The International Research Institute for Climate and Society

Descriptions of the IRI Climate Forecast Verification Scores

The International Research Institute for Climate and Society
The Earth Institute at Columbia University
61 Route 9W, Palisades NY 10964

Table of Contents

Rate of Return

The rate of return tells us how much money would be made if one invested on the forecast with odds corresponding to the probabilities given in the forecast. The amount of money made refers to the average proportion of the initial money gained per individual forecast. This proportion of initial money gained corresponds closely to the extent to which the forecasts assign high probabilities to the later observed outcomes. Such high probabilities given for the later observed results occur when both of two conditions are satisfied: (1) the forecasts are able to distinguish successfully among the observed outcomes, and (2) the forecasts do not have substantial unconditional or conditional biases in their probabilities; i.e., they have good reliability and resolution. Thus, the score reflects discrimination, reliability and resolution.

The rate of return examines the probabilities that were assigned to the later actually observed categories, over time for a given location. It computes the geometric average of all of these forecast probabilities, which requires multiplying them and then taking the root of the product, where the order of the root is the number of forecasts involved. For example, if there are 4 forecasts for a given location, each made at a different time, then the 4th root of the product of the 4 probabilities is taken. Then this same computation is done for the case in which the climatological forecast probability of 0.333 is issued for the same number of forecasts; this produces a geometric mean of 0.333. Next, the geometric mean of the actually issued forecasts is divided by that for the climatology forecasts (0.333), giving a ratio of the two geometric means. When this ratio is 1, the forecasts have no more value than the naïve issuance of climatological probabilities for all of the forecasts. The final step in the rate of return calculation is to subtract 1 from the ratio, so that the result is 0 when the forecasts have no more value than a set of climatology forecasts. When probabilities of 1 (or 100%) are always forecast for what later turns out to be the observed category, the rate of return is 2, since it is (1 / 0.333) – 1, which is 3 – 1, or 2.

A more concrete understanding of the meaning of the rate of return can be gained by considering the proportion of money that a better would gain when starting with some initial amount of money, and betting on each of a set of forecasts with odds that correspond to the forecast probabilities assigned to each of the possible categorical outcomes, and reinvesting all of the resulting balance in this same way for each successive forecast. The rate of return is the average proportion of the initial money gained per forecast, after betting on all of the forecasts, compared with betting just the climatological odds (of 0.333 for each category) on each forecast. The latter betting style would have the same result as not betting at all, and the better would still have just the initial amount of money at the end. One hopes that the ratio of the average proportion of money gained per bet to the average proportion gained when always betting 0.333 would exceed 1.

The following example illustrates the determination of the rate of return for a set of tercile-based probability forecasts. Suppose the following 5 forecasts were given for the below, near, and above normal precipitation categories, and that the category whose probability is shown in red was later actually observed.

B	**N**	**A**
45	35	20
33	33	33
40	33	27
15	30	55
20	40	40

The computation of the rate of return would begin by calculating the fifth root of the product of the five probabilities that were later actually observed, shown in red:

$$\sqrt[5]{(0.35)(0.33)(0.40)(0.55)(0.40)} = \sqrt[5]{0.01016} = 0.399$$

This 0.399 is called the likelihood score (see the description of the likelihood score.) The final step in computing the rate of return is to form a ratio comparing this geometric average to what it would be when always issuing the climatological probabilities (0.333, which would provide no useful forecast information), and then subtracting one:

$$\frac{GeomAvg}{NoSkillGeomAvg} - 1$$

For the example above, with a geometric average of 0.399, the rate of return is

$$\frac{0.399}{0.333} - 1 = 0.198$$

This result means that the better would average 19.8% more money per forecast. This 19.8% (or 0.198) is the rate of return.

A note about the computation of a geometric average: In the example above, the small sample size of forecasts made possible a computation of the geometric average directly, without the use of logarithms. But in many real situations there may be a much larger sample of forecasts, and the multiplication of many numbers of less than 1 may yield a product too small to be represented accurately on a computer (underflow). To circumvent this computational problem, the logarithm of the forecast probabilities may be used instead of the probabilities themselves, and the logarithms would be added rather than multiplied. For more detail about using logarithms for larger sample sizes, see the illustration provided near the end of the description of the likelihood score, in which the geometric mean is also computed.

Likelihood Score

Like the rate of return score, the likelihood indicates the extent to which the forecasts assign high probabilities to the later observed outcomes. Such high probabilities are possible when both of two conditions are satisfied: (1) the forecasts are able to distinguishing successfully among the observed outcomes, and (2) the forecasts do not have substantial unconditional or conditional biases in their probabilities. Thus, the score reflects discrimination, reliability and resolution. The difference between the likelihood score and the rate of return pertains to the scaling of the score. The likelihood score is not scaled to correspond to the proportion of initial monetary investment made per forecast, but rather simply indicates the typical probability level assigned to the category that was later observed.

The likelihood score examines the probabilities that were assigned to the later actually observed categories. It computes the geometric average of all of these probabilities, which requires multiplying them and then taking the root of the product, where the order of the root is the number of forecasts involved (for example, if there are 4 forecasts, then the 4th root of the product of the 4 probabilities is taken). This geometric average, called the likelihood score, is then compared with that expected for issuance of constant climatological forecast probabilities of 0.333, which produces a geometric mean of 0.333. The difference between the geometric mean of the actual forecasts and 0.333 is determined, and this difference divided by the factor 0.667 so that the score is scaled such that an actual forecast geometric mean of 1 (maximally confident forecasts that verify positively) produces a likelihood skill score of 1, and a forecast geometric mean of 0.333 produces a likelihood skill score of 0.

The following example illustrates the determination of the likelihood score for a set of tercile-based probability forecasts. Suppose the following 5 forecasts were given for the below, near, and above normal precipitation categories, and that the category whose probability is shown in red was later actually observed.

