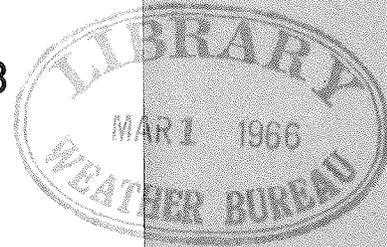


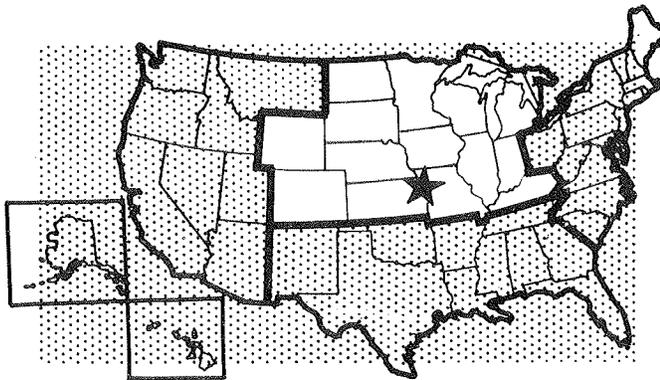
QC
995
UG15
NO.3

TECHNICAL NOTE 20-CR-3



TECHNICAL NOTE 20-CR-3

On the Probability
Forecasting of the Occurrence
of Precipitation



CENTRAL REGION NOTE NO.3

WASHINGTON, D.C.
November 1965

WEATHER BUREAU TECHNICAL NOTES

Regional Notes

The Technical Notes series provides a medium for the documentation and quick dissemination of research and development results not appropriate or not yet ready for formal publication in the standard journals. The primary objective of the series is to provide continuity of documentation for the Weather Bureau's research and development effort. The Regional Notes will report on investigations devoted primarily to regional or local problems.

These notes are for limited reproduction and distribution and do not constitute formal scientific publication. Reference to a note in this series should identify it as an unpublished report.

The reports are available through the Clearinghouse for Federal Scientific and Technical Information, U. S. Department of Commerce, Sills Building, Port Royal Road, Springfield, Virginia 22151.

135 432

U.S. DEPARTMENT OF COMMERCE ● John T. Connor, Secretary

ENVIRONMENTAL SCIENCE SERVICES ADMINISTRATION

Robert M. White, Administrator

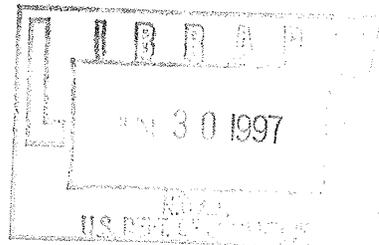
U.S. Weather Bureau

TECHNICAL NOTE 20-CR-3

On the Probability Forecasting of the Occurrence of Precipitation

Lawrence A. Hughes

CENTRAL REGION NOTE NO.3
WASHINGTON, D.C.
November 1965



135 432

M(055)
U587ce
No.3
c.1

CONTENTS

	Page
1. INTRODUCTION	2
2. DEFINITION OF A RAIN OCCURRENCE	2
3. DETERMINATION OF FORECAST PERIODS	5
4. DETERMINATION OF FORECAST VALUES	6
5. VERIFICATION GRAPHS	7
6. VERIFICATION SCORES	9
Brier Score	9
Brier Score Using Climatology	13
a. Score by Subsets	14
b. Total Score Method	16
c. Comparison of the Two Methods	19
d. Other Aspects	21
7. CHICAGO RESULTS	22
2-1/2 yr. Sample	23
Seasonal Variations	30
Words vs. Numbers	31
8. SUMMARY AND CONCLUSIONS	33
ACKNOWLEDGMENTS	35
REFERENCES	36

ON THE PROBABILITY FORECASTING OF THE OCCURRENCE OF PRECIPITATION

Lawrence A. Hughes

U. S. Weather Bureau, Kansas City, Mo.*

ABSTRACT

Problems in forecasting the point probability of the occurrence of measurable precipitation are discussed and results of such an experiment at the Weather Bureau Forecast Center, Chicago, are given based on 2 1/2 yr. of data.

Discussed are definitions, verification graphs, and the Brier score without and with climatology considered. The latter score makes allowance for variations in climatic frequency but does not compensate for all effects of climatic variations, especially climatic variation in areal coverage. Two methods of computation of the Brier score with climatology included are discussed. One gives a score which is relatively laborious to compute but which gives a very clear relationship of skill to the distance of the forecast probability from the climatological frequency.

In the Chicago result, the reliability was high with an average deviation of only 2.1 percent in the first 12-hr. forecast period and 4.8 percent in the fourth period. The resolution was moderate in the first period and dropped rapidly thereafter. Skill dropped almost to 0 for periods beyond about 36 hr. from forecast time. Seasonal variations were minor but were underplayed by the forecasters.

When the probability numbers were compared with probability words prepared concurrently, a wide range of probability numbers was found in each of the four main word categories selected, suggesting that the meaning of the words is not constant among forecasters. In the "no rain" category the number of forecasts had high reliability, thus showing information not communicated by the forecaster. Probability numbers are thus better than probability words.

*Most of the work on this project was accomplished when the author was assigned to Weather Bureau Forecast Center, Chicago, Ill.

1. INTRODUCTION

Weaver [10] in discussing probability says that "Life is an almost continuous experience of having to draw conclusions from insufficient evidence." He notes that rarely do we know that A, B, C, etc., are completely and reliably true, and thus we cannot, by the rules of classical logic, obtain an inevitable and unique conclusion. Instead, he says, there are various alternatives, none of which is certainly correct and none certainly incorrect. Thus there are varying degrees of probability for these alternatives, and the more evidence upon which to determine them, the higher the probability that any one will be true or not true. Weaver's comment is certainly applicable to meteorological conditions, even in this day of radar and computing machines, and is likely to be true for the foreseeable future.

Brier [1] stated the case for the weather forecaster nicely when he said, "The decisions of a rational man will to a large extent depend upon his estimates of the probabilities of the different events and the consequences of them. When he is convinced that the weather forecaster's estimates of these probabilities are better than his own, he will come to him for weather information."

In general the farther ahead we attempt to forecast the weather, the more diverse are the possibilities. Even if the weather for the next hour were forecastable with precision, the events of a week or a month from now are not forecastable at all except as indicated by climatology. Thus the certainly, or probability, of a meteorological event depends on its distance in the future. Even for a few hours in advance there is nearly always some degree of uncertainty. U. S. Weather Bureau forecasters, in recognition of these facts, have long expressed precipitation forecasts in probability terms. The difficulty has been that the words used have not been adequate to the task, with the result that the intended probability has not been clearly communicated to the users.

A number of reports have been published on trial projects to express precipitation forecasts as a numerical probability, (for example: [6, 7, 8, 9, 11]) and the authors found that it can be done. A few Weather Bureau stations, notably San Francisco and Los Angeles have issued such probabilities to the public for some years. The following concerns the problems and results of such an experiment at the Weather Bureau Forecast Center Chicago.

2. DEFINITION OF A RAIN OCCURRENCE

An early need in the experiment was to define a rain occurrence. This involved the decision "How much rain in how many rain gages?" For the amount, a common choice, and the one made at Chicago, is to consider a trace of precipitation (less than .005 in.) as "none". This was felt desirable because there are so few situations in which a trace is significant, especially when it is only a few flakes of snow.

Figure 1 gives the cumulative frequency of 12-hr. precipitation amounts, in inches, for the Chicago official station (Midway Airport) for the 3-yr. period 1958-60. In 53 percent of precipitation occurrences, the recorded amount was "trace" or "0.01 in.", indicating the extent to which the forecaster is faced with the very difficult and, most of the time, impossible decision between the two values. Since the percentage has a minimum value of only about 40 percent in the summer and reaches a maximum of about 65 percent in the winter, the problem is a significant one in all seasons, and is especially critical in winter in Chicago.

The forecaster's problem would not be helped appreciably by choosing a slightly higher dividing value, but it would be eased considerably by including trace as a rain occurrence. While including trace as precipitation would most likely improve the verification scores, it probably would have the opposite effect on the economic worth of the forecasts and thus would be a questionable procedure. Probably the only solution to the forecaster's problem that would benefit the user as well would be a quantitative precipitation probability forecast. Very little has been done along these lines as yet; but such an approach is probably within the ability of forecasters, and offers worthwhile possibilities for the future.

An interim solution would be to forecast the probability of any amount of precipitation, trace or more, and let the user judge the expected amount of precipitation from the worded forecast.

The decision of how many rain gages will be used to verify the occurrence of precipitation is simpler to make, and depends mainly on whether the forecast is for point or areal probability. Point probability is defined as the chance of the occurrence of the event at a particular point (a specified point) in an area, while areal probability is the chance of the event at any point in an area. Since the vast majority of weather-dependent decisions is concerned with a specific area small enough that it will be completely covered by any one small convective shower, and thus may be considered a point, the point probability is a clear choice. The point probability is correctly verified only when the forecast probability is applied to a single gage.

Forecasters may feel that using point probability and a single pre-selected gage to verify is unfair to them. For example, suppose a high probability, say 80 percent, is forecast for Chicago and vicinity, and much of the city gets heavy showers, but the verification station gets only a trace or nothing. If 80 percent were the right choice, such a thing should happen

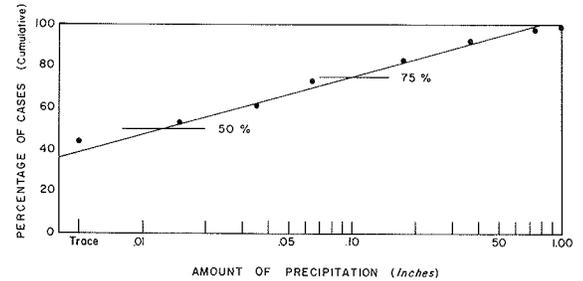


Figure 1. - Cumulative frequency distribution of the 12-hr. amount of precipitation at Chicago's Midway Airport (1958-60). A point on the graph gives the frequency of precipitation equal to and less than the amount indicated. For the frequency of reported amounts, use the highest amount of precipitation that should come under the reporting value.

2 times out of 10. The forecaster could have some reason for concern, if the forecasts were categorical, but he is fairly treated in the verification of probability forecasts, at least for a group of such forecasts.

There is some use of multiple gages for point probability, however. Since the area covered for local forecasts usually consists of several thousand square miles, the use of a single probability value in the local forecast assumes that the probability is the same at any point within the area. If the forecast probability is applied to each of a number of gages (each is verified as if it were the only gage), the amount of verification data is increased, causing the verification result to more quickly reflect the smoothing out of sampling variation. The long-term result will not be significantly changed, however, as long as the area considered is without significant local precipitation-influencing factors -- i.e. is nearly climatologically homogeneous.

It is not believed that the considerable extra effort is warranted, simply to make the sample representative quickly, as even in a month or so a reasonable evaluation could be made from the Chicago data. Should the effort not prove too much for existing manpower, additional benefit could come from such verification, such as information on whether or not the area really is climatologically homogeneous.

Such a system is likely to have more appeal to forecasters, as in the above case eight of the gages might have received a "hit" while two did not, and this appears to give a "better" verification. If they had to do the verification chore, and were to be convinced that the long-term effect was nil, multiple gage use would probably soon lose its appeal.

While multiple gages have limited value for point probability verification, they are required in order to verify areal probability or areal coverage. The relationship between point and areal probability for a single forecast is expressed by

$$P_p = CP_A \quad (1)$$

where P_p is the point probability, P_A is the area probability, and C , the expected areal coverage, is the percentage of the area (such as Chicago and vicinity) expected to be covered by precipitation if precipitation actually occurs in the area.

Note from the equation that the point probability is less than the areal probability unless it is expected to rain over the whole area, and that the point probability is less than the forecast areal coverage, unless it is certain that the area will receive rain.

