The Sensitivity of the Number of Correctly Forecasted Events to the Threat Score: A Practical Application

JOBY L. HILLIKER*

Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania

11 March 2003 and 31 December 2003

ABSTRACT

The sensitivity (S) of the number of correctly forecast events ("hits") to changes in threat score is explored. An emphasis is placed on the practical utility of S for end users of operational systems who may need to further assess a system’s performance whenever its threat score is provided. It is shown that S is a function of three parameters: (i) the actual threat score, (ii) forecast bias, and (iii) number of events in the sample size. All are easily obtainable if the forecasting system’s 2 x 2 contingency table is also provided.

It is also shown that the greatest change in S occurs for threat scores that are relatively low (i.e., close to 0). Conversely, a forecast system that performs well (i.e., has a high threat score) has the least amount of change in threat score. Thus, it is difficult to show continuous improvements for forecast products that perform well already. Examples demonstrating the practical utility of S are presented in this study.

1. Introduction

It is customary to present performance statistics for newly developed forecasting techniques. This information allows the techniques to be assessed through a relative perspective wherein one technique is compared to another using a consistent measure of performance. Performance statistics also permit techniques to be evaluated on their own merit (i.e., through an absolute perspective). For example, a system with a probability of detection (POD) of 0.701 may be viewed as unsatisfactory if the system were forecasting severe thunderstorms in the vicinity of airport runways. Moreover, if the forecasting system were improved, the user’s decision whether or not to accept the enhanced system is undoubtedly linked to his/her understanding of the performance measure(s) presented. If the thunderstorm forecast system previously mentioned were improved to perform at a POD of 0.80, and the user recognized that a 0.10 increase in POD translates into 10 additional thunderstorms (on average) being correctly forecast for every 100 thunderstorms observed, then the user has translated a performance measure into a parameter to which he/she can more readily relate—in this example, the number of correctly forecast events ("hits" hereafter). In some instances, however, considerably more complex performance measures2 are provided, and it is not readily evident to the user how to judge the practical implication of performance change.

One of the more commonly presented performance measures is the threat score, or critical success index (CSI; Bermowitz and Zurndorfer 1979). The definition of the threat score is shown as (1) in section 2. To review, the threat score ranges from zero to one (inclusive), with one representing a forecasting system that can forecast all events correctly, and zero indicating a system that cannot correctly forecast any event. It is now reasonable to draw from the POD example and assess the practical significance of an increase in threat score from 0.70 to 0.80. Does it again imply that 10 additional thunderstorms would be correctly forecast for every 100 observed? Furthermore, does a 0.10 change in the threat score from 0.20 to 0.30, let us say, imply a similar increase in the number of hits as that when the threat score changes from 0.70 to 0.80?

Clearly, the answers to these questions are not as evident as with the POD example, even with a fundamental knowledge of threat score. Similar questions could be addressed if supplemental performance measures are provided. Yet, regardless of the measures presented, the author poses this overriding question: What is the value of these complex measures to a user who

* Current affiliation: Cinergy Corp., Cincinnati, Ohio.

1 This value indicates that for every 100 events occurring, 70 (on average) would be forecast correctly.

Corresponding author address: Joby L. Hilliker, 139 East 4th Street, Mail Drop EA600, Cincinnati, OH 45202.
E-mail: jlh155@hotmail.com

© 2004 American Meteorological Society

2 Summaries of the more popular performance measures are presented in Wilks (1995).
may have limited experience with such measures of accuracy but, nevertheless, must make judgments of the performance of an operational system.

A cursory literature review indicates abundant interest in introducing and refining performance measures for a variety of applications (e.g., Gandin and Murphy 1992; Marzban 1998). Yet, there is a dearth of manuscripts that provide insight into the practical utility of these measures for operational users with all ranges of quantitative and/or scientific backgrounds and experience. A notable exception is from Fritsch et al. (1998), who provide a physical interpretation of threat scores for areal quantitative precipitation forecasts.

In light of the paucity of techniques for translating changes in accuracy measures to practical and direct indicators of change in performance of a forecasting system, an expression is derived—one that provides users the expected change in the number of hits for some change in threat score. Such an expression assesses the sensitivity of the number of hits to the threat score. The threat score is chosen as a template because of its popularity and relatively simple formula.

