TORNADO-WARNING PERFORMANCE IN THE PAST AND FUTURE

A Perspective from Signal Detection Theory

BY HAROLD E. BROOKS

Signal detection theory provides insight into the conflicting goals of increasing probability of detection and decreasing false alarms for tornado warnings.

The National Weather Service (NWS) issues tornado warnings and collects observations to evaluate those warnings. Historically, the evaluation has consisted of the probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI). These quantities can be derived from a 2×2 contingency table (Table 1).[1] POD and FAR are clearly not independent of each other, and CSI provides no additional information. In practice, one could improve POD by issuing warnings on more storms, but that would almost certainly increase the FAR. Increasing POD while decreasing FAR at the same time requires improvements in scientific knowledge or technological application of that knowledge, or improvements in identifying events as tornadic or nontornadic. It would be nice to have a technique to estimate the effects of those changes of issuing additional warnings and improvements in science and/or technology.

In this paper, I will use signal detection theory (SDT) to develop a simple statistical model of NWS current and historical tornado-warning performance for the country as a whole. The model will look at the warning system as a black box, without regard to how any particular individual warning is made, and will focus on overall performance of the aggregate national warning system. This model will be applied to look at possible changes in performance as a result of increasing or decreasing the number of warnings, consistent with current performance, or changing the

AFFILIATIONS: BROOKS—NOAA/National Severe Storms Laboratory, Norman, Oklahoma
CORRESPONDING AUTHOR: Harold E. Brooks, NOAA/National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069
E-mail: harold.brooks@noaa.gov
DOI: 10.1175/BAMS-85-6-837

In final form 5 January 2004

[1] Technically, NWS verification procedures involve calculation of the POD on an event basis (a = total warned events, $a + c$ = total events), the FAR on an areal basis (b = warned counties with no event, $a + b$ = total warned counties), and then calculation of the CSI from the algebraic relationship between POD, FAR, and CSI for a "pure" 2×2 table. The two "a" values are technically not the same. For the purposes of this paper, that distinction will be ignored. In practice, if the CSI and one of the other two quantities is assumed to be true, small changes in the elements of the 2×2 table are required to make the values in the table internally consistent.

TABLE 1. The 2 × 2 contingency table for forecasts and observations and basic definitions (after Doswell et al. 1990).

Forecast		Observed: Yes	Observed: No	Sum
	Yes	a	b	$a + b$
	No	c	d	$c + d$
	Sum	$a + c$	$b + d$	1

Probability of Detection (POD) = $a/(a+c) = \mathrm{erfc}(x) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$ (yes Gaussian model)

False Alarm Rate (FAR) = $b/(a + b)$

Probability of False Detection (POFD) = $b/(b+d) = \mathrm{erfc}(x) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$ (no Gaussian model)

Critical Success Index (CSI) = $a/(a + b + c)$

Fraction of yes events, or climatological frequency $(f) = a + c$

quality of the warning system. Although changing the number of warnings could be done simply by changing decision thresholds, changing the quality of the warning system would require improvements in the basic understanding and application of that understanding, which is a much more challenging task.

SIGNAL DETECTION THEORY BACKGROUND. SDT provides a framework to analyze the performance of the schemes that identify events as "yes" or "no" with uncertain information. The application of SDT to forecast evaluation in meteorology was introduced by Mason (1982). Swets (1996) provides a more complete discussion. A model of the SDT problem involves considering the distribution of the weight of evidence associated with observed yes and no events. Then, a decision threshold is applied, with events being identified as yes or no depending upon whether the value of the weight of evidence is above or below the threshold (Fig. 1). In practice, the threshold for forecast decisions would typically be based upon real or perceived costs associated with