B	N	A
45	35	20
33	33	33
40	33	27
15	30	55
20	40	40

The computation of the likelihood score consists of calculating the fifth root of the product of the five probabilities that were later actually observed, shown in red:

$$\sqrt[5]{(0.35)(0.33)(0.40)(0.55)(0.40)} = \sqrt[5]{0.01016} = 0.399$$

This 0.399 is the likelihood score. It is the geometric average of the probabilities given to the later observed category, and is used to compute both the rate of return (see the description of the rate of return score) and the likelihood skill score (see the description of the likelihood skill score).

A note about the computation of a geometric average is in order. In the example above, the small sample size of forecasts made possible a computation of the geometric average directly, without the use of logarithms. But in many real situations there may be a much larger sample of forecasts, and the multiplication of many numbers of less than 1 may yield a product too small to be represented accurately on a computer (underflow). To circumvent this computational problem, the logarithm of the forecast probabilities may be used instead of the probabilities themselves. The logarithms are summed, and then divided by the number of probabilities. This division replaces the use of a high order root for the product of the original probabilities. At the end of the process, the quotient becomes an exponent applied to the base used for the logarithms, to convert the result from logarithms to non-logarithms. The chosen base may be whatever is desired (e.g., base 10, base 2, or base e). This process applied to the example above follows, using the natural logarithm (base e, or 2.71828):

$$\ln(0.35) + \ln(0.33) + \ln(0.40) + \ln(0.55) + \ln(0.40)$$
$$= -1.050 - 1.109 - 0.916 - 0.598 - 0.916$$
$$= -4.589$$

This sum of the 5 logarithms is then divided by the number of forecasts, to get their average:

$$\frac{-4.589}{5} = -0.918$$

Then this average of the logarithms is exponentiated using the base e, since e was used as the base for computing the logarithms:

$$e^{-0.918} = 0.399$$

This result for the geometric average (i.e., the likelihood score) agrees with that found without using logarithms.

Likelihood Skill Score

Like the rate of return score, likelihood skill score indicates the extent to which the forecasts assign high probabilities to the later observed outcomes. Such high probabilities are possible when both of two conditions are satisfied: (1) the forecasts are able to distinguishing successfully among the observed outcomes, and (2) the forecasts do not have substantial unconditional or conditional biases in their probabilities; Thus, the score reflects discrimination, reliability and resolution. The difference between the likelihood skill score and the rate of return pertains to the scaling of the score. The likelihood skill score is not scaled to correspond to the proportion of initial monetary investment made per forecast, but instead is scaled to reach 1 (or 100%) as the maximum possible level. Like the rate of return score, the likelihood skill score is 0 when the probabilities given to the later observed category are the same as the climatological probabilities (0.333/0.333/0.333).

To compute the likelihood skill score, the difference between the likelihood score (the geometric average) to what it would be when always issuing the climatological probabilities (0.333, which would provide no useful forecast information) is calculated, and this difference is divided by the difference between 0.333 and the maximum possible likelihood score of 1 (i.e., dividing by 0.667):

$$LikelihoodSkillScore = \frac{likelihoodScore_{forecast} - likelihoodScore_{cli}}{1 - likelihoodScore_{cli}}$$

In the above formula, the subscript "forecast" means the likelihood score of the forecasts, and "cli" means the likelihood score of constant climatology forecasts (0.333). Using the example of the five forecasts used in the description of the likelihood score, the likelihood score for the forecasts is 0.399. Then the likelihood skill score is computed as

$$\frac{0.399 - 0.333}{1.000 - 0.333} = \frac{0.066}{0.667} = 0.099$$

This likelihood skill score indicates that the geometric average (the likelihood score) is 0.099 (or 9.9%) of the distance, numerically, between the no-skill level of 0.333 and the maximum possible skill level of 1.000. Some positive skill is therefore indicated. Negative likelihood skill scores are produced when the likelihood score for the forecasts is less than 0.333.

It can be shown that in a tercile-based forecast system the likelihood skill score is always exactly one-half of the rate of return score (see the description of the rate of return score), and thus provides precisely the same basic skill information. The difference is only in the scaling, where the likelihood skill score is set such that 0 indicates no-skill forecasts and 1 indicates the maximum possible forecast skill, while for the rate of return these two benchmark numbers are 0 and 2, respectively.

ROC score

The ROC (relative operating characteristics) curve indicates the degree of correct probabilistic discrimination in a set of forecasts. Discrimination is the ability to distinguish one categorical outcome from another even if the forecast probabilities have biases or calibration problems.

ROC is plotted as a graph with the hit rate shown on the vertical axis and the false alarm rate shown on the horizontal axis. A hit implies the occurrence of an event of interest, such as below normal precipitation, while a false alarm implies the nonoccurrence of such an event. An ROC curve is plotted for each forecast category individually. The ROC curve is created from a set of points on the graph. The first point indicates the hit rate (as a proportion of the total forecasts that are hits) and the associated false alarm rate (also as a proportion) only for forecasts in the highest bin of issued forecast probabilities. Subsequent points indicate the hit versus false alarm rates for cases having successively decreasing forecast probabilities also added into the sample, sequentially. The area underneath the ROC curve is the ROC score. ROC scores above 0.5 reflect positive discrimination skill, 1.0 representing the maximum possible score. Probabilistic biases (i.e., calibration problems) do not degrade the score, as the score isolates discrimination exclusively. More detail about the ROC score and its calculation is provided below.