The areal probability and areal coverage are verified with multiple gages. The former can be verified in the same manner to be discussed later for point probability, with a rain day defined as one with one or more gages receiving precipitation, and a non-rain day as one with no gage receiving precipitation.

Areal coverage is verified from a group of gages, and equals the percent of gages receiving precipitation on a day with precipitation. It is necessary to restrict verification only to days with precipitation because forecast areal coverage is not defined on any other days. Since the average existing areal coverage of precipitation on all days is equal to the point probability, if the area is climatologically homogeneous, a verification including all days (and therefore allowing a zero value for areal coverage) is appropriate to point probability rather than to areal coverage.

3. DETERMINATION OF FORECAST PERIODS

It is customary in U. S. Weather Bureau forecasts to use 12-hr. time periods, generally divided at 6 a.m. and 6 p.m. local standard time. This arbitrary division has some relation to the affairs of man, and partly it is for convenience in wording forecasts. However, the division is not well related to the life of weather systems, for while there are day or night maxima of precipitation in some parts of the United States during some seasons, the difference is slight over much of the country.

Strict use of fixed time periods sometimes adversely affects the probability forecasts in one sense. For example, assume that a rain band associated with a cold front is approaching a station. The rain is expected to last only 2 or 3 hr., and the forecaster feels there is a 100 percent probability that the rain will actually occur at the station. The uncertainty however is in exactly when it will occur. If the probable error of timing is estimated at plus or minus 3 hr. and the expected time of arrival is around the dividing time between periods, three possibilities exist: (1) all the rain may fall in the first period, (2) all of it may fall in the second period, (3) it may rain in both periods. If the most probable time of the middle of the rain period is exactly on the dividing time, each of these possibilities is equally likely; there are two out of three chances that it will rain in the first period (67 percent probability) and the same that it will rain in the second period. Thus the 100 percent probability of occurrence becomes 67 percent for each of the fixed time periods.

This problem would be largely eliminated by use of movable time periods. If the rain period plus the timing error is less than the length of the forecast period, the use of a movable time period could allow higher probabilities than fixed time periods, because the rain may then possibly be placed entirely within one period, even though doubt exists as to the timing. Once the length of the rain period plus the timing error exceeds the length of a forecast period, there is little to be gained by the use of variable forecast periods, because the number of forecast periods involved would not be reduced.

Since the timing errors are greater the farther in advance the forecast is made, the use of fixed periods is essentially a handicap only in the earlier forecast periods and only for occurrences of short rain periods. At stations where short rain periods are frequent, it may be advisable to use movable time periods in the earlier portions of the forecast by giving, when appropriate, the probability for, say, "this afternoon and early tonight". Forecast scores are likely to be improved by such a device, but it is difficult to know the extent to which the forecast is improved for decision making without more knowledge of user needs.

Figure 2 gives information on the duration of precipitation at Chicago, as determined from the Local Climatological Data records. Periods must have been separated by at least 75 min. to be counted as two periods, and each period, to be included separately, must have had a measurable amount of precipitation. The data were tabulated by months for the 5 yr. ending with 1960. The three consecutive months with the maximum number of short rain periods (June, July, and August) were then combined, as were the three with the longest periods (December, January, and February). The curves for these data are shown in figure 2, along with a curve for the remaining 6 mo. of the year. From the figure, assuming the forecast-timing error has no major seasonal variation, one can see that for Chicago the fixed-period problem is most prominent in summer, but is significant throughout the year.

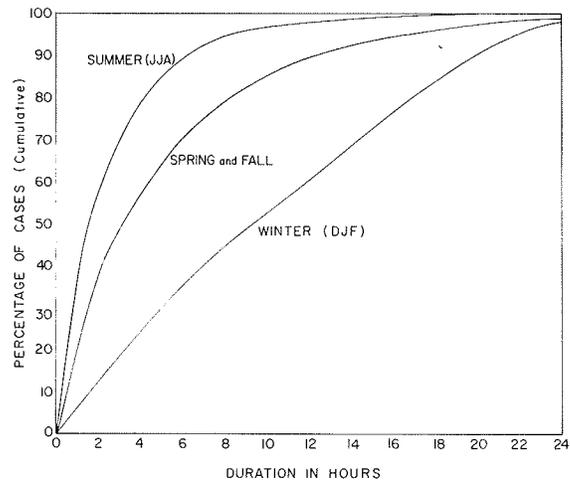


Figure 2. - Cumulative frequency distribution of precipitation duration at Chicago's Midway Airport (1956 through 1960), by seasons. A point on the graph gives the frequency of duration equal to and less than the amount indicated.

In addition there is concern in the choice of forecast periods when the frequency of precipitation has a pronounced diurnal character with relatively high frequency in the period noon to midnight and relatively low from midnight to noon or has a maximum near 6 a.m. or 6 p.m. In such cases it would seem unwise to have probabilities run for the usual day-night periods.

4. DETERMINATION OF FORECAST VALUES

The only real restriction on the set of numbers used to express the probability is that they reach from 0 to 1 (or 100 percent). The main disadvantages of a large set of values, e.g., the integral values 0 to 100, are that verification is more complex, and, for a small group of forecasters, verification is not representative for a long time because of the small number of forecasts in some categories. Both of these can be partly overcome if the forecasts are gathered into groups, e.g., forecast values 25 to 34 percent are grouped under 30 percent.

An obvious first choice for a smaller set would be the set 0, 10, 20, etc. to 100 percent. Forecasters tend to feel, especially at first, that this is as fine a distinction as their abilities and data will allow and possibly is too fine a distinction. But is this true, and is the set the wisest selection? A tabulation of the probability words used by forecasters shows that there are many more combinations of words than the 11 values in the above set of numbers.

Thus forecasters must believe that a set larger than 11 values is possible and desirable. The Chicago group started out with the set 0, 10, 20, etc., but came to realize that some finer division was desirable for the additional reasons pointed out in the following discussion.

As the verifying time of a forecast gets farther into the future, the range of probabilities that the forecasters can meaningfully use must be reduced, so that for a period perhaps a week or so away, only the probability equal to the climatological frequency can be meaningfully used. This suggests, as will be discussed in detail later, that the forecaster's skill is related to his ability to depart meaningfully from the climatological frequency (climat).

Are departures going half way from climat toward 100 percent and toward 0 percent equal increments of skill? To the extent this concept is valid, it suggests that there should be available to the forecaster an equal number of values on each side of the climatological value. Since for Chicago the mean frequency of measurable precipitation for 12-hr. periods is 22 percent, considerable change from the initial set of 11 values would be required.

Within the first month of the experiment, and without yet being aware of the above concept, the group felt the need of an additional value below the climatological value, and the 5 percent value was added. To provide a similar division near the upper extreme, the 95 percent value was added. However, the 95 percent value was later eliminated, partly because it was rarely used and partly to reduce the number of categories above the climatological frequency. Still later it was decided to add the 2 percent and the 15 percent values, with the result that the set now used by the Chicago group is: 0, 2, 5, 10, 15, 20, 30, 40, etc. to 100 percent. This breakdown is nearly equal around the climatological frequency.

There are some indications that having far more forecast values on one side of climat than on the other creates a psychological problem, and significant over (or under) forecasting of precipitation may occur.

Forecasters thus tend to have greater skill at forecasting in numbers than they initially think, and, as will be proven later, they even have skill at some parts of the probability range, in distinguishing between numbers a few percent apart. Whether or not they can show skill in separating probability values 1 or 2 percent apart through the entire range I do not know.

5. VERIFICATION GRAPHS

A convenient type of verification graph is that given in figure 3, which relates forecast probability to observed precipitation frequency and gives an impression of the reliability of the forecasts, i.e., the closeness of the points to the diagonal line. This figure is the composite of all the Chicago probability forecasts for a 2 1/2 yr. period, a total of 12,711 forecasts.

Only a small amount of "overconfidence," as discussed by Root [6] and Sanders [7], is shown in these forecasts because the points lie quite close to the diagonal line of perfect reliability. Early in our probability experiment,

we expected to see overconfidence of rain at probabilities above 50 percent (points below the diagonal line) and overconfidence of "fair" at probabilities below 50 percent (points above the line). This was probably a carryover from categorical forecasts, and it is more likely the crossover point is related to climatology.

There is little in the literature to verify this climatological dependence, but Sander's data [7, 8] supports the idea, as does Root's [6] for San Francisco. Unpublished data from another Weather Bureau station has the crossover point near its 8 percent climat value. However, the best evidence is in unpublished data from still another Weather Bureau station where the year was subdivided into frequency "seasons", with climatological probabilities turning out to be 20 percent, 40 percent, and 60 percent. These seasons had the crossover point at 20 percent, around 50 percent, and 75 percent respectively. A better fit might have been obtained if the frequency seasons were clearly separated instead of having the frequency steadily falling for 6 months and steadily rising for 6 months.

A second type of verification graph is shown in figure 4. It is a frequency distribution of forecast probabilities, which, assuming reasonable reliability, gives a visual impression of the resolution of the forecasts, i.e., the ability of the forecasters to move the probabilities toward the extreme values.

In perfect forecasting, one would have only 0 and 100 percent probabilities, with usually several times more 0's than 100's, depending on the climat frequency. The dashed line in figure 4 might be a reasonable first guess at the expected distribution for the usual (imperfect) forecasts. It has a bimodal distribution, with maxima still at 0 and 100, but with intermediate values. Root's graph for San Francisco [6] has some characteristics of this type in that it has modes both above and below the climat frequency, one actually at 0, and a minimum near the climat value.

The Chicago graph (fig. 4) is quite different in that it has only one mode, and has low frequencies at and near both 0 and 100 percent. It is unlikely that the difference between the Chicago and the San Francisco results is a consequence of the forecasters involved, because they all are experienced men in their areas. Nor is the difference in the frequency of precipitation, as that is about the same, since the San Francisco forecasts were for only the

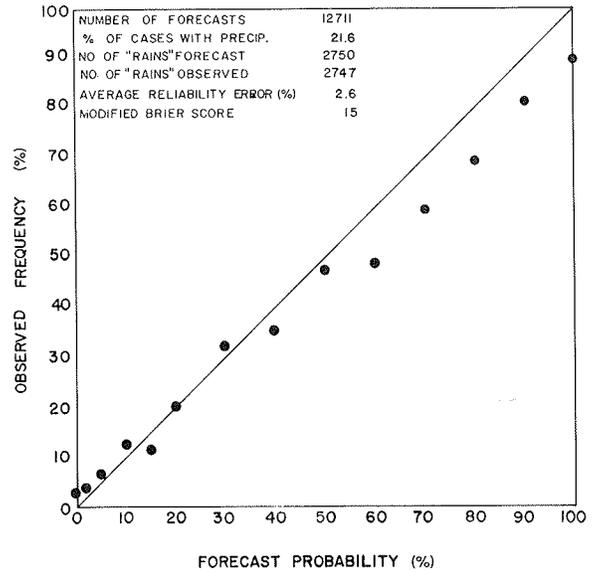


Figure 3. - Forecast probability vs. observed frequency of occurrence of measurable precipitation in a 12-hr. period at Midway Airport (2 1/2 yr. period). The horizontal line is at the long term climatological frequency of precipitation.

"rainy" portion of the year. Instead, the difference is more likely in the types of weather systems. For example, taking an extreme case, although two stations each has a rainfall frequency of 50 percent the frequency distribution of forecast probabilities for the station which may have rain 15 consecutive days and then have no rain for 15 days, is completely different from the one where it rains every other forecast period.

While a graph such as figure 2 is not available for San Francisco, it is known that rain there tends to come in lengthy periods, as also do the dry spells. Thus during a rainy period, high probabilities are likely; when there is a dry period, low probabilities would prevail with a likelihood of 0 values. For this reason the distribution would be expected to be more like the dashed line in figure 4 than like the distribution for Chicago.