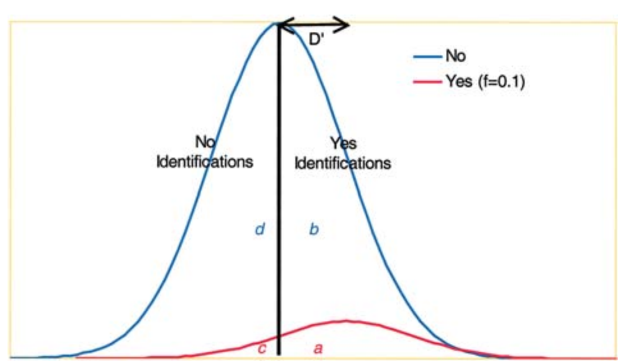

FIG. 1. Schematic model of decision problem. The red Gaussian curve represents the distribution of the value of some quantity associated with observed yes events, and the blue curve represents the distribution associated with no events. In the decision problem, events associated with observed values of the quantity to the right of the vertical line are identified as yes and those to the left are no. Thus, the portion of the red distribution to the right of the line represents correct detections of yes events, and the portion to the left of the line represents missed detections. Similarly, the portion of the blue distribution to the left of the vertical line represents correct detections of no events and the portion to the right represents false alarms. *D′* is the difference between the means in terms of the standard deviation of the two Gaussians. In the illustration, *D′* = 1. The location of vertical line is arbitrary and represents the decision threshold, *a, b, c,* and *d* indicate the regions to the right and left of the threshold associated with the elements of Table 1 with red associated with the yes Gaussian model and blue with the no Gaussian model.

misclassification of events, and then minimizing those over total costs. For any particular threshold, this produces a 2 × 2 contingency table.

The classification scheme can be evaluated in the whole by considering tables from the complete range of thresholds. A particularly powerful way to visualize the performance of the system over the complete range is via relative (or receiver) operating characteristic (ROC) curves (Mason 1982), which plot the POD versus the probability of false detection (POFD) as the decision threshold changes (Fig. 2). As discussed by Wilson (2000), many applications in different areas of decision analysis can be modeled, assuming that the distributions of weight of evidence for yes and no events are both Gaussian. The Gaussian model makes the calculation of POD and POFD simple. The POD is simply the fraction of the Gaussian model associated with yes events to the right of the threshold, and the POFD is the fraction of the Gaussian model associated with no events to the right of the threshold. In the case where the Gaussian models have the same variance, the distance in terms of number of standard deviations between the means of the two Gaussian models (D') provides a simple measure of discrimination between the two events.

Using the Gaussian model for the decisions, hypothetical contingency tables for different decision thresholds can be constructed. Because tornadoes are rare events, even when conditions are favorable enough to issue a warning, it is appropriate to consider the case where the yes events are less frequent than no events. For simplicity, I will assume that the two Gaussian models have their means one standard deviation apart ($D' = 1$), and the frequency of the yes event (f) is 0.1—a value that later will be shown to be consistent with historical tornado-warning performance. An unbiased (number of yes forecasts equals the number of events) forecast system meeting these criteria would produce POD = 0.33 and FAR = 0.67 (Table 2a). The POFD is 0.074 for this case, indicating that the probability of the yes forecast being made, given that an event occurs (POD), is more than 4 times as high as the probability when an event does not occur (POFD). This implies that the hypothetical forecast system has some ability to discriminate between situations when the event does and does not occur, implying that some users could benefit from the system.

Unbiased forecasts are not always desirable, however. If the costs associated with a missed event are higher than those associated with a false alarm, the decision threshold might be set at a much lower level than unbiased forecasts, producing a higher POD. If

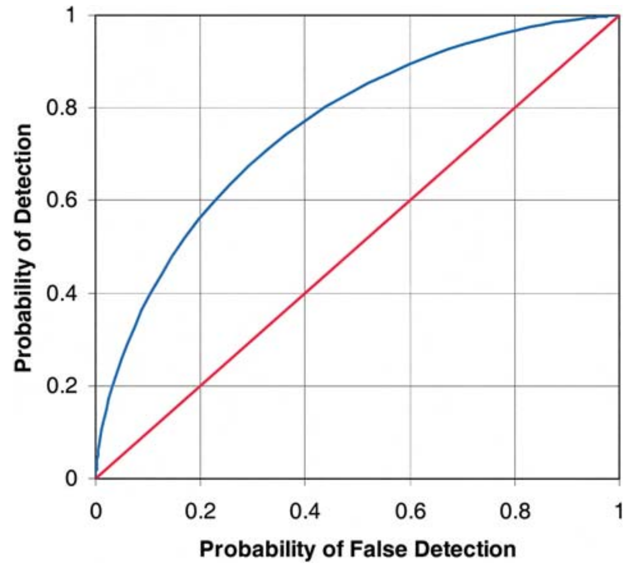

Fig. 2. The ROC curve associated with the model distribution shown in Fig. 1. The curved blue line is the plot of POD vs POFD for each decision threshold, and the 45° angle red line represents no skill.