In a set of forecasts having good discriminative ability for the given forecast category, forecasts with the highest probability bin are expected to have a greater hit rate than those in the next lower probability bin, and so forth through the lowest probability bin in which false alarms would be expected to be most prevalent. Note that in contrast to the case of high forecast probabilities, for low forecast probabilities hits are seen as being undesirable while false alarms are desirable. For a skillful set of probability forecasts, because the points on the ROC curve are created using initially only those cases having the highest forecast probabilities, then growing in increments to include also the bins of successively lower forecast probabilities (and thus containing progressively larger total samples of forecasts), the hit rate is expected to decrease while the false alarm rate is expected to increase. The first few points would then be expected to begin in the lower left corner and to form a curve having a steep upward slope, gradually becoming less steep and ending with a very shallow slope in the upper right corner as the false alarm rate exceeds the hit rate, and as the proportion of hits and false alarms both approach 1. The curve would therefore arch upward with respect to the diagonal line representing equal hits and false alarms, and the area under the curve would be greater than 0.5 (indicating positive skill). Forecast sets having no probabilistic skill would be expected to have roughly the same slope (equaling 1) across the entire plot, and therefore to hover near the diagonal line connecting the lower left and the upper right corners, and to have an area close to 0.5. Forecasts with negative skill would have lower slope in the lower left part of the plot and higher slope in the upper left part, and therefore would be located mainly below the diagonal line and would have an area of less than 0.5 (indicating negative skill). It should be noted that forecasts having good discriminative ability will produce high ROC scores regardless of whether or not the probability values are well calibrated. Thus, for example, if the probability bins represented probabilities that were only one-third as large as their corresponding observed relative frequencies, but their ordering reflected good discrimination, the ROC score would be no different than it would be with correct probabilistic calibration. This feature allows discrimination to be examined without influence from calibration or bias issues, which presumably could be corrected. By examining the ROC curves for the different forecast categories, it is possible to detect cases of conditional discriminative skill, as for example being able to forecast below normal conditions with better discrimination than above normal conditions.

Because ROC analysis involves subdividing a forecast data set into separate probability bins, a fairly large set of forecasts is needed for a robust outcome. Nonetheless, the procedure can be illustrated using an example of just 5 probability forecasts given for the below, near, and above normal precipitation categories, with the later observed category shown by the probability in red.

B	N	A
45	35	20
33	33	33
40	33	27
15	30	55
20	40	40

Suppose we wish to make the ROC diagram and compute the ROC area for the above normal category. First we select a bin size. Since there are only 5 forecasts, each having differing probabilities for above normal, we define category bins that resolve each of the 5 probabilities into one of the five bins. Such a bin definition might be <25, 25-30, 30-35, 35-45, and >45. (With a larger number of forecasts, one could afford to define the bins to have equal widths.) We then form the following table tallying the hit rate and false alarm rate, cumulative in descending order of probability bin. Note that since above normal was observed in two cases and not observed in three cases, there are totals of 2 hits and 3 false alarms.

Bin	**Hit Rate (%)**	**False Alarm Rate (%)**
>45	1 out of 2 (50%)	0 out of 3 (0%)
35-45	1 out of 2 (50%)	1 out of 3 (33%)
30-35	2 out of 2 (100%)	1 out of 3 (33%)
25-30	2 out of 2 (100%)	2 out of 3 (67%)
<25	2 out of 2 (100%)	3 out of 3 (100%)

The ROC graph would then have a dot drawn by default at the (0%, 0%) location in the lower left corner, and then 5 dots drawn at the locations specified in the table above: (50%, 0%), (50%, 33%), (100%, 33%), (100%, 67%), and (100%, 100%). The last point is in the upper right corner. The line connecting the dots is located to the upper left of the diagonal line (at 45º) defining equal hits and false alarms, making for an area exceeding 0.5 and therefore indicating positive discriminative skill. In fact, the area for this example can be calculated to be 0.8, indicating substantial positive skill. This same exercise could be carried out for the near normal and below normal categories.

Generalized ROC Score (GROC)

The generalized ROC score (GROC), like the ROC, shows the degree of correct probabilistic forecast discrimination, even if the forecasts have biases or calibration problems. However, unlike ROC, GROC is generalized to encompass all forecast categories (below, near, and above normal) collectively, rather than being specific to a single category.

The GROC is not plotted as ROC is; however, the score that it produces is fully equivalent to the ROC area, ranging from 0.5 for a set of probability scores without skill to 1 as the maximum possible value. Similar to ROC, probabilistic biases (i.e., calibration problems) do not degrade the GROC score, which reflects purely discrimination ability. More detail about the GROC score and its calculation is provided below.

The GROC is based on a comparison of the forecasts issued for all available pairs of observations of differing category. More specifically, it is the proportion of all such pairs of observations of differing category whose probability forecasts are discriminated in the correct direction. Such a correct discrimination between two observations of differing category would be, for example, cases of near normal and above normal observed precipitation whose respective probability forecasts were 0.20/0.35/0.45 (for below/near/above-normal) and 0.15/0.35/0.50, respectively. Had the two observations been reversed, the forecasts would not have shown correct direction of discrimination. One-half hit is awarded for tied (e.g., identical) forecasts for cases having different outcomes. Again, the final proportion of hits, which is then the GROC score, is fully equivalent to the ROC area for an individual category.

The procedure to calculate GROC can be illustrated using the example of just 5 probability forecasts given for the below, near, and above normal precipitation categories, with the later observed category shown by the probability in red.

Forecast#	**B**	**N**	**A**
1	45	35	20
2	33	33	33
3	40	33	27
4	15	30	55
5	20	40	40

The first step is to identify all pairs of forecasts in which the first one has a lower observational outcome than the second. (This means that if the first one is in the below normal category, the second must be either in the near normal or the above normal category; and if the first one is in the near normal category, the second must be in the above normal category.) Then, for each paired comparison, we determine whether the forecast for the first case is discriminated in the correct direction in comparison with the forecast for the second case—i.e., if it indicated a lower result; or, on the other hand, if it is not discriminated in the correct direction. In the former case a hit is awarded, while in the latter case a hit is not awarded. In some cases, such as when the two forecasts are identical, there is no clear verdict, and one-half hit is awarded.