Graphs like those of both figure 3 and figure 4 are needed for each of the time periods for which forecasts are made in order to study the time deterioration of the forecasts. From such graphs, one can see the systematic error of the group of forecasters for each time period, so it can be at least partially removed in future forecasts. Of more value, when a sufficiently large sample has been obtained, is to make a time breakdown for each forecaster, so he will know how to correct any personal systematic error. Time graphs will be discussed more when the Chicago results are presented.

6. VERIFICATION SCORES

The graphs of figures 3 and 4 gives a visual impression of the quality of the forecasts but they do not give an adequate verification because they do not allow a clear comparison of quality from one forecaster or station to another. A numerical scoring system which has been widely used in the verification of probability forecasts is that of Brier [2]. The ramifications of this score are discussed in detail next.

The Brier Score

For the rain-no-rain situation and for a single forecast, a form of the Brier score can be written as

$$P = (F-E)^2 \quad (2)$$

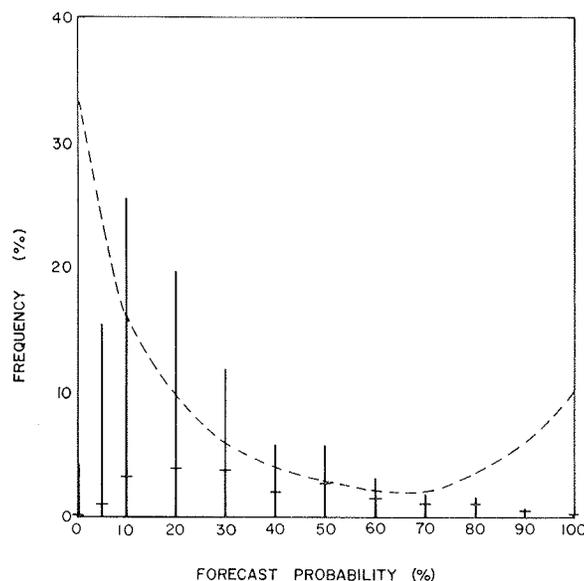


Figure 4. - Frequency of forecast probabilities for the set of probability forecasts of figure 3. The small horizontal line on each bar gives the frequency of precipitation occurrences as a percentage of all cases. For definition of dashed line, see text.

where F is the forecast probability of rain and E is taken as 1 if rain occurs and 0 if no rain occurs. For a group of forecasts, the average value of P is determined.

The range of this score is from 0 to 1, with 0 the best score. The complete Brier score has a range from 0 to 2, because it adds to the above the score for the no-rain probabilities. However, the no-rain score is exactly the same as the rain score, because the sum of the rain and no-rain probabilities is always 1, and the definition of E is reversed for the no-rain situation. Thus it is unnecessary to compute both for the simple rain-no-rain situation. To make the above score compatible with full Brier scores, one need only multiply it by two.

In computing the score, it is not necessary to obtain the differences and square each time verification is made, as for a given set of forecast values these can be obtained ahead of time and recorded. Then when verifying, one need only multiply the appropriate recorded value by the number of forecasts in that category, add all categories, and average to obtain the final score. Hand calculations are thus fast and easy.

Figure 5 is a graph of the score. It is probably not as easy to compute the score from the graph as from tabulated values, but some interesting points concerning the score can be more easily seen from it than from the symbolic form of the score. For example, we can see that for any reasonable and likely reliability, the real range of the score is much more like 0 to 0.3 than 0 to 1. We can see that the forecast of 50 percent probability is going to give a poor score, since it will verify 0.25 regardless of the reliability of the forecast, although if a set of forecasts has an observed frequency of 50 percent, its best verification is still at a forecast value of 50 percent. We can see that reliability is important mainly for the very low and very high values of the forecast probability, as far as getting a good score is concerned.

A single forecast, of course, must have an observed frequency of either 0 to 100 percent, and thus is not concerned with the center portion of the graph. In evaluating the score, one would group forecasts with, for example, all 60 percent forecasts gathered into two groups, one for rain cases and one for no rain. These too would not use the center of the graph. But when the average score for any forecast probability is computed, then the whole graph becomes useful. It is the average for a forecast category that is significant to the forecaster and to the user of the forecast.

We can also see that for a given forecast probability, the best score is not obtained at the point of perfect reliability, i.e., the point where forecast probability is equal to observed frequency, except for the 50 percent value. For forecast probabilities above 50 percent, the best score is achieved if the observed frequency is 100 percent, while for values below 50 percent, the best score is with 0 percent.

It may seem strange that a set of forecasts of the 40 percent forecast probability which verify at the 20 percent observed frequency is better than a similar set verifying at 40 percent observed frequency. However, the set with the poorer reliability has economic advantages, as, from the point of

view of cost-loss ratios, the cost of protection is the same in the two cases since it depends on the forecasts and not on the occurrences, but the losses are different, since users whose C/L ratio lies above 40 percent take a lower loss with the less reliable set of forecasts. This is definitely not to say that reliability is undesirable and unprofitable, for if the forecasts had been placed at 20 percent instead of 40 percent, their score would have been better and the economic value would have been higher. The error, if for a representatively large sample, is personal bias and should occur to a lesser degree in future forecasts, once the forecaster is aware of it.

One decided advantage of the Brier scoring system is that it cannot be "beaten" or played by the forecaster, to his advantage and the users' disadvantage. This is because the score is the sum of the scores for each individual forecast divided by the total number of forecasts. Thus the scores of forecasts that have already been produced cannot be affected in any way, and the way to get the best total score is to make the forecast at hand just as good as possible, i.e., as low as possible if it doesn't rain, and as high as possible if it does rain. Any attempt to "improve" particular categories (as to reliability) by placing forecasts in one category when the forecaster believes they belong elsewhere, will most likely hurt the total score. This is proper too, since there is no way of knowing the subsets of concern to particular users and improvement of a category overall may not be an improvement for a particular user.

The concepts of reliability and resolution deserve more comment. Starting with the symbolic form of the Brier score (equation (2)), the score for a particular forecast category F with an observed precipitation frequency ϕ can be obtained by taking the score for the days with rain as $\phi (F - 1)^2$, and days without rain as $(1 - \phi) (F - \phi)^2$, adding the two and also adding and subtracting ϕ^2 and rearranging to get

$$P_F = (F - \phi)^2 + \phi (1 - \phi) \quad (3)$$

Reliability is a measure of the closeness of the points in figure 3 to the diagonal line and is given numerical value in the above equation by the quantity $(F - \phi)^2$. It is thus the squared deviations from the diagonal line and it quickly approaches 0 as ϕ approaches F. Good reliability ($F \approx \phi$) is rather easy to achieve, and is usually accomplished after a short time, almost regardless of the ability of the forecaster, in agreement with Sanders [8]. This is because reliability can be thought of as a measure of the extent to which the forecaster knows his ability, and the clarity of probability

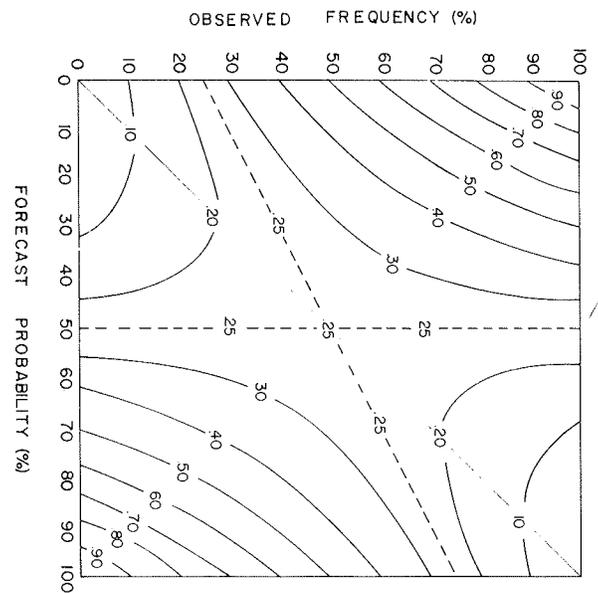


Figure 5. - Graph of the simple Brier score.

verification allows quick evaluation of one's limits. For example, if a forecaster has little ability to forecast rain but is aware of the climatological frequency, he can obtain good reliability by always forecasting the climat value. The score for such a forecast can be obtained at the point where the climatological observed frequency intersects the diagonal line in figure 5 (0.17 for a climat of 22 percent), and a score (for all forecasts) to show skill must be lower than this value. This point will be gone into in detail shortly. A forecaster with a small amount of skill can move some of his points away from climat, but the amount of movement must be small for many points, or large for only a very few. A forecaster with low skill who thinks he has large skill will forecast values far from climat, but must have poor reliability. However, a forecaster could under-estimate his skill and not be aware of it from reliability measures.

Resolution measures the ability of the forecaster to resolve the cases into rain and no-rain days or to move points toward the extreme probabilities of 0 and 100 percent. From the argument just above, it can also be considered as a measure of the ability to move the points away from the value of climat. If reliability is good (the forecaster knows his skill), resolution is a measure of the forecaster's skill. Thus, reliability is a measure of how well the forecaster knows himself (knows the limits of his skill), and resolution is a measure of how well the forecaster knows how to forecast precipitation.

Resolution is given numerical value by the quantity $\phi(1 - \phi)$ in equation (3), and varies between 0 and 0.25, depending only on the observed precipitation frequency in a forecast category. To show the complete separation of resolution and reliability scores, and the complete dependence of the resolution score on the observed frequency, take the case in which forecasts of only 0 and 100 percent are made, and it rains on every 0 forecast and on none of the 100 percent forecasts. These are very poor forecasts, especially if there is a sizable group of them, but skill is shown in making them. This is seen in that there is no resolution penalty in these cases as $\phi(1 - \phi)$ is 0. This is perfectly reasonable, since there has been perfect separation of the rain and no-rain days, but the reliability penalty will be very large.

We have here a case of a man who knows how to forecast precipitation perfectly — he has no resolution penalty — but he is completely unaware of his skill — his reliability penalty is the worst possible. This example fully confirms the word-definition of resolution and reliability just given, for when this man learns that he need just reverse his forecast numbers from those he has been using — when he learns of his ability — he will also have perfect reliability.

If, as is usually the case, the bulk of the forecasts are not in the extremely low and extremely high forecast categories, and the reliability is fairly good, the resolution score is about an order of magnitude larger than the reliability score. It is therefore generally not considered necessary in the simple rain-no-rain situation to make the separation, as the reliabilities shown by experiments so far are such that practically all of the deviation of the score from perfection is due to resolution. This indicates that the forecasters understand their limitations well, as indeed highly experienced forecasters should, but, as will be seen later (in the Chicago results), the

moderate scores and their rapid decrease with time show that the understanding of how to forecast precipitation is not high.

It is easy to see from figure 5 that if reliability is good, the Brier score measures the ability to move the points away from 50 percent instead of from climat. Thus the Brier score is more suited to a climat near 50 percent than to one far from that value, and it therefore may not be adequate in comparisons of sets of forecasts in which the climatological frequency is far from 50 percent or markedly different.

Brier Score Using Climatology

An application of the Brier score which makes some adjustment for variations in the climatic frequency can be made using

$$S = 100 \left(\frac{B_c - B_f}{B_c} \right), \quad \text{or} \quad S = \left(1 - \frac{B_f}{B_c} \right) \times 100 \quad (4)$$

where B_f is the Brier score of the forecasters, computed as above, and B_c is the Brier score of climat, i.e., a score obtained by assuming that a forecast equal to the long-term climatological frequency is made each time a forecaster makes a forecast. This Brier score is comparable to that discussed by Sanders [7, 8].