The objective of this sensitivity derivation is twofold: (i) to offer practical utility to users of an operational system whenever the threat score is presented, and (ii) to provide the meteorological community greater insight into this particular accuracy measure. After a review of some basic concepts and formulas, the next section details the derivation. The utility of the derived expression is then demonstrated in section 3 using two sets of forecasts from an example operational system. Finally, a summary and some concluding remarks are presented in section 4.

2. Sensitivity derivation

This section derives an expression that assesses the sensitivity of the number of hits to the threat score. This expression can then be used, with more practical usage, for the expected change in the number of hits for a specified change in threat score. Before the derivation is detailed, the following presents a review of concepts and formulas germane to the derivation.

Figure 1 shows the standard $2 \times 2$ contingency table, where $a$ represents the number of hits, $b$ the number of false alarms, $c$ the number of misses, and $d$ the number of correctly forecast nonevents.

<table>
<thead>
<tr>
<th>FORECAST</th>
<th>OBSERVED</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>a</td>
</tr>
<tr>
<td>NO</td>
<td>c</td>
</tr>
</tbody>
</table>

FIG. 1. The $2 \times 2$ contingency table, where $a$ represents the number of correctly forecast events (“hits”), $b$ the number of false alarms, $c$ the number of misses, and $d$ the number of correctly forecast nonevents.

$$T = \frac{a}{a + [B(a + c) - a] + c}. \quad (3)$$

A new parameter, $\delta$, is now introduced. It is defined as the total number of observed events in the sample, and should not be confused with the total number of cases in the sample (which would include both events and nonevents). The parameter $\delta$ can be expressed using elements of the $2 \times 2$ contingency table as

$$\delta = a + c. \quad (4)$$

The parameter $\delta$, referred hereafter as the “event frequency,” is representative of the event climatology and is assumed constant for the remainder of the derivation.

If (4) is rearranged in terms of $c$, and then substituted into (3), we obtain

$$T = \frac{a}{(B + 1)\delta - a}. \quad (5)$$

Note that the threat score is expressed in terms of forecast bias and the number of events in the sample, with the event frequency $(\delta)$ acting as a modulator.

Solving (5) for $a$ yields

$$a = \frac{(B + 1)T\delta}{1 + T}. \quad (6)$$

To obtain the sensitivity $(S)$ of the number of hits to the threat score, the derivative of $a$ with respect to $T$ is taken:

$$S = \frac{da}{dT} = \frac{B + 1}{(1 + T)^2}. \quad (7)$$

It is evident that $daldT$, or sensitivity function, is dependent on three parameters: bias, threat score, and event frequency. The properties of $S$ are now explored.

Figure 2 shows a plot of $S$ versus threat score for four different values of bias. Several important properties of the sensitivity function can be extracted from the figure. First, $S$ is maximized at $T = 0$ and minimized at $T = 1$ (with a value one-fourth the maximum), regardless of forecast bias. As an example, if a set of forecasts has no bias (i.e., $B = 1$), the maximum (minimum) value obtained is $2\delta$ (0.5$\delta$). Note that $S = \delta$...
Since verification statistics from different systems change in the number of hits, converting it into finite difference form and solving for the amplified (damped). Therefore forecasts (underforecasts) will have this effect amplifying forecast bias, although a system that consistently overforecasts (overforecasts) events (i.e., \( B > 1 \)) when compared to a system with no bias, regardless of threat score. A more significant property is obtained by exploring the slopes of the curves. Note the curves in Fig. 2 have the greatest slope (i.e., change in sensitivity) for relatively low threat scores (close to 0). The curves then flatten as the threat score approaches 1. This property is a reflection of the inverse squared relationship of threat score to the sensitivity function in (7). Its implication can be summarized as follows: If a system is performing poorly (thus having a relatively low threat score), an incremental change in threat score results in a relatively larger change in the number of hits. If the system is already performing well (thus having a relatively high threat score), the same incremental change in threat score results in a relatively smaller change in the number of hits. This statement is valid regardless of forecast bias, although a system that consistently overforecasts (underforecasts) will have this effect amplified (damped). A more practical version of (7) is obtained by converting it into finite difference form and solving for the change in the number of hits, \( \Delta a \):

\[
\Delta a = \frac{(B + 1)\delta}{(1 + T)^2} \Delta T,
\]

where \( \Delta T \) is the change in the threat score.