In the above example there are 8 pairs of forecasts having differing observed results, listed as follows, with the hit/no hit status shown in the right column.

Pair#	Forecast#s	Hit
1	3 vs 1	0
2	3 vs 2	1
3	3 vs 4	1
4	3 vs 5	1
5	1 vs 2	1
6	1 vs 4	1
7	5 vs 2	0
8	5 vs 4	1

The total number of hits is 6, with a maximum possible number of 8. The GROC score is therefore 6 / 8, or 0.75. Compared with an expected no-skill level of 0.5, positive skill is indicated. Although the hit status is often obvious upon inspection of the two probability forecasts and their respective observed results, there may be less obvious cases in which a hit index formula needs to be used to determine the hit status. The formula is:

$$HitIndex = \frac{\sum_{r=1}^{m-1} \sum_{s=r+1}^{m} p_{k,i}(r) p_{l,j}(s)}{1 - \sum_{r=1}^{m} p_{k,i}(r) p_{l,j}(r)},$$

where m is the number of categories, $p_{k,i}(r)$ is the forecast probability for the rth category, and for the ith observation in category k. When the hit index is >0.5, <0.5 or =0.5, the pair of forecasts is a hit, a non-hit, or a tie, respectively. In obvious cases the hit index is not close to 0.5, while in closer cases it is nearer to 0.5. It is important to use adequate precision for the probability forecasts in using this formula, such as to use at least 3 decimal places for probabilities such as 0.333, to avoid false hits or non-hits that should actually be ties.

Reliability Plot

The reliability plot shows how well the forecast probabilities correspond to the subsequent observed relative frequencies of occurrence, across the full range of issued forecast probabilities. This is examined for each of the forecast categories individually (below, near, or above-normal). It also shows the frequencies with which the various probabilities were issued, revealing forecast boldness (i.e., sharpness) versus conservatism. Together, the above diagnostics indicate the nature of any overall or conditional forecast biases, including whether the forecasts were overconfident or underconfident.

Because a reliability plot contains a wealth of detailed information about the forecasts and their correspondence with the observations for each of a set of issued probability intervals, it requires a large sample of data, and therefore typically pertains to a set of forecasts ranging over a considerable time span and may include a large aggregation of locations (e.g., a continent, all of the tropics or the entire globe). The primary feature of interest in the diagram is the observed relative frequency of occurrence that is associated with the issuance of each of a set of specific probability intervals (bins), shown individually for each of the 3 tercile-based categories. The x-axis shows the forecast probability intervals, while the y-axis shows the corresponding observed relative frequencies of occurrence. The set of dots pertaining to the forecasts for one of the categories (e.g., below normal precipitation) are connected by lines, and separate lines are shown for each of the three forecast categories. For example, one of the dots for the below normal forecast category might show that for the forecast probability interval of 0.45 to 0.55, the observed relative frequency of occurrence is 0.43. Similar dots would show results for 0.55 to 0.65, for 0.65 to 0.75, and so forth. Ideally, the observed relative frequencies would correspond closely with the central values of the forecast probability intervals. A diagonal line that corresponds to perfect reliability is usually shown, facilitating visual interpretation of the deviations of the actual result from the ideal result. When a line has a shallower (smaller) slope than the diagonal line showing perfect reliability, forecast overconfidence is indicated. Overall biases are indicated by overall vertical displacements of the line with respect to the 45º line representing perfect reliability. Often a best-fit regression line is shown for each forecast category, weighted by the frequency of issuance of the probability intervals. For example, points that represent forecast probability intervals frequently issued (such as the 0.35-0.45 interval) influence the position of the regression line more heavily than points representing more rarely forecast probabilities (such as the 0.05-0.15 interval or the 0.75-0.85 interval).

An inset plot is often shown within or below the main reliability plot, indicating the frequency of issuance of each probability interval for each of the forecast categories. This plot reveals how strongly and frequently the issued forecast probabilities depart from the climatological probabilities—a forecast characteristic known as sharpness.

Sometimes a table of corresponding forecast probabilities and their corresponding observed relative frequencies is shown, as well as summary statistics indicating overall forecast bias (overall mean forecast probability given for a given category, versus the overall observed relative frequency for that category). The latter feature may also be shown by marks on the axes of the main graph that indicate the mean forecast probability for a category (on x-axis) and the mean observed relative frequency of the category (on the y-axis).

Heidke Hit Proportion

The Heidke hit proportion is the proportion of forecast cases in which the forecast category assigned the highest probability is later observed. The value of the probability itself is ignored. When more than one category share the highest probability, partial credit is awarded. Thus, the score reflects discrimination, reliability and resolution.

The Heidke hit proportion tallies all of the forecasts (over time, over space, or both) in which the category given the highest forecast probability is observed. The probability itself does not matter, so that, for example, in the tercile-based system when the above normal category is forecast with the highest probability, and is later observed, the same full credit (called a "hit") is awarded whether the forecast probability is only slightly higher than 0.333 (e.g., 0.35) or very high (e.g., 0.90). When two categories share the highest probability and one of them is observed, one-half hit is tallied. When all three probabilities are equal (i.e., the climatological probabilities of 0.333 are issued), one-third hit is tallied. The Heidke score is considered a simplified measure of forecast skill, easily understood by nontechnical users. Its main shortcoming is that it is insensitive to the details of the correspondence of the forecast probabilities with the relative frequencies of observed outcomes, such as probabilistic over- or under-confidence in the forecast probabilities. The following example illustrates computation of the Heidke score for the example of 5 probability forecasts given for the below, near, and above normal precipitation categories, with the later observed category shown by the probability in red.

B	N	A	Hit
45	35	20	0
33	33	33	0.333
40	33	27	1
15	30	55	1
20	40	40	0.5

The Heidke hit proportion, indicating the proportion of hits, is defined by the formula:

$$Heidke.score = \frac{\#\,hits}{total\,\#\,forecasts}$$

The number of hits summed over the 5 forecasts is 2.833. The Heidke score is then

$$\frac{\#\,hits}{total\,\#\,forecasts} = \frac{2.833}{5} = 0.567$$

This proportion indicates skill above that expected by chance, which is 0.333.

