Eighty-Five Percent and Holding—A Limit to Forecast Accuracy?
Daniel L. Smith
National Weather Service Southern Region, Ft. Worth, Tex 76102

Abstract

The significance of Percent Correct Scores for National Weather Service (NWS) probability of precipitation (PoP) forecasts is examined. It is shown that the areal variability of rainfall and the nature of PoP forecasts preclude the achievement of a score of 100%—even for the best possible forecasts. A maximum possible percent correct is defined and radar estimates of rainfall coverage are combined with actual forecasts to determine how closely NWS forecasters approached this limit. Day- and nighttime percent correct scores were 75% and 85%, respectively, for the data examined. These values were close to the respective maximum possible scores of 83% and 90%. Relatively small changes in forecasters' percent correct scores are considered in light of these findings.

How good are weather forecasts? This is a simple question with a not so simple answer. Nevertheless, it is a question we are obliged to answer because forecast improvement depends on reliable measures of forecast accuracy and individual forecaster skill. Verification is complicated by the many kinds of forecasts that are popularly grouped as “weather forecasts” and by the variety of measures that can be applied to assess their accuracy. For example, while it may be rather easy to determine that, on the average, a set of forecasts were accurate to within 3°F, it is more difficult to determine the significance of the forecast skill that achieved this result. For most users, certainly, such an error is much less important at 75°F than it is at 32°F. Other “forecasts,” such as severe weather warnings, are not particularly amenable to simple numerical evaluations but require careful verification nonetheless because of their high impact on the user.

A frequent simple answer to the introductory question is that “... forecasts are about 85% correct ...” Percent Correct is the frequency at which it rained on “rain” forecasts and did not rain on “no rain” forecasts. Cooley (1972; personal communication, 1978) showed that for a large number of locations nationwide, NWS forecasts are about 85% correct if PoP forecasts are converted to “rain” and “no rain” forecasts by using 50% as a threshold value. Percent Correct for NWS forecasts has increased since 1959 when it averaged about 75–80%, depending on forecast lead-time. But what is the significance of this seemingly high degree of skill and what does the relatively small increase over the past 20 years imply in terms of further improvement? We will show in this brief note that an additional increase of 15%—to 100 Percent Correct—should not be considered the ultimate forecast goal. A plateau below this level must exist for probability forecasts and further improvements of only a few percent are much more significant than they might seem.

There are a variety of other measures of forecast accuracy (and forecaster skill)—some of much better diagnostic utility—but experience has shown that Percent Correct is easily understood by the layperson. As it happens, the Percent Correct usually “looks” quite high, but one can quickly appreciate that this is because it generally rains only infrequently and it is much easier in such climates to forecast “no rain” than to forecast “rain.” But the NWS, at least, does not actually forecast “rain” or “no rain”—instead, probability forecasts are used to express the likelihood of precipitation at any given point. While each community (point) may want a “rain” or “no rain” forecast, dissemination facilities (if nothing else) will generally not allow such forecasts for all points. Thus, the NWS issues zone forecasts, for “meteorologically homogeneous” areas roughly the size of a few counties, which express the likelihood of rain at any given point within the zone. Unless rain is considered a certainty at all points within the zone a probability of 100% can not be used. *It is this fact which leads to a limit to “skill” at a level below 100%*. NWS probability forecasts are generally converted to categorical “rain”/“no rain” forecasts by choosing 50% as a threshold value: a probability below 50% is considered a forecast of “no rain” for any chosen verification point in the zone. We will follow this convention in the remainder of this text, although it should be noted that other thresholds have occasionally been proposed by others.

Some will argue that it is only the present state of the system, rather than the possible state of the science, that imposes a limit to skill. It is worth considering for a moment, however, that categorical forecasts—if they are less than 100% correct—suffer in terms of utility by not conveying to the user some measure of uncertainty. This fact has been explored by many authors. It seems obvious that forecasters presently do not have the skill to always forecast 100% when it rains at a point and forecast 0% when it does not; any forecast service that endeavors to provide such a product should be called upon to provide proof of resolution that exceeds that of conventional probability forecasts for the same point. Nevertheless, it can at least be argued that future improvements in human and machine capabilities theoretically may usher in a day when such skill can...
TABLE 1. Percent Correct comparison for Alabama Zones 15 and 16, June to August, 1976.

<table>
<thead>
<tr>
<th></th>
<th>Maximum Percent Correct</th>
<th>Percent Correct (Forecasters)</th>
<th>Percent Correct (Climatology)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day (12-00 GMT)</td>
<td>83 (82)</td>
<td>75</td>
<td>67</td>
</tr>
<tr>
<td>Night (00-12 GMT)</td>
<td>91 (88)</td>
<td>85</td>
<td>82</td>
</tr>
</tbody>
</table>

≥50% when coverages are ≥50% and for PoPs to be <50% when coverages are <50%. The former accuracy, however, is the only way to achieve the best possible Brier or Probability Scores.

