NOTES AND CORRESPONDENCE

Chance Behavior of Skill Scores

UWE RADOK

Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado

26 February 1987 and 1 August 1987

ABSTRACT

The skill score \( S = \frac{R - E}{T - E} \) (representing \( R \) actual and \( E \) expected successful categorical forecasts in a total of \( T \) forecasts) remains a valid tool for assessing the overall quality of current probabilistic long-range forecasts, which start from categorical subdivisions of the forecast area. The skill score definition is modified to become a chi variate with one degree of freedom. Two sets of skill scores computed from forecasts of U.S. monthly precipitation and mean temperature are shown to have frequency distributions of similar shape with nonzero means and standard deviations generally corresponding to smaller independent numbers of verification points than those actually used. The largest skill scores of those examined were obtained for recent precipitation forecasts during a period when forecasts using only climatology were similarly skillful. This suggests that cooperation on part of the climate system remains an essential success ingredient in extended forecasting. A sequential procedure for monitoring the changing level of operational forecasting skill is described.

1. Introduction

The skill score has long been the standard tool for assessing the success of long-range forecasts, alongside more sophisticated methods involving pattern recognition (Somerville, 1977) and empirical orthogonal functions (Bettge et al., 1981). The new probabilistic forecasts which were started in midsummer 1982 (Gilman, 1986) add to the previous categorical subdivision of the forecast area extra lines indicating the forecasters’ level of confidence that not only the right categories have been picked but also the regions of the larger anomalies. The simple categorical choices however remain the first steps in the forecasts. Moreover they provide the basis for historical assessments of changes in forecast skill since the beginning of long-range forecasts. This makes it worthwhile to examine the fluctuations that could occur in the skill score by chance. For this purpose a theoretical sampling distribution for the skill score is derived in this note and compared with the distributions of two sets of operational skill scores.

2. Definitions

The most common skill score definition is based on a contingency table of the form shown in Table 1. Denoting the sum of the diagonal, \( \sum_{i=1}^{3} x_{ii} \), by \( R \) and its expected value by \( E \), the skill score \( S \) is defined as

\[
S = \frac{R - E}{T - E}.
\]  

Different definitions have been used for \( E \), the total number of successful forecasts expected to be obtained by chance. In the original skill score proposed by Panofsky and Brier (1963)

\[
E_i = \frac{\sum_j x_{ij} x_{j}}{T}.
\]  

This has been shown by Livesey and Skilling (1985) to minimize the information of the table (maximize its “Shannon entropy”). The main operational definition now in use by NOAA’s Climate Analysis Center is

\[
E_2 = 0.3(x_{11} + x_{33}) + 0.4(x_{22}).
\]  

This has the (largely cosmetic) drawback that it does not give a unique minimum value to \( S \); writing \( E_2 \) in the alternative form

\[
E_2 = 0.3(T - x_{32}) + 0.4x_{22} = 0.3T + 0.1x_{22}
\]

shows that for \( x_{22} = 0, E_{2\text{min}} = 0.3T, S_{2\text{min}} = -3/7 \); and for \( x_{22} = T, E_{2\text{min}} = 0.4T, S_{2\text{min}} = -2/3 \).

A modified form of \( E_2 \) that avoids this uncertainty and simplifies the theoretical results is

\[
E_3 = \frac{T}{3}.
\]  

Then \( S = 3R/2T - \frac{1}{2} \) and \( S_{\text{min}} = -\frac{1}{2} \), irrespective of the structure of the diagonal.