Note that \( \Delta a \) in (8) is dependent on the total number of events (\( \delta \)) from which the system was developed. Since verification statistics from different systems would likely be calculated from different sample sizes, both sides of (8) are normalized (i.e., divided) by \( \delta \):

\[
\frac{\Delta a}{\delta} = \Delta a' = \frac{(B + 1)}{(1 + T)^2} \Delta T.
\]

This additional step also allows the expected change in the number of hits to be expressed in terms of the total number of events in the sample—a form consistent with that used in the POD illustration in section 1. Hereafter, the expected change in the number of hits is modified to \( \Delta a' \) and is inherently a ratio to the total number of events.

Equation (9) represents the final version of the sensitivity expression, with its practical application demonstrated in the next section. It is, however, worth discussing beforehand the importance of a sufficiently large sample size when choosing a \( \Delta T \) in (9). Larger sample sizes ensure that the chosen \( \Delta T \) (and resultant \( \Delta a' \)) is statistically significant.

The concept of sampling variation can also be applied to the raw values of \( T \) (and \( B \)) in (9). Specifically, the greater the sample size, the greater the “confidence” that the purported estimate of \( T \) (i.e., sample \( T \), denoted as \( T' \) hereafter) will match the unknown population (or true) value (denoted as \( T \) hereafter). If threat score is viewed as a proportion or binomial parameter, this confidence can be expressed via a confidence interval (CI) for \( T \) using the formula (Harnett 1982):

\[
T' \pm z \sqrt{\frac{T'(1 - T')}{n}},
\]

where \( T' \) is the sample threat score, \( z \) is the standardized normal variable and is 1.96 if a 95% (i.e., \( \alpha = 0.05 \)) CI is desired, and \( n \) is the sample size [technically, \( a + b + c \), or the denominator in (1)]. If \( T' = 0.50 \), then a 95% CI for \( T \) for a sample size of 10 is [0.19, 0.81]. Note that for a much larger sample of 2000, the 95% CI (using an equivalent \( T' \)) significantly narrows to [0.48, 0.52].

For this application, however, a more useful version of (10) can be obtained by inverting the equation, and making a conservative estimate of the standard deviation, to solve for the optimal sample size (\( n \)) for some given tolerance (error) between \( T' \) and \( T \) (defined as \( D \)):

\[
n = \frac{z^2}{4D^2},
\]

where \( n \) and \( z \) are defined in (10). If \( D = 0.05 \) (such that a 95% CI for \( T \) could be [0.50 ± 0.05] or [0.45, 0.55]), then \( n \) in (11) is ~400. In other words, the purported threat score of a forecast system whose verification statistics have been compiled from a sample of 400 can be as much as 0.05 from its true threat score and still be statistically significant. Hence, one would need to choose a \( \Delta T \) greater than this threshold to ensure...
that the resultant change in $\Delta a'$ is statistically meaningful.

Yet, the above statement is valid only if sampling variations affecting $T$ were considered. Since (9) also contains a bias term, a more thorough analysis is required to account for sampling variations affecting $B$. This analysis becomes even more complex when the following are considered: (i) the interaction of sampling variations of $B$ with those of $T$, and (ii) the inverse squared relationship $T$ has to $\Delta a'$. Although such an exhaustive analysis dedicated to quantifying the impact of sampling variability on $\Delta a'$ is beyond the scope of this paper, (11) provides some guidance. In any event, a sufficiently large $\Delta T$ is recommended (especially for smaller sample sizes) when (9) is used so that the user avoids the risk of ascertaining a statistically insignificant value of $\Delta a'$.