Table 2. Contingency tables with $D' = 1$ and climatological frequency (f) = 0.1 for (a) unbiased forecasts, (b) forecasts with POD = 0.75, and (c) forecasts with FAR = 0.25.

(a)

		Observed		
		Yes	No	Sum
Forecast	Yes	0.033	0.067	0.1
	No	0.067	0.833	0.9
	Sum	0.1	0.9	1

POD = 0.33, FAR = 0.67, POFD = 0.074

(b)

		Observed		
		Yes	No	Sum
Forecast	Yes	0.075	0.335	0.410
	No	0.025	0.565	0.590
	Sum	0.1	0.9	1

POD = 0.75, FAR = 0.82, POFD = 0.372

(c)

		Observed		
		Yes	No	Sum
Forecast	Yes	0.0006	0.0002	0.0008
	No	0.0994	0.8998	0.9992
	Sum	0.1	0.9	1

POD = 0.006, FAR = 0.25, PFOD = 0.0002

the goal for POD was set at 0.75 for the same D' and frequency of yes events, or climatological frequency of the event (f), the resulting FAR would be 0.82 and the POFD would be 0.37 (Table 2b). If, on the other hand, the costs of false alarms are higher than that of missed events, the threshold might be set based on the FAR. The reduction in FAR that is associated with the increase in POD in the previous example would be to make it 0.25. The corresponding POD with that FAR would be 0.006, and the POFD would be 0.0002 (Table 2c). Thus, in this hypothetical situation, a low tolerance for false alarms (high costs) leads to very low POD values. If missed events are considered costly, however, much higher FAR values must be accepted. For a constant D', it is impossible to make improvements in POD and FAR at the same time.

MODEL. My goal is to develop a simple model of the tornado-warning system that reproduces much of the observed behavior. In one sense, this model treats the warning system as a black box, only considering the outputs, with no consideration of the process that goes into producing a warning. It will look only at the results of the behavior leading to warnings, not at the behavior itself. The model produces relationships between the various elements of the 2 × 2 table, with the assumption that the POD and FAR are known, in order to apply SDT to the warning evaluation problem. If the elements of the table are considered to be probabilities, so that $a + b + c + d = 1$, then three equations are required to determine all elements of the table. The POD and FAR relationships provide two of the equations, so only one more is necessary. A logical choice, given the two aspects on the ROC diagram, is to relate POFD to the other quantities. The fraction of all elements associated with yes events or climatological frequency $f = a + c$ is useful for the derivation.

The definitions of POD, FAR, and POFD provide the starting point. From the definitions of POD and FAR, we have $a = f\text{POD}$ and $b = a\text{FAR}$ (1 – FAR). Plugging in for a in the latter expression,

$$b = f\,\text{POD}\frac{\text{FAR}}{1-\text{FAR}}, \qquad (1)$$

and, thus,

$$\text{POFD} = \text{POD}\left(\frac{f}{1-f}\right)\left(\frac{\text{FAR}}{1-\text{FAR}}\right). \qquad (2)$$

With some manipulation, (2) becomes

$$\frac{1}{\text{FAR}} = 1 + \left(\frac{\text{POD}}{\text{POFD}}\right)\left(\frac{f}{1-f}\right). \qquad (3)$$

All four elements of the 2 × 2 table and, from that, any quantity associated with the 2 × 2 table, can be determined by knowing any three of the following: POD, POFD, FAR, and f.