Heidke Hit Proportion for 2nd Most Likely Probability

Although the Heidke hit proportion was originally intended for use with the forecast category assigned the highest probability, it can be applied to the other categories also. The Heidke score for the 2nd highest probability is the proportion of forecast cases in which the forecast category assigned the second highest probability is later observed. The value of the probability itself is ignored.

In similar fashion to the standard Heidke score, when two categories share the second highest probability and one of them is observed, one-half hit is tallied. When all three probabilities are equal (i.e., the climatological probabilities of 0.333 are issued), one-third hit is tallied. When there are 3 probabilities, the 2nd highest probability is also the 2nd lowest probability, and the Heidke score for the 2nd highest probability might be expected to be close to the climatological level of 0.333. However, the Heidke score for the 2nd highest probability may be above the climatological level when the forecasts are heavily hedged toward the near normal category, due to forecaster cautiousness, and the second highest probability may be more strongly related to the true belief of the forecaster regarding what is most likely to happen, and to still be somewhat elevated above the climatological level. For example, a forecaster who believes that below normal precipitation is most likely but is hesitant to indicate this directly in the forecasts (whether due to fear of being "wrong" or for political or other reasons), might issue probabilities such as 0.40/0.45/0.15. Using the same example of the five forecasts used earlier, the Heidke score for the 2nd highest probability is computed as follows:

B	N	A	Hit
45	35	20	1
33	33	33	0.333
40	33	27	0
15	30	55	0
20	40	40	0.5

The number of hits for the 2nd highest probability summed over the 5 forecasts is 1.833, making a proportion of 1.833 / 5 = 0.367. This is slightly higher than the proportion expected by chance, which is 0.333. It is not straightforward to interpret this result in terms of implied forecast skill, because in some cases the 2nd highest probability may be closer to the highest probability than to the lowest probability, while in other cases the opposite may be true. A tendency for hedging toward the near normal category is not particularly marked in this set of forecasts, so interpretations involving such forecaster behavior may not be warranted here.

Heidke Hit Proportion for Least Likely Probability

The Heidke hit proportion for the least likely probability is the proportion of forecast cases in which the forecast category assigned the lowest probability is later observed. The value of the probability itself is ignored. Thus, in similar fashion to the standard Heidke score, this score reflects discrimination, reliability and resolution.

As with the standard Heidke hit proportion, when two categories share the lowest probability and one of them is observed, one-half hit is tallied. When all three probabilities are equal (i.e., the climatological probabilities of 0.333 are issued), one-third hit is tallied. The Heidke score for the lowest probability would hopefully be below the climatological level, because a higher score would indicate a poor set of forecasts. It may be computed in order to compare it to the standard Heidke score (for the category assigned the highest probability) to assess the degree to which it is lower than it. Using the example of the five forecasts used earlier, the Heidke score for the lowest probability is computed as follows:

B	**N**	**A**	**Hit**
45	35	20	0
33	33	33	0.333
40	33	27	0
15	30	55	0
20	40	40	0

The number of hits for the lowest probability summed over the 5 forecasts is 0.333, making a proportion of 0.333 / 5 = 0.067. This is much lower than the proportion expected by chance, which is 0.333. Together with the proportion of hits for the highest probability (0.567) computed earlier, this result corroborates the conclusion that some skill is present in this set of forecasts.

Heidke Skill Score

The Heidke skill score is based on the Heidke hit proportion, which is the proportion of forecast cases in which the forecast category assigned the highest probability is later observed. Hence, this score reflects discrimination, reliability and resolution. The Heidke skill score places the Heidke score in a framework in which the level reflecting no-skill is scaled to be 0 instead of 0.333, while the level reflecting all hits remains at 1. The rescaling for the Heidke skill score is done by first subtracting 0.333 from the Heidke score, and then multiplying the result by 1.5 (or, identically, dividing the result by 0.667), as reflected in the formula:

$$Heidke.skill.score = \frac{\#hits - \exp ected\#hits}{total\#forecasts - \exp ected\#hits}$$

Using the same example of the five forecasts used earlier, the Heidke skill score would use the same tally sheet as used for the Heidke score, which found a sum of 2.833 hits over the 5 forecasts. The number of hits expected by chance is 5 / 3, or 1.667. Using the formula we get:

$$\frac{\#hits - \exp ected\#hits}{total\# - \exp ected\#hits} = \frac{2.833 - 1.667}{5 - 1.667} = \frac{1.167}{3.333} = 0.350$$

Note that in the numerator, the amount subtracted is one-third of the total number of forecasts, and the denominator is two-thirds of the total number of forecasts.

Exceedance of Heidke Hit Proportion above 0.333

The Exceedance of Heidke Hit Proportion above 0.333 is based on the Heidke hit proportion, which is the proportion of forecast cases in which the forecast category assigned the highest probability is later observed. The Exceedance of Heidke Hit Proportion above 0.333 rescales the Heidke hit proportion by subtracting 0.333 such that the level reflecting no-skill becomes 0 instead of 0.333, and the level reflecting all hits becomes 0.667 instead of 1.

The purpose of this rescaled score is to show the difference of the Heidke score from the expected level of 0.333, the expected proportion of hits for random (no skill) forecasts or of constant issuance of climatology (0.333) forecasts. In this way, Heidke scores greater than 0.333 become positive scores, while those falling short of 0.333 become negative scores. Using the example of the 5 forecasts used earlier, we had found the sum of the hits to be 2.833 out of a possible 5, resulting in a Heidke score (or proportion of hits) of 0.567. The Heidke score of difference from 0.333 is then 0.567 minus 0.333, or 0.234. Note that a negative result occurs when the hits proportion is less than 0.333.