3. Practical application of sensitivity expression

This section demonstrates the practical utility of the derived sensitivity expression in (9) using hypothetical verification statistics from forecasts generated by two different thunderstorm forecast systems. Table 1 shows the threat score, bias, $\delta$, and four values ($a$–$d$) that compose the $2 \times 2$ contingency table for each group of forecasts.

The operational user would likely note several aspects of each forecast system’s performance. First, the threat score is relatively higher (0.50) for the first system, and rather low (0.16) for the second system. Moreover, the bias is $>1$ for both forecast sets, indicating a tendency for the forecast system to overforecast convection. Also evident are the infrequency of observed thunderstorms and the overwhelming number of correctly forecasted non-events (i.e., $a + b + c \ll d$).

A practical question that may be posed is the significance of the 0.50 threat score with the first forecast group. One way to address this is to determine, quantitatively, its difference in performance from a system having a threat score of 0.60, for example. Addressed another way would be to determine the additional number of thunderstorms correctly forecast for this 0.10 increase in threat score.

To obtain these answers, values of bias and threat score are input into the sensitivity expression (9). After substitution, the expression simplifies to

$$\Delta a' = (0.947)(\Delta T),$$

indicating that the change in the number of hits is obtained by multiplying the desired incremental change in threat score by 0.947 ($S$, in essence). Thus, for a $+0.10$ change in threat score, $\Delta a' \sim 0.09$, meaning that nine additional thunderstorms are expected to be forecast correctly for every 100 storms observed.

A corresponding equation can also be obtained for the second forecast system using the relevant values from Table 1:

$$\Delta a' = (2.47)(\Delta T).$$

Note the value of $S$ for this second system has almost tripled (2.47 vs 0.947). Although this is partially the result of the increased bias of the forecasts, the bigger effect comes from the much lower threat score. This is consistent with the discussion in section 2, which highlighted the greater sensitivity for relatively low threat scores.

Thus, for the second set of forecasts, a 0.10 increase in threat score would mean that for every 100 thunderstorms observed, an additional $\sim 25$ ($\Delta a' \sim 0.25$) would be correctly forecast.

4. Summary and concluding remarks

The sensitivity of the number of correctly forecasted events (“hits”) to the threat score is explored. The motivation for presenting this study stems from a perceived need to provide practical utility of more complex performance measures specifically to end users of operational forecast systems. To address this objective, an expression is presented that translates a more complex accuracy measure (e.g., the threat score) into parameter(s) (e.g., number of hits) to which a user can more easily relate.

It is shown that the sensitivity ($S$) of the number of hits to the threat score is a function of three parameters: (i) the threat score itself, (ii) the bias in the set of forecasts, and (iii) the total number of events in the sample. Because $S$ is dependent on multiple parameters, it is more complex than the POD sensitivity to the number of hits [determined to be a constant—namely, the total number of events ($\delta$)]. In addition, the dependence on the above three parameters makes it imperative that the $2 \times 2$ contingency table, in addition to threat score, be available so that $S$ can be computed.

The maximum value of $S$ occurs at a threat score of 0, with the minimum at 1 (at one-fourth the maximum value), regardless of forecast bias or event climatology. Moreover, a system that overforecasts (underforecasts) events has a higher (lower) $S$ when compared to a system with no bias. The total number of events is also positively correlated to $S$.

Furthermore, the greatest change in $S$ occurs for relatively low threat scores, the result of its inverse squared relationship with threat score. Thus, for a given incremental increase in threat score, a relatively lower increase in the number of hits is realized for a forecast
system that performs well (when compared to one that performs poorly). The implication is that it becomes difficult to show continuous improvements for those systems that perform well already.

The threat score was specifically chosen for this exploration because of its popularity and relatively simple formula. Using this derivation as a template, similar (but more complicated) sensitivity expressions could be obtained for more complex performance measures (e.g., Heidke’s skill score, Kuipers skill score).

Acknowledgments. The author would like to thank George Young, J. Michael Fritsch, and Chetan Sharma for providing valuable input into the manuscript. The author also is indebted to the three reviewers for their suggestions and improvements. This work has been supported under the USWRP/National Science Foundation Grant ATM-9714154 and the U.S. Department of Transportation FAA Grant 2001-G-006.

REFERENCES