A fundamental difficulty is that it is impossible to know with certainty how many correct forecasts of nonevents (element d of Table 1) there are. As a result, f is unknown without making some assumptions. To overcome this problem, the forecasts can be stratified (Murphy 1995). An appropriate stratification is to divide all possible warning situations into those that are trivially easy to determine whether there will be a tornado and those that require that a possibly difficult decision be made. For instance, a weak radar echo in the middle of winter when the atmosphere is below freezing at all levels is unlikely to be considered potentially tornadic, but a strong radar echo with a hook echo in the middle of a tornado watch will require that a decision be made about whether to issue a warning. It is assumed that almost no tornadic events would occur in the "trivially easy" situations, but there is no way of estimating that number. Focusing on the difficult situations, f can be considered to be the difficult situations that have a tornado. From the stratified perspective, an entry is made in one of the four elements of Table 1 each time a forecast (warning/no warning) and its corresponding observation (tornado/no tornado) are made. In this same context, D' can be thought of as a proxy for the quality of the total warning system in the sense that it measures how well tornadoes can be discriminated in the warning process. Note that for a particular threshold on Fig. 1, as D' increases, the area given by a increases, so that the POD increases at the same time that the FAR [$b/(a + b)$] decreases. The quality of the "warning system" includes, but is not limited to, the science of understanding the phenomena, development of spotter networks, the technology to look at the atmosphere, and the ability of the human forecasters to use the technology to apply the science to the decision problem at hand.

There is no obvious a priori way to determine D' and f. For a particular value of D', POD and POFD for any threshold x can be derived simply from the complementary error function $erfc(x)$, calculating the area to the right of the threshold for the Gaussian curve, from

$$erfc(x) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \qquad (4)$$

For any D', then, POD and POFD can be derived from the appropriate Gaussian model. By assuming a value for f, all four elements of the 2×2 table can be determined. Each value of D' will be associated with a particular ROC curve and, given the assumed value of f, a curve in the space of FAR and POD. The appropriate one for any application can be found by finding the line that goes through whatever values of FAR and POD are desired. Because D' and f are related quantities, they yield a unique curve in FAR–POD space for each combination. For a specified point in FAR and POD space, the relationship is such that a larger value of D' would be associated with a smaller value of f. In the STD framework, then, f is related to the difficulty of separating the two events, as measured by D'.

APPLICATION. As a starting point, the performance for the year 2001 is of interest. An infinite number of lines pass through POD and FAR of the year (POD = 0.69, FAR = 0.71). A selection of those lines is given in Fig. 3, each derived from a particular D' and f. In order to go through the point, small values of D' are associated with large values of f, indicating that the tornado-warning-decision problem is hard, but that relatively few storms are considered potentially tornadic. For $D' < 1.7$, the FAR changes more slowly than POD over a broad range of values. Also, as f decreases for a constant value of D', the line moves toward a higher FAR (not shown).

Some qualitative bounds on f can be drawn. Because the probability of a tornado given any detected circulation identified by the National Severe Storms Laboratory Mesocyclone Detection Algorithm (Stumpf et al. 1998) is ~0.01 (G. Stumpf 2003, personal communication), it seems unlikely that $f \leq 0.01$, so that $D \leq 2$. Assuming that human forecasters can outperform a low standard, such as any circulation, it seems more likely that f is on the order of 0.1, in which case $D' \sim 1.3$ for 2001.

Using the STD interpretation, moving toward the left on the POD–FAR curve is associated with raising the threshold for issuing a warning. Using that, I can estimate the effects of changing the decision threshold on POD and FAR. If it is decided that the goal for FAR is 0.50 and that changes will follow a single line on an ROC curve, the associated POD would be about 0.30 for $f = 0.1$.

Adding the performance statistics from 1986 to 2000 and 2002 provides additional insight into the warning system and provides support for the notion that $f \sim 0.1$ for the tornado-warning problem (Fig. 4). Clearly, performance in 2001 was about as good as any time in the period. The years from 1990 to 1998 and 2000 fall close to the line associated with $D' = 1$, $f = 0.1$, with the latter years being associated with a higher POD. Note that FAR increases very little along the line for most of the range. From POD = 0.30 to POD = 1.0, FAR only changes from 0.65 to 0.90. Assuming that the estimate of $f \sim 0.1$ is close to the truth, FAR is insensitive to large changes in the decision threshold for the current performance. As such, any

