores was 15.6 per cent. Moreover, the best individual student score represented an improvement of 17.8 per cent over climatology. Here is clear evidence that 12 heads are better than one.

The rather small dispersions shown in Figs. 7 and 8 may be attributable in part to the similarity of background and training among the 12 students. By way of comparison, the probabilities offered by Forecasters A and B during the Fall 1961 series were compared. These individuals were both experienced but were trained separately and had worked together only briefly during their professional careers. Differences between their forecast probabilities for the various 24-hr period are illustrated by the histograms in Fig. 9. On the whole the deviations are not inconsistent with the distributions shown in Figs. 7 and 8. If deviations from the group mean are randomly distributed among the students the root-mean-square difference between two student forecasts for the distributions shown in these figures would be 0.17. The near unaminity in the probabilities for the first period in Fig. 9 is due in part to the near certainty due to availability of data subsequent to the initial time. At sufficiently long time ranges the dispersion would again become small, since both forecasters would doubtless acknowledge their inability to offer advice other than the climatological likelihood, but this point had evidently not been reached at 72 hr.

Even with these experienced forecasters, the advantages which might have been gained from consultation were apparent. The score for the mean of the two probabilities on each forecast represented an improvement over climatology of 84.6 per cent for the period from zero to 24 hr, 44.5 per cent from 24 to 48 hr, and 37.0 per cent from 48 to 72 hr. This result represents greater skill than that shown by either individual forecaster except for the first period, as can be seen by inspection of the comparable data in Fig. 6.

The value of "consensus" forecasting was explained to the students prior to the Spring 1962 series. Extensive consultation accompanied the preparation of their forecasts and evidently had a beneficial effect, since their performance as a group was better relative to the instructors' in this year than in 1961. (See Table 4.)

8. Summary and conclusions

The foregoing evidence clearly indicates that forecasters are capable of making skillful statements about the probability of occurrence of a wide variety of meteorological events over ranges up to 72 hr. Skill is here measured relative to a climatological control forecast. These predictions, when made by experienced forecasters, are highly valid in the sense that the relative frequency of occurrence tends to correspond closely to the stated probability. Much of the skill, together with reasonable validity, develops early in the career of novice forecasters.

Use of the Brier score has been found completely satisfactory as a method of encouraging the making of skillful valid probability statements and of evaluating the results. This score may be expressed as the sum of two contributions, one of which is a measure of the validity of the probability forecasts and the other of which is a measure of their sharpness (i.e., nearness to certainty or deviation from climatological expectancy). The latter contribution is predominant and its general level for a given type of forecast is set by the state of the forecasting art though substantial variations among individual forecasters occur.

Sharpness and validity are clearly identified with the two aspects of subjective preparation of probability forecasts, respectively, *sorting* all instances into an ordered set of categories of likelihood of occurrence, and *labeling* each category with a specific likelihood, or probability, of occurrence. Differences between the skills of individuals stem from varying abilities in both of these aspects of forecasting, though variations in sorting ability appear to be larger.

Different forecasters tend to assign similar probabilities in a specific forecast instance. The root-mean-square deviation from a group-mean forecast probability is between 0.1 and 0.2. There is merit in consultation among forecasters, however, since the group-mean probability forecast is found to be a more skillful statement than the probability forecast of the most skilled individual, provided the experience levels of the individuals are roughly similar.

Analysis of probability forecasts for a wide variety of surface weather elements over various regions of the Northern Hemisphere indicates that:

1) Forecasts for a specific instant (spot forecasts) are roughly half as skillful as forecasts for 24-hr periods in the 24- to 48-hr range.

2) Skill in spot forecasting of individual elements varies with the directness with which the element can be inferred from prognostic charts. This forecast skill is greatest for wind direction and least for precipitation.

3) Skill is small or absent in forecasts of meteorological events which have extremely high or low climatological frequencies of occurrence.

4) Forecasts over the North Atlantic Ocean are more skillful than those in Europe and Asia, presumably because of the greater synoptic representativeness of the observations and the greater synoptic clarity of the circulation and weather patterns.

It is generally agreed that there is an urgent need for unambiguous yardsticks for the measurement of forecast performance. The use of the Brier score, together with application of a suitable climatological or persistence control, is a highly flexible and meaningful verification technique which appears to fill this need. This score may not be a suitable measure of the effectiveness of the forecast in the making of a particular operational decision, but as a simple overall measure it has considerable merit.

Use of the Brier score requires that predictions be in probabilistic form. Though forecasters do not now generally do so, it is felt that they are fully capable of making such predictions without compromising their ability or integrity. Should they be reluctant to do so because of personal inclination or lack of time, their categorical statements can readily be converted to probability distributions, as has been done by the U. S. Weather Bureau,[2] though not without some likely sacrifice of skill which they would be capable of demonstrating. Some continuously varying elements lend themselves only awkwardly to probability statements of the type discussed here, but a forecast of the most probable value together with an estimated error distribution would serve the purpose. For example, a forecast of the most probable temperature, together with an estimate of the standard error in a normal error distribution could be converted to a probability forecast for an arbitrary set of five-degree categories.

Aside from the merits or disadvantages of offering quantitative probabilities to the general public or other users of meteorological information, it is urged that probability be acknowledged as the proper internal language of forecasters.

Acknowledgment. The author wishes to express his gratitude to the students and instructors of the M.I.T. synoptic laboratory program who provided much of the experimental data and assisted in its analysis.

REFERENCES

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

——, 1957: Effect of errors in estimating probabilities on the usefulness of probability forecasts. *Bull. Amer. meteor. Soc.*, **38**, 76–78.

——, and R. A. Allen, 1951: Verification of weather forecasts. *Compendium of meteorology*, Boston, Amer. Meteor. Soc., 841–848.

Bross, I. D. J., 1953: *Design for decision*. New York, The Macmillan Co., 47–52.

Dexter, R. V., 1962: Confidence factors are fictional. *Weather*, **17**, 123–133.

Gringorten, I., 1958: On the comparison of one or more sets of probability forecasts. *J. Meteor.*, **15**, 283–287.

Miller, R. G., 1962: Statistical prediction by discriminant analysis. *Meteor. Monogr.*, **4**, No. 25, 54 pp.

Root, H. E., 1962: Probability statements in weather forecasting. *J. appl. Meteor.*, **1**, 163–168.

[2] U. S. Weather Bureau, 1961: Verification of the Weather Bureau's 30-day outlooks. *Technical Paper No. 39*, 6–7.