We can see from either form of the above score that if the Brier score of the forecasters is higher (poorer) than the Brier score of climat, then the above score will be negative. Also, if $B_c = B_f$, the score will be 0. Thus with a range from 100 percent for a perfect score (when $B_f = 0$) to less than 0 for poor scores, the above score may be thought of as the percentage improvement of the forecaster's probability over the use of the climatological frequency as a forecast. It thus compensates for climatological variations in precipitation frequency. It does not fully compensate for climatological differences, however, as will be seen later.

There are two ways in which this score can be formed from a set of forecasts. Both have meaning and usually will give different results, although the differences may not be large. In one method, the score of each forecast category subset is obtained, with the total score for all forecasts taken as their weighted average (subset method). This is the method used by Sanders [7, 8]. In the second method (total score method), the numerator and denominator are each computed for all forecasts before the division is made to get the score. In symbolic form, using the B_f/B_c form of the score, the first method uses $\frac{B_f}{B_c}$, while the second uses $\frac{B_f}{B_c}$. Each of these scores has its advantages, and each will be discussed.

a. Score by Subsets

Computing the score by subsets $\left(\frac{B_f}{B_c} \right)$ is the more cumbersome method,

requiring about triple the effort needed for the Brier score alone, because the Brier score must be computed for both the numerator and denominator in a fashion similar to the Brier score alone, and then the scores for subsets must be obtained and weighted according to the number of forecasts in the subset before averaging.

Graphs of this score for various values of climat are transformations of the Brier graph of figure 5. For a climat of 50 percent, the pattern of figure 5 is unchanged, with the 0.25 values becoming 0, the upper right and lower left corners becoming 100, and the lower right and upper left corners becoming minus 300. Note that the end points of the sloping 0 line are each halfway from the value of climat toward perfection.

The graph for the Chicago climat of 0.22 is given as figure 6. The zero lines shift to meet at the new value of climat, with the end points of the sloping 0 line having the same relationship as before. The value in the upper right and lower left corners remains the same (100) for all climat values, but the values in the other two corners change. When the values on the ordinate and abscissa of figure 6 are changed to read in the opposite direction, it becomes a graph for a climat of 78 percent.

Figure 5 clearly suggests that the set of forecast probabilities should have equal increments on either side of the climat value, but the non-uniform spacing of figure 6 suggests the need for smaller increments below climat than above. A roughly equal quantity of reasonably spaced values on either side of climat would satisfy the requirements, thus substantiating the earlier argument to that effect.

The graphs clearly show that a forecast of the climatological frequency will give a 0 score if the set is representively large, and that score (skill) increases as reasonably reliable points are moved away from the climat value.

Since this score uses the climatological frequency as a base for measuring skill, what effect will seasonal variations in the climatic frequency have on the scores? Figures 5 and 6 provide some of the answer. Assume that the bulk of the forecasts are in forecast categories within 15 or 20 percent of the value of climat for each frequency "season". This assumption will be very

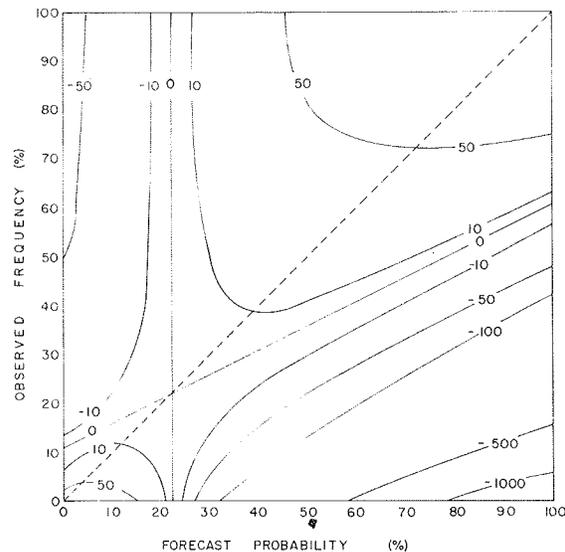


Figure 6. - Graph of the Brier score with climatology included in percent, when computed by forecast category subsets, for a climat of 22 percent.

good for the forecast periods farthest in the future, and reasonably good for the first forecast period. As an example, take a location with three frequency seasons of 22, 50, and 78 percent each. We can see from figures 5 and 6 that if the 50 percent graph is used for all forecasts, better scores will result in both the low and high frequency seasons than would result if the graph appropriate to the season had been used in each case. Of course, the smaller the range of climat values the less important that separate graphs be used for computation. Climat values within 5 percent or possibly 10 percent of each other may have no appreciable effect on the scores, so the variations among forecasters at the same station (± 3 percent of climat at Chicago) are not likely to be significant in the ranking of these forecasters.

The concepts of reliability and resolution can be carried over into this score. Taking the form of the score with the ratio B_f/B_c , which can be considered as a penalty to be subtracted from 1, the numerator can be separated into reliability and resolution by the method discussed under the Brier score. If this is done, we get a penalty due to lack of reliability and a penalty due to lack of resolution. Graphs of these penalties for a climat of 0.22 are given as figure 7.

The portion of the reliability penalty graph near the 0 line will change little through the likely range of climat values, but the lower right and upper left corners will change a great deal. Thus for a fairly reliable set of forecasts, this graph would be essentially correct for any climat, and the penalty would rarely exceed 10 percent and would likely be less than 5 percent as the average for all forecasts.

The resolution penalty graph is a series of horizontal lines each having the same value for all forecast probabilities, since it is dependent only on the observed precipitation frequency (ϕ) in any forecast probability subset. (The numbers on the extreme right will be discussed later.) The upper and lower borders of the graph have a 0 value for all climat values, while the 100 value lies at the observed frequency equal to the value of climat, thus clearly showing the reduced penalty as the observed precipitation frequency of the forecast sets is made to move away from climat. The resolution graph may change considerably with differing values of climat. For example, reversing the coordinate labels again creates a graph for a 0.78 climat. The resolution graph for any other climat is easily visualized.

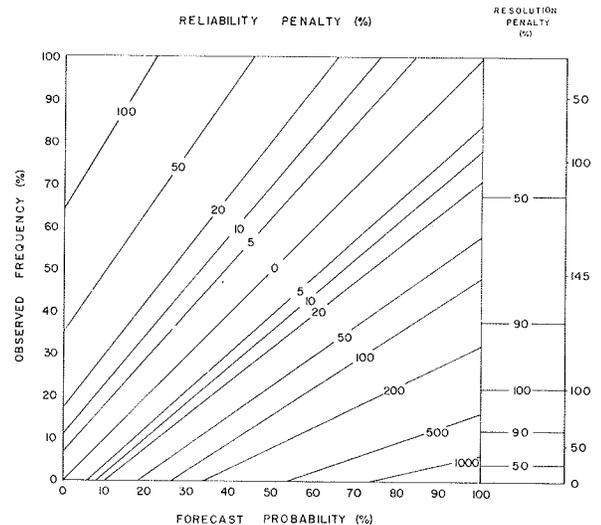


Figure 7. - Graph of the reliability penalty and of the resolution penalty for the Brier score with climat considered when computed by subsets. The numbers on the extreme right give the resolution penalty when the score is computed for all forecasts together (see text).

It can be easily seen from these graphs that unless the bulk of the forecasts is near the 0 or 100 percent observed frequencies, the resolution penalty is likely to be many times the reliability penalty. Thus while the concepts of reliability and resolution have meaning, the deviation of the total score from perfection is usually almost entirely due to poor resolution, and it is likely that computation of the separate penalties is not necessary.

The main advantage of this subset method of forming the score is the clear indication that score (and skill) is directly related to the amount of deviation of the forecast probability from climat, with all reasonably reliable subsets having better scores than a constant forecast of the climat value. The main disadvantage is the large amount of calculation necessary, especially if one computes scores for different forecast periods, different years, and for each forecaster.

b. Total Score Method

The total-score method is far simpler to use. The numerator B_f is computed as if only the simple Brier score were desired, except that averaging the scores is not necessary or desirable, and the figure for the total score is used. The denominator B_c requires very little computation, since only one value -- the long term climat -- is allowed as a forecast value, and all forecasts are grouped for the computation (rain days and no-rain days are computed separately, then combined as for the numerator). The total score is also used in the denominator. Performing the division accomplishes the averaging, so the score by this method requires only a little more effort than the simple Brier score - a decided advantage if hand computations are made, or even if a lot of scores must be computed by machine.

As the quantity of forecasts being verified increases, and the observed frequency of the sample ϕ approaches the long-term climatic frequency C , it can be seen by substituting C for F in the Brier score (equation (2)) and expanding (see equation (3) with $F = \phi$) that the Brier score of climat (B_c) approaches a constant equal to $C(1 - C)$. Thus in comparing plain Brier scores from different geographic locations, some adjustment for climatic variations can be made very simply by dividing each score by its appropriate limiting value of $C(1 - C)$.

When scores are computed for individual forecasters at a single station, B_c will also approach $C(1 - C)$. If it equaled $C(1 - C)$ for all forecasters, the ranking of forecasters would depend only on the Brier score itself. But in the 2 1/2 yr. at Chicago this state was not yet reached, as the precipitation frequency varied among forecasters from 18.9 percent to 24.6 percent. These variations could be significant in the ranking of forecasters, and the extent to which they can, as well as other items to be discussed, can be seen from figure 8. The figure shows only the variations in B_c and not in the total score, because the B_f score is not subject to variable interpretations as long as one holds to the unadulterated Brier score, and because the variations in the total score depend on B_f .

The straight line gives the full B_c score for a climat (C) of 0.22 and all observed rain frequencies ϕ (from equation (3)). We can see from this line that even for a variation of ϕ among forecasters of only ± 3 percent

(comparable to that at Chicago), the B_c score changes by about ± 0.017 , or about 10 percent of the 0.17 values at a C of 0.22. These variations are likely to have an effect on the ranking of forecasters and they did at Chicago. Therefore, the B_c score of $C(1 - C)$, where C has the same value for all, generally should not be used in computing scores to rank forecasters. The effect of these variations on the total score becomes more significant as the value of C goes toward 0 and as the value of B_c goes toward 0.

While I consider it inappropriate, I have heard it said that since it is only chance that determines whether a forecaster is required to forecast on a group of days with above or below average frequency of precipitation, his forecasts should be verified using a C appropriate to his set of forecasts instead of the long-term value used for all forecasters. In this case $C = \phi$, and the curved line in figure 8 gives the B_c score for any ϕ . Note that for variations of a few percent around the long term value of C , the B_c score is essentially the same on both the straight and curved lines and the difference between these scores would not likely affect the ranking of forecasters. The effect on the ranking would probably be negligible for any group of forecasters who regularly alternate forecast shifts for a representatively long period (a year?). But here, too, the problem becomes more critical at the very low values of climat. One should note, however, that the B_c score via the long form is higher (total score better) for climatic frequencies below and above the average (long term) value, so forecasters with either a minus or a plus deviation will have a slight advantage in the score.

Using C as the sample frequency (and therefore always equal to ϕ) instead of the long-term climatic frequency gives an interesting sidelight. Such a system cannot be a true no-skill base for forecasting, as the value C cannot be used for day-to-day decision making since it is not known until the sample has been obtained. In this sense it is kin to the Chance Skill Score. In fact, if the form $\phi(1 - \phi)$ is used for B_c , and the square root of the modified Brier score is taken, the result is equivalent to the chance skill score and may be compared with such scores obtained from previous categorical forecasts. Also, if the probability forecasts are converted to categorical forecasts using a dividing criterion that produces exactly as many categorical rain forecasts as observed, the chance skill score of the categorical forecasts so produced is exactly the same as the square root of the modified Brier score. Even if B_c is computed from its long form, the result is close to the chance skill score and can still be compared to such scores.