3. The sampling distribution of \( S \)

A statistical significance test appropriate for the full three-way contingency table in Table 1 with expected...
value $E_1$ is given in most textbooks and involves the 
chi square distribution with 4 degrees of freedom. 
However, this definition of $E$ places too much weight 
on the details of the failed forecasts which are of interest 
mainly for a detailed examination of the forecast 
method used. In the operational context the skill score 
especially compares two observed numbers $o_1, o_2$ with 
two expected numbers $e_1, e_2$. This comparison defines 
a chi square with one degree of freedom 
\[ x_{(1)}^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2}. \] (3)
Here $o_1 = R$, $o_2 = T - R$, $e_1 = E$, $e_2 = T - E$ so that 
\[ x_{(1)}^2 = \frac{(R - E)(T - R - E)^2 + E(T - E)}{E(T - E)} \]
\[ = \frac{R(T - E)^2}{E(T - E)}, \] (4)
and since from (1) 
\[ R - E = S(T - E) \]
\[ \Rightarrow \frac{T - E}{E/T} = S^2. \] (5)
For the special case of $E = E_3 = T/3$ this takes the 
simple form 
\[ x_{(1)}^2 = 2TS^2. \] (5')
Now the chi square distribution with one degree of 
freedom has the form 
\[ F(x_{(1)}^2)dx(x_{(1)}^2) = \left[ \frac{1}{2} \Gamma(1/2) \right]^{-1}(x_{(1)}^2)^{-1/2}e^{-x_{(1)}^2/2}d(x_{(1)}^2). \] (6)
With $\Gamma(\frac{1}{2}) = \pi^{1/2}$ and $x_{(1)}$ as variables in place of $x_{(1)}^2$ 
so that $d(x_{(1)}^2) = 2x_{(1)}dx(x_{(1)})$, and separating positive and 
negative values of $x_{(1)}$, this becomes 
\[ f(x_{(1)})dx(x_{(1)}) = \pi^{-1/2}e^{-x_{(1)}^2/2}d(x_{(1)}), \] (7)
the standard Gaussian distribution with zero mean and 
unit variance. According to (5), therefore, chance skill 
scores can be expected to be normally distributed with 
a variance that depends on the total and expected 
forecast numbers, viz. $\sigma^2 = E/[T(T - E)]$. For $E = T/3$ 
this simplifies to $\sigma^2 = (2T)^{-1}$.\(^1\)

4. Observed and chance skill scores

Two sets of operational skill scores have been used 
to test the results of the preceding section. The first set, 
kindly made available by Dr. D. L. Gilman, covered 
the period from winter 1974/75 through summer 1979 
and consisted of 100 station scores and expected success 
numbers $E_2$ for each of 19 seasons. Each score was 
computed from six forecasts of the monthly mean 
temperature, starting the forecast period from the 
beginning and from the middle of each month. To reduce 
the effect of spatial coherence only half the stations 
were used here. The second set of skill scores, kindly 
made available by Dr. K. Hanson, consisted of 21 sets 
of monthly precipitation and mean temperature skill 
scores for the official AC3 forecasts as well as for 
base and climatology ones, all of them computed 
with $E = E_3 = T/3$ from one precipitation and one 
temperature for each of the 48 contiguous states of the 
United States.

The analysis involved rearranging the scores or 
equivalent chi values in order of ascending magnitude. 
As chance results, the resulting cumulative frequencies 
would then be expected to approximate straight lines 
in probability coordinates, constructed to make 
\[ \int_{-\infty}^{\infty} e^{-x^2/2}dx \] a linear function of $x$.

Figure 1 shows the chi values of the first set plotted in 
this manner separately for the four seasons and the 
year, with offset abscissa scales. The full straight lines 
are regression lines fitted to the nine ordinates representing 
cumulative probabilities from 0.1 through 0.9, 
while the broken straight lines are the expected 
theoretical distributions, $N(0, 1)$. The observed distributions 
can be seen to have means (50% intersections) 
different from zero; they also have steeper slopes which 
imply that the variances $\sigma^2$ of the observed $X$ values 
are larger than their chance value, unity. A plausible 
explanation is that the six overlapping forecasts in each 
skill score are correlated with one another. An effective 
number of $T_{eff}$ of independent forecasts can then be 
defined by $T_{eff}/T = 1/\sigma^2 x$, so that in the present case 
$T_{eff} = 6/\sigma^2 x$.