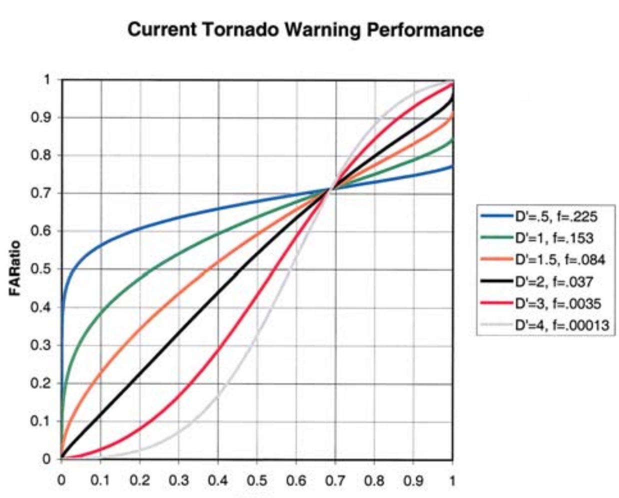

FIG. 3. Curves associated with different combinations of f and D' that pass through 2001 warning performance.

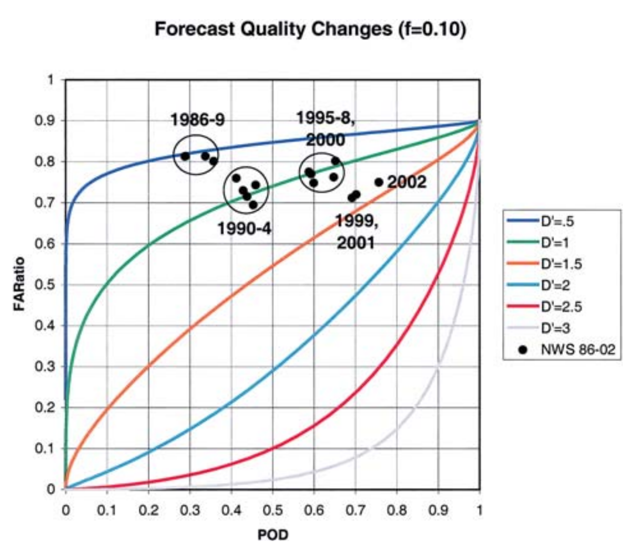

FIG. 4. Annual, national FAR and POD statistics for tornado warnings for each year from 1986 to 2002 with a variety of lines with constant D', assuming $f = 0.1$.

goals for performance associated with FAR are unlikely to be useful, unless the quality of the system (D') improves dramatically. If, for instance, a goal of POD = 0.80 and FAR = 0.50 is set for the warning system twenty years in the future, $D' \sim 2.2$ would be required.

Assuming $f = 0.1$, the NWS warning statistics can be plotted on a ROC diagram with curves for different D' values (Fig. 5) that are consistent with different eras. One possible interpretation of these curves is that performance improved from the late 1980s ($D' = 0.55$) into the 1990s ($D' = 1$). The change from the early 1990s to the late 1990s is consistent with a change in the threshold at which decisions are made, with more warnings being issued. However, as seen in Fig. 4, the primary effect was to improve POD, with a small increase in FAR. The years 1999, 2001, and 2002 clearly showed better performance than earlier periods lying on the $D' = 1.35$ line. Assuming that this value truly represents current performance, reaching the hypothetical future system with $D' = 2.2$ requires a change in D' over the next 20 yr at a rate equal to the change since the late 1980s.

CONCLUDING REMARKS. Some cautionary remarks are necessary. Performance varies from location to location and situation to situation, so that the relationships apply to overall national performance; inferences about particular situations must be made with care. In addition, nothing can be said about changes in lead time for warnings. There is little information on what an appropriate lead time is for optimal response and the simple model here cannot provide any insight. Decision models could be developed that estimate the value of changes in lead time, but they are far beyond the scope of the work here.