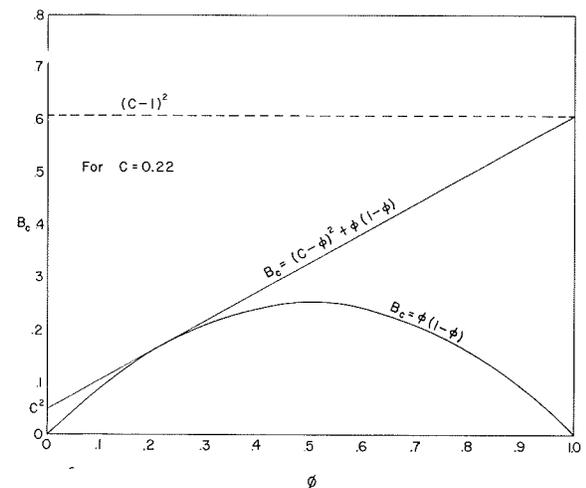


Figure 8. - Graph of the complete and the simplified B_c score, for a climat value C of 0.22.

Another use of the curved line is in the consideration of the significance of seasonal variations in climatology in the B_c score. If seasonal values of the long-term climat are within about 5 percent of the annual value of climat, there would be little gained by separating the sample into frequency "seasons", as the score on the straight line would be essentially the same as that on the curved line. But for larger values, especially at low values of climat and for variations of 10 percent or more, the effect would be significant, and the sample should be separated into frequency seasons. If this is not done, the scores for those seasons with above average and below average precipitation frequency will yield better total scores than they would if the seasonal factor were considered.

Forecast category subsets can be obtained for this score even though the score is computed for all forecasts combined. To do this, the B_f part of the penalty is computed in the same way as for the previous scoring method, except that the average value for the subset is obtained instead of the total value. For the denominator B_c , the average value for all forecasts, not just those in the subset, is used for each of the subsets. Assuming $\phi = C$, as will be the case eventually as has been discussed, we can obtain a graph for this score, because the score for each subset is simply the Brier score for that subset divided by the constant $C(1 - C)$.

The graph of this modified Brier score has exactly the same pattern as the Brier score in figure 5; for all values of climat, however, the lines take on different values depending on the value of the climat. For a climatic frequency of 22 percent, $C(1 - C)$ equals about 0.17. Dividing the graph by 0.17 and considering the subtraction from 1 and the multiplication by 100 produces the following: the 0.17 line on the Brier graph becomes 0, with all higher values becoming negative and all lower values becoming positive; the lower left and upper right corners are perfection and take on the value of 100; the lower right and upper left corners take on values which depend on the climatic frequency, and would be minus 485 for the climatic frequency of 22 percent. Values can be easily worked out to give the full set for any climat value needed.

For a climatic frequency of 50 percent, the area with positive score on the graph would reach a maximum, with the 0.25 lines on figure 5 becoming 0, and the lower left and upper right areas between them having positive values. As the climatic frequency approaches 0, the positive areas in the lower left and upper right also approach 0.

A graph like figure 5 but for a low value or a high value of climat points out a factor of importance in the selection of the set of forecast probability values. With a climat of 10 percent, all reasonably reliable forecasts of 10 percent through 90 percent give negative scores; thus the positive score must come almost entirely from the correct identification of the no-rain days and the use of probability values below 10 percent. It would probably be unwise in such a case to use a set of forecast values with 10 percent increments, as only one low value (0) will yield positive scores. This supports the earlier argument that the set of forecast probabilities should depend on the climatological frequency.

c. Comparison of the Two Methods

The scores for the two methods — subset and total — can be the same under two conditions. First, if in the subset method, the B_c score is the same for each subset and the same as the average B_c for the total score method. These conditions are true for a climat of 50 percent, because of the unique feature of the Brier score that the 50 percent forecast category is the same for all observed frequencies. (The B_c scores for the subset method can be obtained from fig. 5 as the values along the vertical at the value of climat.) Thus the two systems give the same score, or nearly the same score, for at least the climat values around 50 percent.

The second way for the scores to be the same is for the score of each of the subsets of the subset method to be the same, and equal to that for all forecasts. Taking as an example a climat of 20 percent, we can compare the B_c scores on figure 5 along the vertical at the 20 percent forecast value, with the B_f scores for points along the perfect reliability line. Notice, for values within 15 or 20 percent of climat, that the penalty ratio B_f/B_c is nearly equal to 1, and even for the forecast value of 60 percent, the ratio is still not down as far as 0.5. Thus if the bulk of the forecasts is fairly close to the value of climat, the subset scores will be close to the same value (0) as will the score for all forecasts (see fig. 6). Such a grouping of forecasts is not desirable, as it shows low skill, but this was the grouping at Chicago, for reasons discussed in detail later, and the score for all Chicago forecasts together was exactly the same by the two methods.

The total score form of the score can also be separated into reliability and resolution. Taking the form of the score containing the ratio B_f/B_c , the numerator can again be separated into reliability and resolution scores but this time each part will have the average B_c score of all forecasts as a denominator. This will show the contributions to the penalty of the reliability and of the resolution of the forecasts. The reliability graph for the limiting condition of $B_c = \phi(1 - \phi)$ would be very similar to figure 7, as the 0 line would be the same, but, since B_c is the same for all subsets, the other lines would be parallel to the 0 line with the same gradient on both sides of that line to a penalty for a ϕ of 0.22 of 585 in the lower right and upper left corners. (There is no resolution penalty on the top and bottom lines, so the reliability penalty in percent is 100 minus the total score.) This means that for a fairly reliable set of forecasts, there is little difference in the reliability penalty of the two scoring systems.

The resolution graphs for the limiting case, however, are quite different in the two scoring systems. For the total score method, the resolution graph has a maximum penalty at 50 percent observed frequency, lowering in a gradually increasing gradient to 0 percent at observed frequencies of 0 and 100 percent, going through a penalty of 100 percent at the value of climat C , as in figure 7, but also at the value of $(1 - C)$. The numbers on the extreme right of figure 7 show this graph in skeletal form. One can see that for a series of points around the value of climat, each of the scoring systems is high on one side and low on the other, so compensation is possible and the total penalty for forecasts in that area need not be appreciably different in the two systems. But the differences are so large point for point that large differences are

possible. However, in this scoring system as well as in the other, practically all of the deviation of the total score from perfection is due to resolution error and thus the separation need not be made explicitly.

The main advantage of the total score method is the ease of computation, although its close relationship to the chance skill score is also in its favor. Its main disadvantage, is its loss of the clear indication that score is directly related to amount of deviation from climat. This last is so because the poorest score for a perfectly reliable set is not at the value of climat but at 50 percent regardless of the value of climat. Thus it seems more suited to a climat near 50 percent.

While the poorest score occurs at 50 percent for a perfectly reliable set of forecasts, the best score for a set of forecasts with an observed frequency of 50 percent is still with a forecast probability of 50 percent, so that reliability is still a goal. One might argue that it is inappropriate to look at the forecasts by subsets instead of all together, as it seems unfair to climat to force it to take a reliability and resolution determined by the forecasters' subsets. By the subset method, climat has some resolution but has poor reliability, while by the total score method it has no resolution but perfect reliability — the latter a more reasonable situation.

One might also say that in the total score method, for each forecast above climat which gives a negative score, there must be several forecasts below climat which give positive scores, if climat is below 50 percent and a reasonable reliability is assumed. The non-linearity of the resolution penalty in the total score method yields a greater gain for a point shifted above climat, so there is a net gain when all points are considered.

One cannot generalize to the extent of saying that one score will always be higher than the other. Assuming that the score is almost exclusively determined by the resolution penalty, we can see from figure 7 that in the total score, forecasts above the value of climat (except for observed frequencies of 100 percent) will have higher scores while forecasts below climat will have lower scores than in the subset method. One could say that the total score method tends to be a "fair weather" method, since it gives negative scores on most rain forecasts but since a whole set of forecasts will have some forecasts both above and below climat, the total score by the two methods could easily be about the same. It was the same for the Chicago forecasts.

Even though the total score method might be considered a fair weather method, the subset method for all forecasts had only 30 percent of the penalty with the no-rain days, even though there were about four of these to each rain day. This clearly shows the strength of the forecasts is in correctly forecasting the no-rain situation.

I would like to recommend one of these scores at this point, but I cannot do so. If one is short of help to compute scores, the total score method would be the clear choice. I lean toward the subset method because of its clearer indication of skill, but have computed scores mainly by the total score method because of simplicity.

d. Other Aspects

The Brier score is related to the economic worth of the forecasts.¹ If one assumes that there are the same number of users in each of a set of uniformly-spaced forecast probabilities (probably an unreal assumption), and that each user takes the same loss if an unprotected rain occurs, then it can be shown that the simple Brier score is directly proportional to the cost of using a set of probability forecasts, above the cost required for perfect forecasting. The Brier score with climat score considered, under the same conditions, gives the percentage of possible savings of the forecasts over the use of the climatic frequency as a forecast.

The effective range of the Brier score with climat considered is from a bit below 0 to 100, which is considerably less than the possible range with the simple Brier score. Earlier, the range of both these scores was reduced from that theoretically possible, by raising the lower limit because of the good reliability of most forecasts. The upper limit obviously cannot be reached either, because perfect forecasting cannot be done. However, it can be reached for a large group of forecasts, as for example, the 0 category in a very dry climate.

A factor which affects the working upper limit in a major way is the areal coverage of precipitation. Since, for a rain occurrence, the point probability equals the areal coverage, as indicated earlier, it is impossible to reach the upper limit of the score with a rain situation unless the areal coverage is 100 percent.

The Chicago and vicinity area, as defined, covers about 7000 square miles, and the precipitation coverage in this area is frequently far from 100 percent. The solid curving line in figure 9 shows the cumulative frequency distribution of areal coverage over Chicago and vicinity for 24-hr. periods, based on 16 stations reporting in the period 1956 through 1960. Even for the 24-hr. period, the average coverage on a rainy day was only a bit over 50 percent, and this figure is impossible to exceed for the 12-hr. period used in the forecasts and will generally be smaller for the 12-hr. period.

The dashed line in figure 9 represents a regime which generally has widespread rains so that low areal coverages are infrequent. The dotted line represents a regime which has isolated showers as its common condition and high area coverages are not common. The Chicago curve is close to the diagonal line of uniform distribution of areal coverages, but has relatively greater

¹One of the reviewers of this paper brought to my attention a very recent proof of this statement derived by Allen Murphy of the University of Michigan and to be published.

frequencies in the low and in the high percentages than in the middle values, and therefore shows a slight tendency toward both isolated showers and wide-spread rains.

From the above it is clear that if the three curves were for regimes having the same climatic frequency of precipitation, Chicago would be expected to have the lowest Brier score with climat considered because of the larger number of areal coverages in the middle values. Thus this score compensates for only part of the effect of climatic variation. This effect of the variation in areal coverage of precipitation has not been given the attention it deserves in comparisons of sets of probability forecasts.

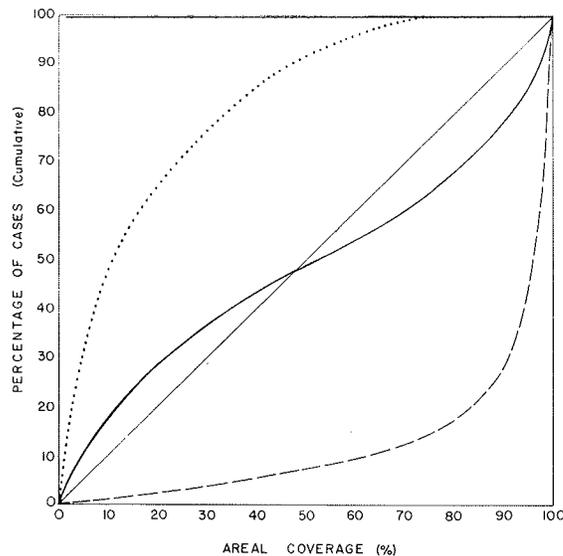


Figure 9. - Cumulative frequency distribution of the areal coverage of precipitation in Chicago and vicinity. A point on the graph gives the frequency of areal coverage equal to or less than the amount indicated. For definition of dashed curve and dotted curve, see text.