Ranked Probability Score (RPS)

The ranked probability score (RPS) measures the squared forecast probability error, and therefore indicates to what extent the forecasts lack success in discriminating among differing observed outcomes, and/or have systematic biases of location and level of confidence. Thus, the score reflects the degree of a lack of discrimination, reliability and/or resolution.

Successful discrimination among outcomes would be shown, for example, if cases in which above normal temperature occur are generally given higher probabilities for above normal than cases when above normal temperature do not occur. Systematic biases of location would be indicated, for example, if most of the forecasts for above normal temperature have probabilities that are too low, and those for below normal are too high, even if some discrimination for above and for below normal temperature is present. Systematic biases of confidence level would be indicated, for example, by forecasts that indicate more confidence (show probabilities farther from 0.333) than is justified by the degree of discrimination. More detail about the RPS is provided below.

The ranked probability score (RPS) is based on the squared probability error, cumulative across the three forecast categories in order from lowest to highest:

$$RPS = \frac{1}{ncat - 1} \sum_{icat=1}^{ncat} (Pcumfct_{icat} - Pcumobs_{icat})^2$$

Where icat is the category number (1 for below normal, 2 for near normal, 3 for above normal), ncat is the number of categories (3 in a tercile-based system), Pcumfct is the cumulative forecast probability up to category icat, and Pcumobs is the comparable term for the cumulative observation “probability”. The error is the squared difference between the cumulative categorical forecast probability and the corresponding cumulative observed “probability” in which 1 is assigned to the observed category and 0 is assigned to the other categories. For example, suppose the forecast was 0.20/0.35/0.45 (for below/near/above-normal), and the observed category was above normal. The cumulative forecast probabilities would then be 0.20, 0.55 and 1.00, while the cumulative observed “probabilities” would be 0.00, 0.00 and 1.00. The RPS would then be computed as $(0.00 – 0.20)^2 + (0.00 – 0.55)^2 + (1.00 – 1.00)^2$. These add up as 0.04 + 0.3025 + 0 = 0.3425. This cumulative squared probability error is then divided by one less than the number of categories, which for a tercile-based system is 3 – 1, or 2. This division is done to account for the dependence of the cumulative error on the mechanical artifact of the number of categories. (With more categories, the expected cumulative error increases, given the same basic probabilistic forecast quality.) Dividing by 2, the RPS becomes 0.3425 / 2 = 0.17125.

However, another final adjustment is still needed. When the near normal category is observed, the RPS tends to be smaller (indicating less forecast probability error) than when the below or above normal categories are observed, due simply to the fact that the forecast error is more limited when a high probability is forecast for an adjacent category as opposed to a category on the opposite extreme. For example, our forecast of 0.20/0.35/0.45 shows a clear tendency toward above normal. If above normal is later observed, the RPS is 0.17125, while if near normal is observed the RPS is 0.12125. This “better” result for a near normal observation may initially seem counterintuitive, but can be understood when considering the RPS resulting from each possible observational outcome when the climatology forecast (0.333/0.333/0.333) is issued. For the climatology forecast, RPS is 0.1111 when near normal is observed, but is 0.2778 when above or below normal is observed. If a large number of forecasts are issued and if approximately one-third of the stations (or grid points) are observed to be below, near, or above normal, then the impact of the strong tendency toward lower (better) RPS for the observed near normal cases would be inconsequential. However, such a balance of observed relative frequencies among the three categories cannot be guaranteed either between or within forecasts, and some forecasts might have lower RPS largely because more near normal observations were observed for them. Adjustment for this undesirable feature is accomplished by multiplying the RPS for forecasts having near normal observations by 2, and multiplying RPS for forecasts having above or below normal observations by 0.8. This results in an expected RPS of 0.2222 for climatology forecasts regardless of which category is observed, and creates intuitively reasonable RPS values for non-climatology forecasts regardless of the observed relative frequencies of the three categories. Applying this adjustment to the RPS for our example forecast, we multiply 0.12125 by 2, obtaining an RPS of 0.2425.

Note that higher RPS indicates greater forecast probability error. The RPS is sensitive not only to the forecast probability given to the observed category, but also to the probabilities given to the other categories. For example, in the case of a poor forecast in which the above or below normal category was observed, the RPS is relatively greater (indicating a poorer forecast) when the category on the opposite side of near normal has a high forecast probability than when the near normal category has a high forecast probability. This feature of the RPS is in contrast to the rate of return, likelihood score or likelihood skill score, where the probabilities assigned to categories that were not observed are ignored.

Ranked Probability Skill Score (RPSS)

The ranked probability score (RPSS) is a skill score based on a comparison of the cumulative squared probability error (i.e., the ranked probability score, or RPS) for an actual set of forecasts with the RPS that would result from constant issuance, for all forecasts, of the climatology forecast of 0.333 probability for each category. In either case, the RPS measures the squared forecast probability error, and therefore indicates the extent to which the forecasts lack success in discriminating among differing observed outcomes, and/or have systematic biases of location and level of confidence. Positive RPSS implies that the RPS is lower for the forecasts than it is for climatology forecasts. Thus, the score reflects discrimination, reliability and resolution.

In the comparison between the actual forecasts and the constant climatology forecasts, the orientation of the RPSS is reversed from that of RPS, where now higher scores indicate forecasts having higher skill levels. The RPSS therefore indicates to what extent, compared with constant climatology forecasts, the actual forecasts are successful in discriminating among differing observed outcomes, and are free of systematic biases of location and level of confidence. More detail about the RPSS is provided below.