What is the MPC—or, at least, what are characteristic values? Since the MPC is linked to the climatological variation of rainfall coverage it must vary from region to region and probably varies with season in any given region. At “85%” how close are we to the MPC? Are we perhaps there? The MPC approaches 100% in a climate where rain frequency approaches zero. However, in a climate where daily scattered showers cover 30%-70% of the area and extremes of coverage are rare, the MPC ~ 50%. An accurate determination of the MPC can only be made if we have knowledge of the areal coverage of rainfall. Such determinations seem to be scarce, but Smith (1977) collected a data set that lends itself well to determination of the MPC for a near-coastal regime in summertime. For the eight forecast zones shown in Fig. 1, radar estimates of rainfall coverage were made for 12 h day- and nighttime periods during June, July, and August 1976. Rainfall observations from Dothan and Mobile in Alabama Zones 15 and 16, respectively, provide rainfall frequency verification and show that radar estimates can accurately reveal coverage of rainfall. For a full discussion of the data see Smith (1977).

Table 1 shows summer day- and nighttime values of the MPC as estimated from radar data for Alabama Zones 15 and 16. The MPC was determined by summing all coverages ≥50%, adding the complement of all coverages <50%, then dividing by the number of observations; or

\[
MPC = \frac{\sum C_A + \sum (100 - C_B)}{(N + n)}
\]  

where \(C_A\) is the number of rain occurrences on periods with coverage ≥50%, \(C_B\) is the number of no rain occurrences on periods with coverage <50%, and \(N\) is the total number of observations; or

\[
MPC = \frac{A + B}{(N + n)}
\]  

where \(A\) is the number of rain occurrences on periods with coverage ≥50%, and \(B\) is the number of no rain occurrences on periods with coverage <50%. If coverage estimates are accurate and the chosen raingage is representative of the area, then Eqs. (1) and (2) yield the same result. In fact, for the Alabama data set they
were very close—the parenthetical MPC values in the table are from Eq. (2), using observations from Mobile and Dothan.

Also shown in the table are Percent Correct scores from forecasters' first period (00-12 h) probability (PoP) forecasts. It can be seen from the table that forecasts for the summer of 1976 were close to the "85%" level; a little lower during the daytime "scattered shower" regime, as we might expect. Of critical importance, however, is that during the daytime the Maximum Percent Correct is only about 83%. If this value is "scaled" to 100% by multiplying by 1.205 and the forecasters' Percent Correct (75%) is adjusted likewise, we see that forecasts were within 10% of "perfection"—that is, they achieved 90% of the possible "skill." An adjusted Percent Correct (for those who like to think in terms of 100% as the upper limit) shows that these forecasts were, in fact, better than 85% correct! If we take (from the table) a nighttime MPC of 90%, the forecasters' "adjusted" Percent Correct was 94%. It has generally been conceded that the "last 15%" would be hard won as forecasts improved beyond the 85% level. It appears that we might well be farther along on the road than we realized—no wonder the going is tough!

Even though the Alabama forecasts look quite good, a measure of "skill" is still lacking. After all, if it never rained it would be easy to achieve the MPC of 100%. A benchmark is provided by Percent Correct for climatological "forecasts," shown in the table. Since 12 h day- and nighttime rain frequencies for the area are <50%, climatological "forecasts" are always "no rain"; thus the climatological percent correct is simply the complement of the rain frequency, measured at a representative point in the area. Graded for improvement over climatology the Alabama forecasters scored quite high. The total range between skill of climatology and Maximum Percent Correct for summer daytime forecasts in the study area was only 16% (83%-67%). At 75%, forecasters fell exactly halfway between the limits. Thus, an apparently small percentage improvement over climatology (from 67% to 75%) is seen in its proper perspective when considered against "how far we have to go"! More importantly, if the summer of 1976 can be considered representative, it can be seen that small changes (1%, or so) in the Percent Correct score represent significant changes in skill because the range of skill is only about 16%, not 33% (100%-67%) as might be expected. For nighttime forecasts in the study sample the total range of skill was only 8% (90%-82%), indicating that small changes in Percent Correct are even more significant. Figure 2 illustrates the concept of a limit below 100% imposed by the MPC. Forecaster "skill" approaches the MPC with time. It is an interesting prospect that at some future time forecasters may be able to specify areal coverage perfectly, thus achieving the point A, which represents perfect probability forecasts for points within an area. Given a homogeneous distribution of showers in an area, one must know the distribution (coverage) of showers in order to specify with precision that a particular point will have rain. Thus, point A in Fig. 2 must be reached before point B. Perfect PoP forecasts, in other words, will come before perfect categorical forecasts for individual points. We will not speculate on when point A will be reached!

The literature contains conflicting, but generally somewhat unimpressive, evidence that only slight increases in Percent Correct have been achieved in recent years. It is likely that by combining data from diverse geographical areas and for many seasons we have diluted the evidence that indeed significant achievements have been made. Moreover, a true assessment of our relative progress is hampered by an implication that 100% is our ultimate goal. The geographical and seasonal variability of rainfall, and the attendant variability of the MPC, argues for regional determination of the Percent Correct (and, for that matter, all scores that are based on a conversion of PoP forecasts to categorical forecasts). Next, without great difficulty, scores could be "normalized," allowing comparison of forecaster skills for such meteorologically different areas as Birmingham and Los Angeles. Most importantly, a true evaluation of past and present states of forecaster skill can lead to further improvement in the future.

Acknowledgment. This brief study is based largely on conversations between the author and Paul Moore, Chief of Scientific Services Division, NWS Southern Region.

References