7. CHICAGO RESULTS

The results of the Chicago experiment should be representative of the output of a group of forecasters who have considerable experience and who produce probability forecasts in numbers for the first time without objective aids. They had, of course, produced probability forecasts in words for a long time, but the transition to numbers involved problems, some of which have been discussed previously.

Objective aids were felt to be desirable, but none existed. The extent to which they will help a group of highly experienced forecasters is a moot question. As expressed by Gringorten [5], "Objective methods or devices become tools for the less experienced forecaster." While this is probably still the state of the art today, the use of high-speed computing machines with statistical and dynamical techniques is likely to create objective methods of sufficient quality and quantity in the near future to improve the product of the experienced forecaster.

The Chicago experiment was started without objective aids, on the basis stated nicely by George [4]: "The formulation of methods for the mathematical prediction of weather probabilities is likely to be a slow process involving intricate calculations, but fortunately it is not necessary to await this optimum result before making a start on this highly necessary change. To begin with, the forecaster can usually make a fairly skillful subjective estimate as to his confidence for any particular forecast". The Chicago

experiment was aimed at finding out just how skillful experienced forecasters would be in their subjective estimates. The following will show that they can make such estimates with consistent and worthwhile skill and that their numerical probabilities are better than the probabilities expressed in worded forecasts.

The 2 1/2 Yr. Sample

A set of probability forecasts was produced four times a day, with a set consisting of the point probability of precipitation amounting to 0.01 inch or more in Chicago and vicinity during each of three or four consecutive 12-hr. periods. The probability was intended to apply to any point in the approximately 7000 square miles of Chicago and vicinity, although possibly more attention was paid to the somewhat smaller highly urbanized area. The official station, Midway Airport, was used for verification. The forecaster was free to choose his probability value as he saw fit, unencumbered by pressure to maintain time continuity of numbers, or to make the probability words released to the public fit the probability numbers recorded but not released.

Graphs such as figure 10 were prepared monthly for the first year or so and then less frequently. Some written discussion of the results was given to the forecasters in each of the first few months (then less frequently) to make them aware of their weak points and limitations and to thrash out problems that arose, as discussed here earlier. After about 6 months, when a reasonably large sample had accumulated, a time breakdown was made of each forecaster's forecasts, so he could evaluate his personal bias for each time period.

The Brier score with climat considered was used for verification, but the implications of this score were not fully understood at the beginning, and some poor forecasts probably resulted as forecasters attempted to maintain reliability without realizing that this hurt their total score and the economic worth of their forecasts. Winter snow flurries were particularly troublesome, because of the high frequency of trace amounts, as has been mentioned earlier.

The forecasts were determined an hour or so later than if they had been released to the public, but this should have no significant effect, except possibly in the first half of the first period forecasts.² Both three and four consecutive 12-hr. periods were used, with the forecast period extending farthest into the future ending about 56 hr. from the time the probability was recorded, or 60 hr. from the synoptic map upon which the forecast was based.

²Only half the first period forecasts are involved, because half were recorded about 2 hr. before the start of the forecast period, and half were recorded about 8 hr. before, and only the first group would likely be affected.

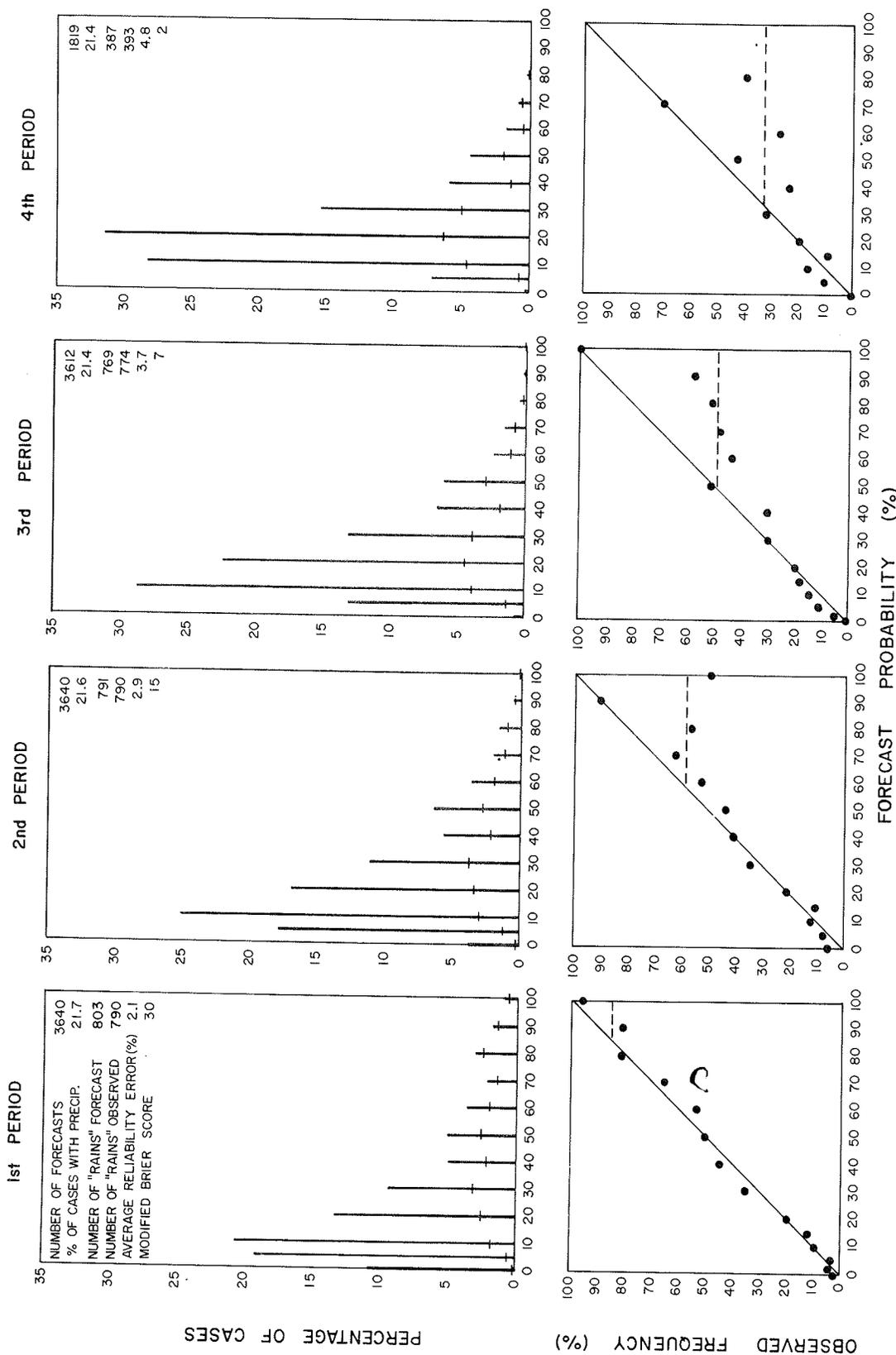


Figure 10. - Time period breakdown of probability forecasts of figures 3 and 4. The periods are successive 12-hr. periods, with the first starting about 2 or 8 hr. after time of forecast.

Figure 10 covers the first 2 1/2 yr. of the experiment for each of the four periods.³ The graphs for all forecasts combined were given earlier as figures 3 and 4. A number of things can be seen from figure 10. Let us consider the lower set of graphs first.

For all forecasts (see also fig. 3) the average weighted error from perfect reliability is 2.6 percent. If the error is given an appropriate sign, the algebraic average becomes almost 0 (actually 0.1 percent.) If the reliability had been perfect or if the average error when sign was considered had been 0, the forecasters would have forecast exactly the same number of rains as occurred. The number of rains forecast can be computed by considering that, for example, a 40 percent forecast is a forecast of 0.4 of a rain. Such a computation yielded 2747 forecast "rains" in the sample, with 2750 rains observed — almost perfection on this point. The number of rains forecast is very close to the number observed in each of the four periods as well. The number of actual 12-hr. periods with measurable precipitation was 392, with exactly as many night periods as day periods with precipitation.

The lower set of graphs of figure 10 also shows the time deterioration of the forecasts. As time progresses, one can see that a straight line fitted to the points rotates around roughly the climat value, and the very low and very high points are gradually eliminated. The main loss in reliability was in the high probability values, and this indicates that the rate of elimination of the high values as the forecast period reached farther into the future was not rapid enough. Even so, the average reliability error increased only from 2.1 percent in the first period to 4.8 percent in the fourth period.

Brier [3] discussed the economic effect of reliability errors. His argument is that the only users affected by reliability errors are those whose cost-loss ratios lie between the forecast probability and the observed frequency for that group of forecasts. Thus small errors affect only a small number of users, and if the forecast probability and the observed frequency are both on the same side of climat, even those that are affected get some gain over the use of the climatological frequency as a forecast. Thus the errors in reliability shown in figure 10 are of concern only in the high values and the later periods; the following discussion suggests how these can be eliminated.

The horizontal dashed line on the bottom graphs gives the average rain frequency for all forecast probabilities higher than the value of the line (in the fourth period, the 30 percent value was also included). This line represents the approximate percent upper limit of the forecast ability of the Chicago forecasters for each period, for the following reason. Assume a forecaster decided that a particular third period had a forecast probability of at least 50 percent, but instead of his choosing which value (50 percent

³On the bar graphs, the 2 percent and the 15 percent bars are omitted, as forecasts with these values were made only during the last 6 months and thus their frequencies would be unrepresentatively low. These values are retained in the dot graphs, however.

or more) to record, one of these higher values was chosen randomly. In such a case, the points for each of the probabilities of 50, 60, 70, etc., would fall on the 49 percent line for this sample. Thus the probabilities higher than 50 percent in the third period would be little better than a random selection of the higher numbers, and thus little skill would be shown in distinguishing one from another. Skill is shown in selecting the days as higher than average rain days, but the best choice of a probability would have been 50 percent, with no higher values. This suggests how the set of forecast probabilities should be reduced from the total set, as time increases. Based on the above reasoning, the following set was suggested for Chicago, starting after the end of the second year.

1st period - All values (listed earlier)
 2d period - 02 through 70
 3d period - 05 through 50
 4th period - 10 through 40

These ranges were not mandatory. Should particularly dry or particularly wet conditions prevail, a downward or upward shift respectively would be in order. When all these sets are considered, there are on the average just about as many values above climat as below climat, in keeping with the point discussed much earlier.

The set of probability values available to the forecasters during the first two years of the experiment did not contain the values 2 and 15. While possibly more of academic interest than economic value, the value 2 was felt desirable because some forecasters were reluctant to use the 0 value, as they were rarely certain that there would be no precipitation, even though they felt the next higher value, 5, was too high. Perhaps rightly so, they were not persuaded to use the 0 value for a probability not truly 0 but much less than 5 percent, i.e., they would not accept a verification even as low as 1 percent for the 0 value.

Also, the verification bore out the idea of using a value of 2, because on graphs such as those of the bottom row of figure 10, it was noticed, in the first and second periods, that the 5 percent point fell below the line whereas the 0 point fell above the line. This meant that some points from the 5 value could be shifted to a lower value, say 2, and, if done well, it would improve the verification of the 5 value (make it more reliable). Similarly, some points from the 0 value could be shifted to 2 and improve the 0 value. Thus the 2 value fulfilled a need that was reflected in the verification. No such argument was applicable to the use of the 15 value (or any other value), but it was felt that in the later forecast periods, when the values were crowding more nearly to the climat value, finer breakdown near the climat value was needed. Possibly this should have included a 35 and/or a 25 value as well, but these last two values were not used.