The ranked probability skill score (RPSS) is based on a comparison of the squared probability error, cumulative across the three forecast categories in order from lowest to highest, with the same computation applied to constant climatology (0.333/0.333/0.333) forecasts. (Note that using climatology forecasts as the reference for comparison is not the only reasonable option. Other options may be random non-climatology forecasts, or damped persistence forecasts from the previous season's observations.) Specifically, RPSS is a comparison between the ranked probability score (RPS) for the forecasts to the RPS for constant climatology forecasts. In either case, high RPS indicates either or both of lack of successful discrimination among observed outcomes, and the presence of biases in location or forecast confidence. Successful discrimination among outcomes would be shown, for example, if cases when above normal temperature do occur are generally given higher probabilities for above normal than cases when above normal temperature do not occur. Systematic biases of location would be indicated, for example, if most of the forecasts for above normal temperature have probabilities that are too low, and those for below normal are too high, even if some discrimination for above and for below normal temperature is present. Systematic biases of confidence level would be indicated, for example, by forecasts that indicate more confidence (show probabilities farther from 0.333) than is justified by the degree of discrimination.

The RPS for either the actual forecasts or the constant climatology forecasts is computed as the squared difference between the cumulative categorical forecast probability and the corresponding cumulative observed "probability" in which 100% is assigned to the observed category and 0% is assigned to the other categories. For example, suppose the forecast was 20/35/45 (for below/near/above-normal), and the observed category was above normal. The cumulative forecast probabilities would then be 0.20, 0.55 and 1.00, while the cumulative observed "probabilities" would be 0.00, 0.00 and 1.00. The RPS would then be $(0.00 – 0.20)^2$ + (0.00 –

$0.55)^2 + (1.00 - 1.00)^2$. These add up as 0.04 + 0.3025 + 0 = 0.3425. Then, upon dividing by one less than the number of categories (i.e., 2), the RPS becomes 0.1712. As will be explained below, in computing RPSS it is not necessary to adjust the RPS further according to which category was observed (i.e. to multiply by 2 if near normal is observed, and by 0.8 if below or above normal is observed). Higher RPS indicates greater forecast probability error. Note that unlike the rate of return or the likelihood skill scores, RPS is sensitive not only to the forecast probability given to the observed category, but also to the probabilities given to the other categories. To compute RPSS, the RPS of the climatology forecast with the same observational outcome is also computed. In this case the climatology RPS is $(0.00 - 0.333)^2 + (0.00 - 0.667)^2 + (1.00 - 1.00)^2$. These add up as 0.111 + 0.444 + 0 = 0.555, and upon dividing by one less than the number of categories (i.e., 2) it is 0.278. The formula for RPSS is based on the RPS of the forecasts and the RPS of the constant climatology forecasts, as follows:

$$RPSS = 1 - \frac{RPS_{fct}}{RPS_{cli}}$$

where RPS_{fct} and RPS_{cli} are the RPS for the forecasts and for the climatology forecasts, respectively. When RPS_{fct} and RPS_{cli} are equal, RPSS is zero, and when RPS_{fct} is zero, RPSS reaches its maximum possible value of 1. In the case of the example above, with $RPS_{fct} = 0.1025$ and $RPS_{cli} = 0.222$, RPSS is computed as

$$RPSS = 1 - \frac{0.1712}{0.278} = 0.38$$

It is noted that in computing RPSS, the adjustment performed for the RPS (multiplying by 2 if near normal is observed, and by 0.8 if below or above normal is observed) is not necessary because it would not change the quotient on the right side of the equation, as both numerator and denominator would be multiplied by the same factor. It is also noted that in the formula for RPSS, the orientation is reversed from that of RPS such that higher RPSS indicates higher skill, and that the score is scaled such that forecasts with cumulative squared error equaling that of the climatology forecast have RPSS of 0, and forecasts with no squared error have RPSS of 1. In the current example, since the RPSS is somewhat higher than zero, the RPSS indicates skill somewhat better than it would be if the climatology forecast had been issued. It is possible for RPSS to be more strongly negative than -2 in some severe cases. It is noted that for a tercile-based system, when the observation is in the near normal category, RPS_{cli} is always 0.1111 (as in the above example), and when it is in either of the two outer categories, RPS_{cli} is always 0.2727.

Brier Score

The Brier score measures the squared forecast probability error, and therefore indicates to what extent the forecasts lack success in discriminating the occurrence of a specific observed category (such as above normal temperature), and/or have systematic biases of location and level of confidence in the probability of occurrence of that category of outcome. Thus, the score reflects the degree of a lack discrimination, reliability and/or resolution.

Successful discrimination for a specific outcome would be shown, for example, if cases when above normal temperature is observed are generally given higher forecast probabilities for above normal than cases when above normal temperature is not observed. Systematic biases of location would be indicated, for example, if most of the forecasts for above normal temperature have probabilities that are lower than their observed rate of occurrence, even if some discrimination for above normal temperature is present. Systematic biases of confidence level would be indicated, for example, by forecasts that indicate more confidence (show probabilities farther from 0.333) than is justified by the degree of discrimination. More detail about the Brier score is provided below.

The Brier score is based on the squared probability error for the given forecast category, such as above normal temperature. Thus, in a 3-category system there are three Brier scores—one for each of the three categories. The squared probability error is the squared difference between the categorical forecast probability and the corresponding observed "probability" in which 1 is assigned if the category is observed, and 1 is assigned if it is not observed. For example, suppose the forecast was 0.45 for the above normal temperature category, and the observation indicated that above normal temperature did not occur. The squared probability error would then be $(0.45 - 0.00)^2$, which is 0.2025. This is the unadjusted Brier score. Note that higher unadjusted Brier scores indicate greater forecast probability error. If above normal temperature did occur, the unadjusted Brier score for this case would be $(0.45 - 1)^2 = 0.3025$, indicating larger forecast probability error than if above normal did not occur.