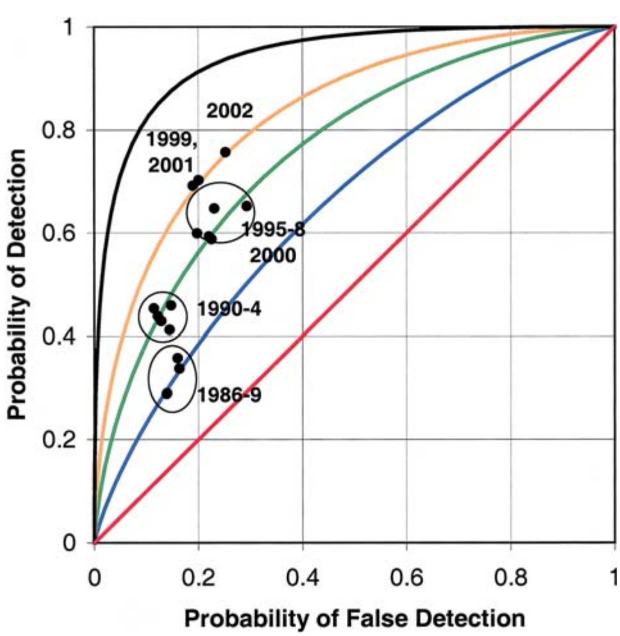

Fig. 5. ROC curves associated with historical tornado-warning performance ($D' = 0.55$ in blue, 1 in green, 1.35 in orange) and hypothetical future performance ($D' = 2.2$) associated with POD = 0.8 and FAR = 0.5.

Historically, it appears that NWS forecasters issuing tornado warnings have, on aggregate, behaved in a way that can be modeled by a relatively simple decision model. Improvements in tornado-warning performance can be demonstrated relatively easily. It appears that the current quality of the system is such that large reductions in FAR could only be accomplished by very large reductions in POD.

Future improvements in the quality of the warning system could change the relationship between FAR and POD. If D' increases enough, the POD will become less sensitive than FAR to changes in decision thresholds. Such an increase could occur if changes in D' continue to occur at the rate they have over the last 20 yr. The past 20 yr have seen a major field program to study tornadogenesis, the Verification of the Origin of Rotation in Tornadoes Experiment (VORTEX) (Rasmussen et al. 1994), deployment of the Weather Surveillance Radar-1988 Doppler network, a program to train forecasters on how to use the radar and make decisions using the new scientific understanding, and improvements in guidance forecasts from the Storm Prediction Center. The radar system's deployment is roughly coincident with the early 1990s' improvement in quality. The dissemination of tornado-warning guidance from the National Severe Storms Laboratory and the NWS's Warning Decision Training Branch, based, in part, on results from VORTEX, may be responsible for the improvement in the last few years. It is not clear what mix of changes in science, technology, training, and guidance would be necessary to lead to future major improvements in the quality of warnings. The simple model here cannot distinguish between changes in the various components of the system, only their effects in the aggregate. It seems likely that continued significant, or even enhanced, investments in all of the areas will be necessary. Large improvements in quality can occur, but they are unlikely to come for free. Performance in most of the 1990s represents a period where the quality was relatively constant, with only changes in the decision threshold.

ACKNOWLEDGMENTS. A seminar by Jeff Kimpel, director of the National Severe Storms Laboratory, about

future radar systems was the inspiration for this work. Discussions with a number of scientists from the various meteorological groups in Norman, Oklahoma, were helpful in formulating the ideas.

REFERENCES

Mason, I. B., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

Murphy, A. H., 1995: A coherent method of stratification within a general framework for forecast verification. *Mon. Wea. Rev.,* **123,** 1582–1588.

Rasmussen, E. N., R. Davies-Jones, C. A. Doswell III, F. H. Carr, M. D. Eilts, D. R. MacGorman, and J. M. Straka, 1994: Verification of the Origins of Rotation in Tornadoes Experiment: VORTEX. *Bull. Amer. Meteor. Soc.,* **75,** 995–1006.

Stumpf, G. J., A. Witt, E. D. Mitchell, P. L. Spencer, J. T. Johnson, M. D. Eilts, K. W. Thomas, and D. W. Burgess, 1998: The National Severe Storms Laboratory Mesocyclone Detection Algorithm for the WSR-88D. *Wea. Forecasting,* **13,** 304–326.

Swets, J. A., 1996: *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers.* Lawrence Erlbaum Associates, 308 pp.

Wilson, L. J., 2000: Comments on “Probabilistic predictions of precipitation using the ECMWF ensemble prediction system.” *Wea. Forecasting,* **15,** 361–364.