The upper set of graphs in figure 10 gives the frequency distribution of forecast probabilities by time periods. All graphs show the single mode near the climatological frequency and low values at both extremes. Notice how few 100's there are compared with 0's and how far the ratio of 100's to 0's is from the 1 to 5 suggested by climatology. In the first period the ratio is 1 to 16 while in the third period it is 1 to 70. It is evidently

much harder than would be expected by climatology to be sure of rain than to be sure of fair (no rain). This is probably because of the spotty nature of many rains in the Chicago area (point probability must be less than the area probability) and the widespread nature of fair weather.

The number of forecasts in the fourth period is only half that of the other periods because this forecast is made only every other shift. The number of forecasts in the third period is a bit lower than in the first and second periods because forecasts for this period were not made on one of the forecast shifts during the first month of the experiment.

The small horizontal line on the bars in the upper set of graphs gives the frequency of precipitation in that category. Only about one-third of the rains occur with forecast probabilities of 40 percent or more. This is partly because so few forecasts of these higher categories are made in the longer forecasts, and partly because the low areal coverage of precipitation forces low probabilities even in the early forecast periods, although the uncertainty of position and timing must contribute significantly too. In the first period 60 percent of the rains occur with probabilities of 40 percent or more, while the figure is 22 percent for the fourth period, and eventually would necessarily decrease to 0 as the forecast period extended more into the future and a value as high as 40 percent could no longer be forecast.

The score given on figure 10 (and on fig. 3) is the Brier score with climat included (total score method). The score for all forecasts appears to be a stable value, as it was essentially the same in the second 12 months of the experiment as in the first 12 months.

The time deterioration of the score of the forecasts is given in figure 11. Rather surprising is the great reduction from the first half to the second half of the first period, brought on by an additional 6 hr. of lead-in time at the 10 a.m. and 10 p.m. forecast time. Evidently even a few hours is very important in the quality of the forecasts for this first period. This definitely indicates that it should benefit the forecaster to wait just as long as possible before releasing a forecast for a time period close at hand. It also suggests that the additional time and detail available to a field station are likely to provide a better first period forecast than that produced by a center such as NMC.

Figure 11 also provides a basis for determining how far into the future forecasts should be made available to the public. The last two points on the figure show no appreciable skill over climatology, and the third from last shows little skill. Elimination of these three points would allow a forecast always to extend through the tomorrow period and consist of two or three probabilities in each forecast, depending on the time of day issued.

A common question concerning forecasts is "What is the percent correct?" This question can be answered concerning probability forecasts and this is possibly the only way to consider a single probability forecast. A 100 percent forecast on a rain period or a 0 forecast on a no-rain period are the only forecasts that are completely correct. All other forecasts are only partially correct. For example, a 70 percent rain probability forecast is 70 percent right (and 30 percent wrong) for a period in which it rains, while

it is only 30 percent right (70 percent wrong) for a period in which it doesn't rain. The average percent correct for all Chicago forecasts is 72.3 percent. While such a score is interesting, it could not be used by itself as a measure of skill, because a constant forecast of 0 percent will yield a score equal to the climatic frequency of no rain (about 78 percent for Chicago).

Another subsidiary figure of some interest is the percentage of times the forecaster deviated from climatology in the right direction, i.e., had a higher forecast probability on a rain day and a lower one on a no-rain day. This figure for the Chicago group for all forecasts is 74 percent.

A final point in the overall verification of the Chicago forecasts concerns whether or not the forecasters are transferring to the probability forecasts the amount of skill they have previously demonstrated in categorical forecasts.

In categorical forecasts the number of rains forecast is usually very close to the number of rains observed. This is desirable, and would be so in perfect forecasting. Categorical forecasts can be converted to probability forecasts which have the same number⁴ of rains observed as forecast by plotting two points with perfect reliability. These two points could help experienced forecasters shift skill from categorical to probability forecasts, by taking the higher value — the percent correct on an expected rain day — as the first approximation to the probability for a rain day, with subsequent modification according to the current situation. The lower value would likewise be the first approximation on a no-rain day.

Categorical forecasts continued to be made at Chicago while the probability experiment was in progress. The comparable categorical forecasts were made only once a day and for only the second and fourth of the eight periods given in figure 11. These categorical forecasts had essentially the same number of rains observed as forecast, as did the probability forecasts for the same periods. However, when these probability forecasts were converted to categorical forecasts, with a rain forecast for each probability forecast of 50 percent or more, the number of rains forecast dropped by about 25 percent. To convert these probability forecasts to categorical, using a separating criterion that caused the number of rains forecast and observed to be

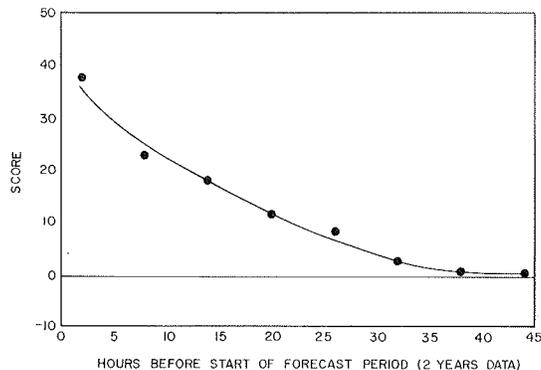


Figure 11. - Time variation of the modified Brier score with climat considered for forecasts of figure 10.

⁴The probability for one point is determined from the ratio of the number of correct precipitation forecasts to the total number of precipitation forecasts, while the other probability is 1 minus the similar thing for the no-precipitation forecasts.

equal, would require the dividing line to be about 40 percent instead of 50 percent.

It would appear, therefore, that the probability forecasts did not carry over the skill demonstrated in the categorical forecasts. Table 1 gives additional information by comparing the converted probability forecasts using the 50 percent dividing line (CP) with the comparable categorical forecasts made at the same time (CC), and with all (19 mo.) of comparable categorical forecasts made in the period just prior to the experiment (BC).

The observed rain frequency was about the same in the 19 mo. prior to the experiment (22.6) as during the experiment (21.6) so the periods should be comparable. The frequency rain was forecast in the converted probability forecasts (16.4) was much lower than in the two sets of categorical forecasts (22.5), where the latter about equaled the observed rain frequency. The chance skill score was lower for both types of forecasts during the experiment (0.33), than for the earlier forecasts (0.40). The percent correct of the no-rain forecasts was essentially the same in the three sets of forecasts, although a bit lower during the experiment. The percent correct of the rain forecasts was much lower on the categorical forecasts during the experiment (46.6) than before (54.6), while for the converted probability forecasts it was only a bit lower (52.6). All this suggests that the categorical forecasts were poorer during the experiment than before, and while the probability forecasts made up some of this deficiency, the probability forecasts still did not have the full skill of the earlier categorical forecasts.

The probability forecasts would be about equal in quality to the earlier categorical forecasts if the number of rains forecast and observed had been about equal, as the others figures are comparable. To make these figures more nearly equal, the forecasters would have had to shift some of the 40 percent forecasts, and possibly some 30 percent forecasts, to 50 percent or more.

TABLE 1. - Comparison of Probability and Categorical Forecasts (percent)

	CP	CC	BC
Rain forecasts (percent correct)	52.6	46.6	54.6
No rain forecasts (percent correct)	84.6	85.8	86.2
Observed rain frequency	21.6	21.6	22.6
Forecast rain frequency	16.4	22.5	22.6
Chance skill score	0.33	0.33	0.40

But in order to maintain the percent correct of the rain forecasts, the shifted forecasts could not be a random selection of the 40 and 30 percent forecasts, but must be a group which verify over 50 percent observed frequency. Such a shift would be quite difficult to make, and would, of course, lower the rain frequency in the remaining 40 and 30 percent forecasts. Figure 10 shows that this might be possible, since the 30 and 40 values in the first and second periods average above the line and thus could be improved by the removal of more rain cases than no rain cases.

From the above it would appear that when the forecasters felt that a categorical rain was the best forecast, they sometimes became a bit more conservative in the probability forecasts and recorded a value of 40 percent instead of one of 50 percent or more. The availability of the many values in probability forecasting allowed this conservatism, but it should probably be resisted or the quality of the probability forecasts will suffer by not reflecting the skill previously demonstrated in categorical forecasting.

Seasonal Variations

The seasonal variations proved to be minor, but a few points can be mentioned. The seasons were defined with winter as December-January-February; summer as June-July-August, etc. The summer and fall had an observed frequency of precipitation of about 18 percent (which is below the yearly average), and the forecasts for each of these two seasons called for about 8 percent more rains than actually occurred. The winter and spring seasons averaged a precipitation frequency of about 26 percent each (above the yearly average), while the forecasts called for fewer rains than occurred, with about 2 percent fewer in winter and over 11 percent fewer in spring. Thus, the forecasters were undercompensating for the seasonal variations.

Part of the large error in spring is no doubt a result of starting the experiment in the spring, because the first two of the 8 months tabulated for spring contained well over half the 0 forecasts, and they had an average observed frequency of over 11 percent. This overconfidence of the forecasters in the no-rain condition was soon overcome, however. The main error in the spring forecasts came from low forecast values which were too low, while the main error in the summer and fall came from high forecast values which were too high, the latter suggesting insufficient reduction for the reduced areal coverage of shower type rains. This overconfidence of the low values in the high frequency season and overconfidence of high values in the low frequency season is likely to be a characteristic of forecasters at first and is a group bias that knowledge and time will reduce.

The ratio of the number of 0 forecasts to 100 percent forecasts is about 12 to 1 for the high frequency winter and spring seasons, and would probably drop to about 8 to 1 if the many unwise 0's from the first 2 mo. of the experiment were more reasonably proportioned. On the other hand, during the low frequency months of the summer and fall season, the ratio was about 80 to 1. The climatological frequency would suggest about 4 to 1 for the "wet" seasons and 5 or 6 to 1 for the "dry" seasons. The ratio in the wet seasons

is thus off by a factor of 2, while that in the dry season is off by a factor of about 15. The difference between these is probably due to the differences between point and areal frequency of precipitation.

Words vs. Numbers

An attempt to find the relationship between the forecaster's choice of a probability number and a probability word was made using the first 2 yr. of data from the experiment. The forecasters were not informed that an attempt was being made to find such a relationship, although they could have become aware of it through general conversation. Also, the forecasters received no instructions to make the word and number forecasts consistent.

Such a relationship is hard to define because of the large variety of words and word combinations used by the forecasters, and grouping was necessary. Four groups were finally selected; namely (1) no precipitation, (2) chance of precipitation, (3) precipitation likely, and (4) unqualified precipitation. The "no precipitation" category gave no problem, as it was pure and large. The other categories were purified by selecting only those forecasts which contained, for example, the word "chance" without other modifiers of any type. While there were nearly 5500 forecasts in the no-precipitation category, the relatively low frequency of precipitation and the purification process left only about 350 to 500 forecasts in each of the other categories. However, the results given in figure 12 are mainly reasonable.

In the no-precipitation category, there are a few forecasts that are too high to be reasonable. No certain reason for this was found, but it may result from the fact that the worded forecast was prepared about an hour before the number forecast, or from the forecaster's being aware of the location of the verifying gage, or, in the first period, being able to distinguish between different parts of the city, or making a poor forecast because of misunderstanding the verification system. At any rate, the forecaster was clearly able to make a significant breakdown of the no-precipitation forecasts into number categories of 40 percent and less, with high reliability. Since the number of forecasts in this category represents well over half the forecasts issued, these not-released number forecasts represent considerable information not communicated by the forecaster. This in itself would appear to be a strong argument for the issuance of probability numbers.

In the other categories, the range of values selected is also surprisingly high, while the reliability, except for the "likely" category, is quite high. This range of values is another strong reason for the use of numbers, as the words are just not conveying a consistent meaning.