When the forecast probability for a given category is less than 0.5, the unadjusted Brier score is smaller (indicating less forecast probability error) when the category is not observed than when it is observed. Suppose a forecast is 0.45/0.35/0.20, showing a clear forecast tendency toward below normal compared with the other categories. If below normal is later observed, the unadjusted Brier score is 0.3025, while if it is not observed the score is 0.2025. This "better" result for a category not being observed than being observed when its forecast probability is greater than 0.333 but less than 0.5 may initially seem counterintuitive, but can be understood when considering the unadjusted Brier score resulting from the "yes" or "no" observation outcome when the climatology forecast (0.333/0.333/0.333) is issued. For the climatology forecast, the unadjusted Brier score is 0.1111 when the category is not observed, and 0.4444 when the category is observed. If a large number of forecasts are issued and approximately one-third of the stations (or grid points) are observed to be below, near, or above normal, then the impact of the strong tendency toward lower (better) unadjusted Brier score when a given category is not observed (as noted in particular when its forecast probability exceeds 0.333 but is less than 0.5) would be inconsequential. However, such a balance among the observed relative frequencies of the three categories cannot be guaranteed either between or within forecasts, and some forecast categories might have lower unadjusted Brier scores largely because those categories were observed less than one-third of the time. Adjustment for this undesirable mechanical feature can be accomplished by multiplying the unadjusted Brier score for forecasts whose given category was not observed by 2, and multiplying the unadjusted Brier score for forecasts whose given category was observed by 0.5. This results in an expected adjusted Brier score of 0.2222 for climatology forecasts regardless of whether the given category is observed or not observed, and creates intuitively reasonable score values for non-climatology forecasts regardless of the observed relative frequencies of the categories. In the example above, since the above normal category was not observed, the unadjusted Brier score of 0.2025 would be multiplied by 2 and would become 0.4050. Note that if the above normal category did occur, the unadjusted Brier score would be 0.3025, but this would be multiplied by 0.5 to become 0.15125, indicating less error (in the context of the tercile system with its 0.333 baseline probability) than if the above normal category did not occur.

Brier Skill Score

The Brier skill score is a skill score based on a comparison of the squared probably error (i.e., the Brier score) for an actual set of forecasts with the Brier score that would result from constant issuance, for all forecasts, of the climatology forecast of 0.333 probability for a given forecast category (e.g., below normal precipitation). Since either Brier score reflects squared forecast probability error, it indicates the extent to which the forecasts lack success in discriminating among differing observed outcomes, and/or have systematic biases of location and level of confidence. Thus, the score reflects discrimination, reliability and resolution. A positive Brier skill score implies that the Brier score for the forecasts is lower than it is for the climatology forecasts.

In the comparison between the actual forecasts and the constant climatology forecasts, the orientation of the Brier skill score is reversed from that of the Brier score, where now higher scores indicate forecasts having higher skill levels. The Brier skill score therefore indicates to what extent, compared with constant climatology forecasts, the actual forecasts are successful in discriminating the occurrence versus nonoccurrence of the given category, and are free of systematic biases of location and level of confidence.

Successful discrimination of the occurrence of the given category outcome would be shown, for example, if cases when above normal temperature occur are generally given higher probabilities for above normal than cases when above normal temperature do not occur. Systematic biases of location would be indicated, for example, if most of the forecasts for above normal temperature have probabilities that are too low, even if some discrimination for above normal temperature is present. Systematic biases of confidence level would be indicated, for example, by forecasts that indicate more confidence (show probabilities farther from 0.333) than is justified by the degree of discrimination. More detail about the RPSS is provided below.

The Brier skill score for either the actual forecasts or for constant climatology forecasts is computed as the squared difference between the categorical forecast probability and the corresponding observed "probability" in which 1 is assigned if the category is observed, and 0 is assigned if the category is not observed. For example, suppose the forecast for above normal temperature is 0.45, and the observed category was in fact above normal. The unadjusted Brier score would then be $(1.00 – 0.45)^2$, which gives 0.3025. As will be explained below, in computing the Brier skill score it is not necessary to adjust the Brier score further according to whether the category was observed or not (i.e. to multiply by 2 if it is not observed, and by 0.5 if it is observed). Therefore, the unadjusted Brier score of 0.3025 is considered the final Brier score for the forecast.

To compute the Brier skill score, the Brier score of the climatology forecast (0.333/0.333/0.333) with the same observational outcome is also computed. In this case the climatology Brier score is $(1.00 – 0.333)^2$, which gives 0.444. The formula for the Brier skill score is based on the Brier score of the forecasts and the Brier score of the constant climatology forecasts, as follows:

$$BSS = 1 - \frac{BS_{fct}}{BS_{cli}}$$

where BS_{fct} and BS_{cli} are the Brier scores for the actual forecasts and for climatology forecasts, respectively. When BS_{fct} and BS_{cli} are equal, the Brier skill score is zero, and when BS_{fct} is zero, RPSS reaches its maximum possible value of 1. In the case of the example above, with BS_{fct} = 0.3025 and BS_{cli} = 0.444, the Brier skill score is computed as

$$BSS = 1 - \frac{0.3025}{0.444} = 0.32$$

It is noted that in computing Brier skill score, the adjustment performed for the Brier score (multiplying by 2 if the given category is not observed, and by 0.5 if the given category is observed) is not necessary because it would not change the quotient in the second term on the right side of the equation, as both numerator and denominator would be multiplied by the same factor. It is also noted that in forming Brier skill score, the orientation is reversed from that of the Brier score such that a higher Brier skill score indicates higher skill, and that the score is scaled such that forecasts with squared error equaling that of the climatology forecast have a Brier skill score of 0, and forecasts with no squared error have a Brier skill score of 1. It is possible for RPSS to be as strongly negative as -9 in the most severe case—when a probability of 1 (or 100%) is issued for a category and that category is not observed. It can also be noted that for a tercile-based system, when the observation is not in the given category, RPS_{cli} is always 0.111, and when the observation is in the given category (as in the above example), RPS_{cli} is always 0.444.