The mean values of $x$ for the four seasons and for 
the entire first set are given in Table 2, together with 
their standard deviations and independent forecast 
numbers $T_{eff}$ which also appear in Fig. 1. The pairs of 
average skill scores $S$ in Table 2 represent the extremes 
of $E_2$, which can range from $E_2 = 2.4$ (for $x_2 = T$) 
for $E_2 = 1.8$ (for $x_2 = 0$). As means of 250 scores (200 
for the fall season) the $S$ values are significantly positive 
for winter and spring only, according to the Student's 
t-tests (Hoel 1962, section 11.5) in the last two lines of the 
table. The numbers of independent forecasts range

\(^{\text{1}}\) A reviewer has pointed that, in this special case, $S$ is a binomial 
variate which approaches the normal form for large $T$. The $x_{(1)}$ variate 
is not subject to such restrictions.
from a minimum of 3.5 in summer to a maximum of 5.3, only a little less than the actual number of forecasts (6), in spring.

The same type of representation has been used in Fig. 2 for the skill scores themselves of the official and climate forecasts in the second dataset. For economy of space, the cumulative temperature skill scores are plotted rising from right to left, and those of precipitation descending in that direction. Again the plots are reasonably straight; their means and standard devia-

**TABLE 2.** Characteristics of the distribution of skill scores for six monthly mean temperature forecast for 50 U.S. stations (D. Gilman’s data). The skill scores have been converted to their \( x_{10} \) equivalents, as described in the text, for the two extreme possibilities of the expected success number, \( E_{\text{max}} = 2.4 \), \( E_{\text{min}} = 1.8 \). The \( T_{\text{en}} \) are the numbers of independent forecasts in the 6 making up each skill score. The last two lines test the statistical significances of the \( \bar{x} \) with Student’s \( t = \bar{x}/\sigma \), where \( \sigma = \sigma_{\bar{x}}N^{-1/2} \).

<table>
<thead>
<tr>
<th></th>
<th>Spring</th>
<th>Summer</th>
<th>Fall</th>
<th>Winter</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Scores (N)</td>
<td>250</td>
<td>250</td>
<td>200</td>
<td>250</td>
<td>950</td>
</tr>
<tr>
<td>( \bar{x} )</td>
<td>0.322</td>
<td>0.063</td>
<td>-0.038</td>
<td>0.356</td>
<td>0.261</td>
</tr>
<tr>
<td>( S_{\text{max}} = \bar{x}/(T-E_{\text{max}}/E_{\text{max}})^{1/2} = \bar{x}/3 )</td>
<td>0.107</td>
<td>.021</td>
<td>-0.013</td>
<td>0.119</td>
<td>0.087</td>
</tr>
<tr>
<td>( S_{\text{min}} = \bar{x}/(T-E_{\text{min}}/E_{\text{min}})^{1/2} = \bar{x}/3.742 )</td>
<td>0.086</td>
<td>.017</td>
<td>-0.010</td>
<td>.095</td>
<td>0.070</td>
</tr>
<tr>
<td>( \sigma_{\bar{x}} )</td>
<td>1.061</td>
<td>1.310</td>
<td>1.220</td>
<td>1.203</td>
<td>1.118</td>
</tr>
<tr>
<td>( T_{\text{en}} = \frac{6}{\sigma_{\bar{x}}^2} )</td>
<td>5.33</td>
<td>3.50</td>
<td>4.03</td>
<td>4.15</td>
<td>4.80</td>
</tr>
<tr>
<td>( t = \bar{x}N^{1/2}/\sigma_{\bar{x}} )</td>
<td>4.80</td>
<td>0.76</td>
<td>-0.44</td>
<td>4.68</td>
<td>7.20</td>
</tr>
<tr>
<td>Chance probability of exceeding ( t )</td>
<td>&lt;0.001</td>
<td>&gt;0.40</td>
<td>&gt;0.60</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
skill. The test was designed to discriminate between several alternative levels of skill. In the operational context these are expressed more conveniently in terms of the forecast success ratio which for $E_3 = T/3$ has the form