The surprisingly large range of values in all categories is hard to explain. Part of it may be that the worded forecasts sometimes express the area probability and sometimes the point probability. Another reason could be that the forecaster used the words differently, perhaps more casually, in the later periods compared with the earlier periods. This latter was tested by making a time breakdown of the categories into the four 12-hr. forecast periods. The test (see table 2) was rather inconclusive for all except the no-precipitation category, because of the small sample size, but there was

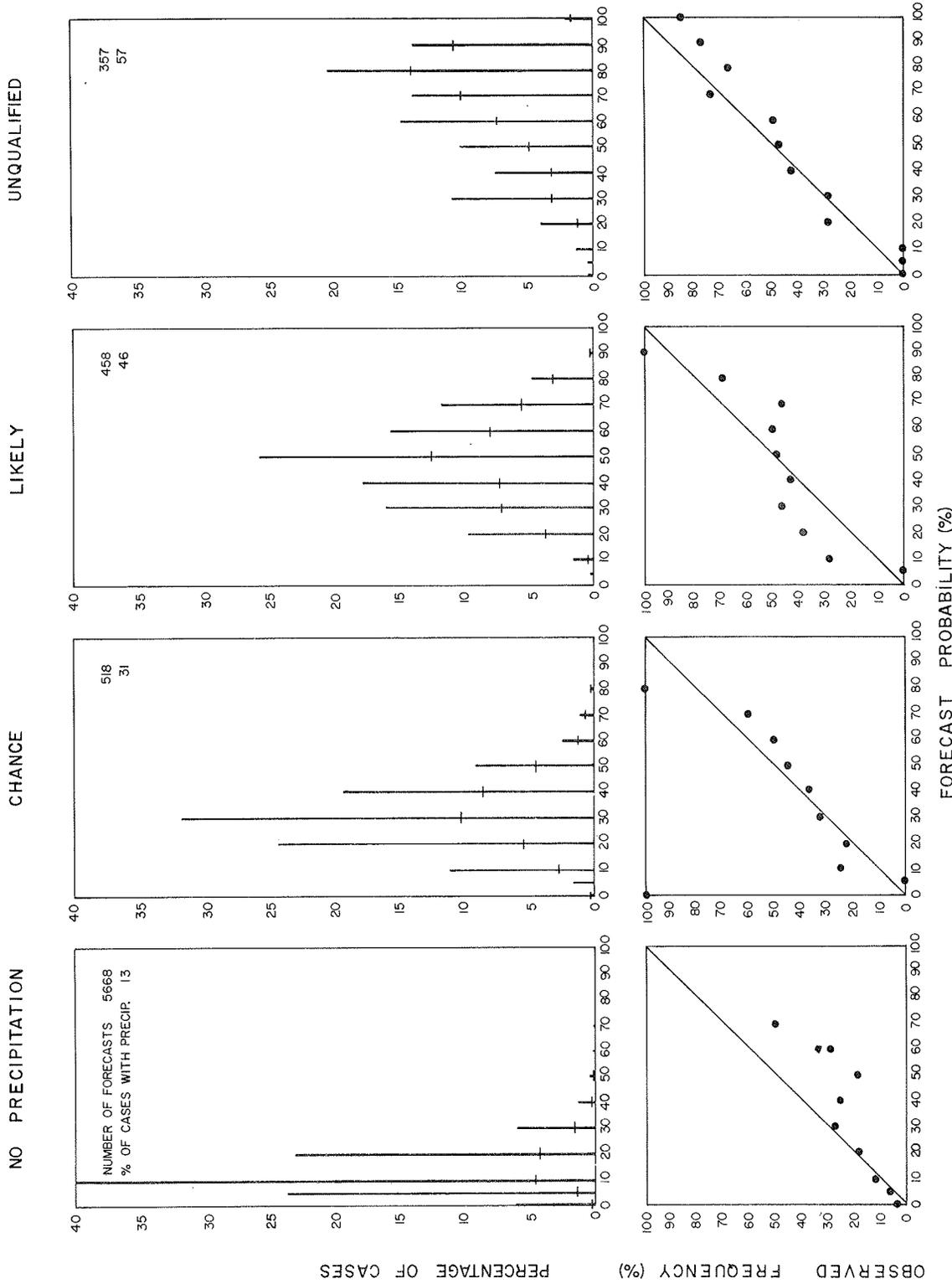


Figure 12. - Similar to figure 10, but for a 2-yr. sample of groups of forecasts selected according to the probability word used in the forecast.

TABLE 2. - Percent of Rain Cases by Categories

Category	F o r e c a s t P e r i o d				
	1	2	3	4	All
No Rain	6	12	16	18	13
Chance Rain	29	26	28	34	31
Rain Likely	52	43	49	44	46
Rain Unqualified	67	60	38	38	57

indication that unqualified rain was defined by the forecasters as a higher number in the first period than in the fourth, with the opposite condition in the no-precipitation category. The chance and likely categories did not show any significant time variations. The likely category is different from the others in that it verified about 50 percent no matter what probability number was used by the forecaster. There is no obvious reason for this.

The percentage of precipitation cases in each category is reasonable, with 13 percent for the no-rain category, 31 for chance, 46 for likely, and 57 for unqualified, as given in table 2.

The detailed time breakdown for the no-precipitation category (not shown) was representative because the sample size was large. The reliability showed slow deterioration but was fairly good even in the fourth period, where 5 percent verified 12, 10 percent verified 15, 20 percent verified 19, and 30 percent verified 26 (this was the bulk of the forecasts).

8. SUMMARY AND CONCLUSIONS

Probability forecasts in numbers superior to probability forecasts in words can be made by experienced forecasters without objective aids. These forecasts will yield precipitation in the frequency promised, that is, they will be reasonably reliable, but on the other hand they may not promise much, because of low resolution.

Point probability of precipitation is likely of more value to the vast majority of users than area probability, and herein lies much of the cause of the low ability to forecast precipitation. The reason is that when there is precipitation in an area such as Chicago and vicinity, the point probability equals the areal coverage, and therefore the maximum point probability may be far from 100 percent in shower situations. Thus when comparing

forecasts from different stations, in addition to considering the difficulty of the problem of the development and movement of weather systems, and the climatic frequency of precipitation, one must also consider the differences in the average space coverage and average time span of precipitation.

The exclusion of trace amounts as a rain occurrence causes the forecaster to be confronted with a nearly impossible forecast situation all too frequently, as at Chicago precipitation amounts of trace and 0.01 in. in 12 hr. occurred in more than 50 percent of the periods receiving trace or more. Using a trace as a precipitation occurrence would probably ease the forecaster's problem and improve the forecast score, but may adversely affect the value of the forecasts to the user. A forecast giving the probability of various precipitation amounts seems to be the only way out of this problem, and it is felt that this could be done with significant skill.

The use of predetermined forecast periods creates a problem not inherent in the forecast situation, but this can be overcome to some extent by using a variably positioned forecast period in the earlier part of the forecast.

The set of forecast values used should be rather limited for psychological reasons, and, for a small group of forecasters, in order to get a large enough quantity in most forecast categories for a representative verification. The set should be selected considering the climatological frequency, such that about half the values are above climat and half below. The set should decrease as the forecast period reaches farther into the future. This decrease is accomplished at a station with a low climatological frequency of precipitation by gradually eliminating mainly the high forecast values. For the Chicago forecasts, the upper limit of forecast probability reduced to about 50 percent for the 12-hr. period starting about 30 hr. from the time of the synoptic map upon which the forecast was based.

The Brier score is easy and fast to compute by desk calculator or slide rule and is adequate to rank forecasters (or groups of forecasters) who forecast under similar climatic conditions considering both the frequency of precipitation and the spotty nature of the precipitation or its average time span. This score is ideally suited to a climatic frequency of 50 percent, but a frequency this high is uncommon in the United States.

The Brier score with an effect of climatology included is given in the form

$$S = 100 \left(\frac{B_c - B_f}{B_c} \right)$$

where B_f is the Brier score of the forecasts, and B_c is the Brier score of "climatology". One method computes the score by subsets and averages. This method has the advantage of having a clear relationship of skill to the distance the forecast is moved away from climat, but it requires about three times the effort necessary to compute the simple Brier score. In the second method, the total score of the forecasts B_f and of climat B_c are computed before dividing. This method requires but little more computing than the simple Brier score, but loses some of the clarity in relating skill to climat. If climat values for portions of the year deviate by as much as 10 percent

from that for the whole year, the score should be computed by frequency "seasons", no matter which scoring method is used.

These scores compensate for the different frequency of precipitation among regions, and thus allow better comparison of scores under such conditions. However, they do not adjust for variations in the areal coverage of precipitation nor for variations in the time span of precipitation, and therefore they only partially compensate for climatic variations among stations.

For a sufficiently long period, the B_c score in the total score method approaches a constant equal to $C(1 - C)$. This constant can be computed from knowledge of only the climatic frequency (C) and thus when comparing scores among stations it could be used to partially adjust for climatic variations. Such use has some value even if the subset method is used to compute the score. The score is relative to a commonly accepted no-skill base, and the Chicago forecasts showed very little skill above this base for periods beyond 36 hr. from the map time upon which the forecast was based.

A final advantage of probability forecasts in numbers is the more precise verification possible. Such a verification soon overcomes the usual over-confidence of forecasters by showing them the limits of their ability. They thus promise no more than they can deliver, and this, coupled with the fact that the user of the forecast can also easily check on the relationship of what is promised vs. what is delivered, will make the user more likely to treat the forecast with respect. Obviously maximum acceptance of such forecasts by the public must come from their education in the need for probability forecasts. Basically the need arises from the non-cyclic nature of weather systems, leading to errors in placement and timing, uncertainty on whether or not a weather system will be born, or survive, and the inability to forecast the exact position of the spotty (shower) type precipitation. Finally one might say that while forecasters can't forecast well the occurrence of precipitation, they can forecast quite well the chance of precipitation occurrence.

ACKNOWLEDGMENTS

The author is indebted to Mr. J. F. Hoehn, former research assistant, for his ideas, helpful discussions, and many of the computations, and to Mr. J. R. Fulks, Meteorologist-in-Charge, Chicago, for a careful review of the manuscript. Particular thanks are due to Mr. J. E. Hovde, Supervising Guidance Forecaster, for his many helpful and sometimes heated discussions of many aspects of this paper, particularly the sections dealing with verification scores, forecast periods, and forecast values.

REFERENCES

1. G. W. Brier, "Verification of a Forecaster's Confidence and the Use of Probability Statements in Weather Forecasting," U. S. Weather Bureau Research Paper No. 16, 1944, 10 pp.
2. G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," Monthly Weather Review, vol. 78, No. 1, Jan. 1950, pp. 1-3.
3. G. W. Brier, "The Effect of Errors in Estimating the Probabilities on the Usefulness of Probability Forecasts," Bulletin of the American Meteorological Society, vol. 38, No. 2, Feb. 1957, pp. 76-78.
4. J. J. George, Weather Forecasting for Aeronautics, Academic Press, New York and London, 1960, 673 pp.
5. I. I. Gringorten, "Methods of Objective Weather Forecasting," Advances in Geophysics, vol. 2, 1955, pp. 57-92.
6. H. E. Root, "Probability Statements in Weather Forecasting," Journal of Applied Meteorology, vol. 1, No. 2, June 1962, pp. 163-168.
7. F. Sanders, "The Evaluation of Subjective Probability Forecasts," Scientific Report No. 5, Department of Meteorology, Massachusetts Institute of Technology, June 1958, 57 pp.
8. F. Sanders, "On Subjective Probability Forecasting," Journal of Applied Meteorology, vol. 2, No. 2, Apr. 1963, pp. 191-201.
9. M. J. Schroeder, "Verification of Probability Fire-Weather Forecasts," Monthly Weather Review, vol. 82, No. 9, Sept. 1954, pp. 257-260.
10. W. Weaver, "Probability: The Odds are High that it Affects You," Think, vol. 27, No. 4, Apr. 1961, pp. 3-6.

NOAA CENTRAL LIBRARY
CIRC M(055) US87ce no.3
Hughes, Lawr. On the probability forecasts



3 8398 0006 8158 9