$$R/T = \frac{2S - \frac{1}{3}}{\frac{1}{3}}$$  \hspace{1cm} (8)$$

Corresponding values of $R/T$ and $S$ are given in Table 4. The procedure used is the "sequential probability ratio test" (e.g., see Hoel, 1962, section 14.1). For discriminating with optimum efficiency (minimal sampling) between alternative means of a normal distribution with unit variance, progressive sums of the variable $\left(\frac{2T_{\text{eff}}}{T}\right)^{1/2}S$, in the present case are calculated and plotted as function of time in relation to pairs of parallel straight lines, representing two mean skill scores $\bar{S}$ (i.e., $\sum \left(\frac{2T_{\text{eff}}}{T}\right)^{1/2}S_i < \left[\sum \left(\frac{2T_{\text{eff}}}{T}\right)^{1/2}S_i\right]_2$). If the higher of these mean scores ($\bar{S}_1$) is true the cumulative sum should move above that line in all but a small number $\beta$ of cases; if the lower mean score is true, the cumulative sum should fall below the $\bar{S}_2$ line in all but a small number $\alpha$ of cases. No decision between the two mean scores can be made while the cumulative sum remains between the two lines; instead further scores must be awaited until one of the lines is crossed, although a cumulative sum rising faster (slower) than the two lines can be taken as suggesting an upward (downward) trend in skill.

For the present case the limits are defined by the inequalities (derived in Hoel, 1962, p. 355)

$$\frac{1}{2T_{\text{eff}}N} \log e \frac{\beta}{1-\alpha} = \frac{m}{\sum S} < \frac{m}{\sum S}$$

$$< \frac{1}{2T_{\text{eff}}N} \log e \frac{1-\beta}{\alpha} + \frac{m}{\sum S},$$  \hspace{1cm} (9)$$

where $\bar{S}$ is the difference between the two prescribed mean scores and $\sum S$ the sum; $m$ is the progressive number of skill scores tested.

The limits corresponding to (9) have been calculated with $\alpha = 0.05$ and $\beta = 0.1$ for the forecast skill scores

5. The operational assessment of skill scores

The second dataset has also been used to simulate an operational test for possible trends in forecasting skill. Their statistical significant skill scores of 0.17 and 0.18 as means of 21 values for precipitation, and no skill for the temperature forecasts. In this case we expect $\sigma^2 = \left(2T\right)^{-1}$; the actual standard deviations are equivalent to independent forecast numbers $T_{\text{eff}} = \left(2\sigma^2\right)^{-1}$ ranging from 18 to as low as 4, compared to the actual value of $T$ (48), presumably due to the spatial coherence of the forecasts. The large scatter (low $T_{\text{eff}}$ values) of the persistence forecasts points to especially large patterns as responsible for successes or failures.

### Table 3. Characteristics of the distribution of skill scores for 21 forecasts of the mean monthly temperature (T) and precipitation (P) for the 48 contiguous states (K. Hanson’s data). The $T_{\text{eff}}$ are the number of independent forecasts in the 48 making up each skill score. For other details see text.

<table>
<thead>
<tr>
<th>Forecasts</th>
<th>Official</th>
<th>Climate</th>
<th>Persistence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>$T$</td>
<td>$P$</td>
<td>$T$</td>
</tr>
<tr>
<td>$\sigma_S$</td>
<td>0.04</td>
<td>0.17</td>
<td>-0.03</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.225</td>
<td>0.166</td>
<td>0.164</td>
</tr>
<tr>
<td>$T_{\text{eff}} = \frac{1}{2\sigma^2}$</td>
<td>9.9</td>
<td>18.1</td>
<td>18.6</td>
</tr>
<tr>
<td>$t = \frac{S\left(2T\right)^{1/2}}{\sigma}$</td>
<td>0.81</td>
<td>4.69</td>
<td>-0.08</td>
</tr>
<tr>
<td>Probability exceeding $t$ with 20 degrees of freedom</td>
<td>&gt;0.30</td>
<td>&lt;0.001</td>
<td>&gt;0.90</td>
</tr>
<tr>
<td>$S$</td>
<td>0.05</td>
<td>0.08</td>
<td>0.237</td>
</tr>
<tr>
<td>$\sigma_S$</td>
<td>8.9</td>
<td>4.0</td>
<td>3.97</td>
</tr>
</tbody>
</table>
Table 4. Skill Score ($S$) and forecast success ratio ($R/T$) equivalents.

<table>
<thead>
<tr>
<th></th>
<th>$S$</th>
<th>$0.1$</th>
<th>$0.25$</th>
<th>$0.40$</th>
<th>$0.55$</th>
<th>$0.70$</th>
<th>$0.85$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R/T$</td>
<td>$0.33$</td>
<td>$0.4$</td>
<td>$0.5$</td>
<td>$0.6$</td>
<td>$0.7$</td>
<td>$0.8$</td>
<td>$0.9$</td>
</tr>
</tbody>
</table>

of the second dataset, transformed to chi variates with $T_{eff} = 48$ and $T_{eff} = 16$. In Figs. 3 and 4 the progressive sums of the $X$ equivalents of $S$ are shown in relation to their limits for the entire period of record. In practice a test would become conclusive (with the assurance set by $\alpha$ and $\beta$) when the $X$ plots cross one of the limits. Such crossings happened for both the official and climatology forecasts of precipitation when the full number (48) of forecasts was used (Fig. 3). This is shown more clearly by a numerical version of the test (Table 5). In a truly operational context the test would end with a crossing and begin anew from that point; since the limits are parallel straight lines this simply means transferring the origin to the time ($m$) of the crossing. Treated in this manner the last four climatology forecasts of temperature ($\chi$) suggested a significant increase in success ratio to over 50%.

![Fig. 3](image1.png)

**Fig. 3.** Cumulative sums of chi values computed from the skill scores of Fig. 2. The parallel lines are sequential limits for discriminating between the mean forecast success percentages indicates, as explained in the text, assuming each score to represent $T = 48$ independent verification points, $O$: official NOAA forecasts, $P$: forecasts based on persistence, $C$: forecasts based on climatology.

![Fig. 4](image2.png)

**Fig. 4.** As in Fig. 3, but assuming each score to represent only 16 independent verification points.

With a smaller effective number Fig. 4 shows that the decision limits move farther apart; the forecast sequences then remained in limbo between the 40% and 50% success rates for the entire period of 21 months.

6. Conclusion

The sequential analysis represents one way of continually monitoring changes in skill and assessing their statistical significance. From a physical point of view the concurrent behavior of the official and climate skill scores for precipitation in the second dataset examined is quite revealing. These precipitation skill scores had comparable positive mean values. One possible interpretation is that the official forecasts gave a good deal of weight to climatic mean values, and succeeded because the climate during the period studied "played the game", with small anomalies. At the current state of the art of extended forecasting such conformism or persistence may be prerequisites for more than occasional successes. Viewed in this way, the forecasts follow a fixed track while climate weaves around their predictions, approaching and following them for a time now and then before diverging again. If this is a realistic appraisal, a monitoring procedure revealing sudden changes in skill, in the manner here outlined, ought to become a stock in trade of long-range predictions.

A modified procedure for probabilistic forecasts
TABLE 5. Sequential testing for 10% differences in forecast success ratios. The observed sums have been computed from the precipitation scores of the official forecasts for May through September 1985.

<table>
<thead>
<tr>
<th>m</th>
<th>$\sum_{i} \sqrt{zTS}$</th>
<th>Hypothetical $R/T$ limits %, $T = 48$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>40 vs 50</td>
</tr>
<tr>
<td>1</td>
<td>3.72</td>
<td>0.18</td>
</tr>
<tr>
<td>2</td>
<td>6.76</td>
<td>1.90</td>
</tr>
<tr>
<td>3</td>
<td>10.78</td>
<td>3.61</td>
</tr>
<tr>
<td>4</td>
<td>11.60</td>
<td>5.33</td>
</tr>
</tbody>
</table>

Interpretation: $R/T = 50\%$ if the alternative is chosen to be 40%; the test remains undecided between 50% and 60% until $m = 5$ when 50% is favored. The alternative of 70% is rejected at $m = 2$.

might use a larger number of classes and make the score give also some credit for forecasts which miss by only one class (e.g., predicted "much above" when "above" occurred). It would be worthwhile to carry out a detailed comparison of the distribution of the predicted probabilities with that of the verified frequencies, so far made available only as a three-year summary by Gilman (1986).

Acknowledgment. A second reviewer's comments have helped greatly to clarify the intended meaning of this note.

REFERENCES


