The Critical Success Index as an Indicator of Warning Skill

JOSEPH T. SCHAEFER

National Weather Service Central Region, Scientific Services Division

(Manuscript received 27 February 1990, in final form 20 July 1990)

ABSTRACT

A form of the critical success index (CSI) is used by the National Weather Service to indicate the value of warnings. This verification statistic assumes that the times when an event was neither expected nor observed are of no consequence. It can be shown that the CSI is not an unbiased indicator of forecast skill but is proportional to the frequency of the event being forecast. This innate bias is demonstrated theoretically and via example. An unbiased verification statistic appropriate for forecasts of rare events is presented and applied to severe convective weather warnings. Comparisons of this score to the CSI show the extent of the penalty the CSI extracts from forecasters who work in areas that are not climatically prone to given events.

1. Introduction

A form of the critical success index (Donaldson et al. 1975), the CSI, is used by the National Weather Service (NWS) as an indicator of the value of severe thunderstorm warnings (National Weather Service 1982). This verification statistic considers only those situations where a forecasting problem existed. In using the CSI, it is assumed that instances when an event was neither expected nor observed are of no consequence.

The CSI, under other names, has been proposed as a verification tool for over a century. One of its early applications was to the tornado forecasts reported by Finley (1884). In Finley's forecast experiment, scheduled tornado forecasts were produced for 18 geographic regions during the spring of 1884. A total of 2803 predictions were made. Of these, 100 were for tornadoes and the remainder were for "no-tornadoes." The associated storm data showed that 28 of the tornado forecasts and 2680 of the no-tornado forecasts verified. Accordingly, Finley claimed a verification rate of 96.6% (2708/2803). Gilbert (1884) noted that because this statistic was dominated by the no-tornado cases, it was not indicative of the skill of the forecasts. In fact, the verification rate would have increased to 98.2% (2752/2803) if all the experimental forecasts had been for no-tornadoes. In an effort to overcome this problem, Gilbert proposed using a score that ignores non-occurence forecasts that verify. Gilbert called this score the "ratio of verification," but its defining formula is the same as the one for the CSI.

Much later, Palmer and Allen (1949), while examining categorical precipitation forecasts, found that even in wet climates a vast majority of cases are no-precipitation-forecast/no-precipitation-observed episodes. Because of this, they felt that a verification score must ignore non-episodes. The "threat score" that they developed is the same statistic that had been used by Gilbert.

The question is whether ignoring no-forecast/no-event cases is fair to the forecaster. At times, the non-issuance of a warning (or forecast) is far from a mindless task. For instance, one of the factors that motivated Lemon to develop his radar severe thunderstorm identification criteria (Lemon 1977) was the frequent overwarning associated with conventional radar signatures of severe thunderstorms.

Further, as emphasized by Mason (1989), there is a strong dependence of the CSI on the occurrence frequency of the forecast event. In this paper, some of the work originally presented by Gilbert will be developed. It is demonstrated that the CSI is not an unbiased indicator of forecast skill. An unbiased version of the CSI is then developed and used to show the penalty that the CSI extracts from forecasters who work in areas that are not climatically prone to given events.

2. The CSI

Consider a set of forecasts that can have only two alternatives (e.g., yes, no). Let:

- X denote the number of positive (yes) forecasts that correspond to an occurrence of the event,
- Y denote the number of events that occurred in conjunction with a negative forecast,
- Z denote the number of positive forecasts that were not accompanied by an event, and
- W denote the number of negative forecasts that did not have any associated events.

If only forecasts of event occurrence are considered

Corresponding author address: Joseph T. Schaefer, NOAA, 601 E. 12th Street, Room 1836, Kansas City, MO 64106.

significant, X is the number of hits, Y is the number of misses, and Z is the number of false alarms. A four-cell contingency table can be constructed which depicts the relationship between the forecasts and the events (Table 1). Further, as can be seen from Table 1, the total number of positive forecasts (P) is

$$P = X + Z. \quad (1)$$

The total number of events (E) is

$$E = X + Y. \quad (2)$$

The total number of cases (T) is

$$T = X + Y + Z + W, \quad (3)$$

and the frequency (F) of the event is

$$F = E/T.$$

These counts are used to form several standard verification statistics. The probability of detection (POD), or the prefigurance (Panofsky and Brier 1965), is simply the ratio of events that are correctly forecast to occur to the total number of events:

$$\text{POD} = X/(X + Y) = X/E. \quad (4)$$

Simply stated, the POD is the percent of events that are forecast.

The false alarm rate (FAR) is a measure of the failure of the forecaster to exclude non-event cases, and is the ratio of the number of false alarms to the total number of predicted events. More formally, it is the ratio of unsuccessful positive forecasts to the total number of positive forecasts:

$$\text{FAR} = Z/(X + Z) = Z/P. \quad (5)$$

Stated positively, rather than negatively, the success ratio (SR), or post agreement, is defined as the ratio of hits (correct positive forecasts) to the total number of event forecasts:

$$\text{SR} = X/(X + Z) = X/P = 1 - \text{FAR}.$$

The CSI (or ratio of verification or threat score) is simply the ratio of successful event forecasts (X) to the total number of event forecasts that were either made ($X + Z$) or needed (Y):

$$\text{CSI} = X/(X + Y + Z) = X/(P + E - X). \quad (6)$$

In the simplest of terms, the CSI is the ratio of the number of hits to the number of events plus the number of false alarms. It varies directly with the number of correct event forecasts (hits), and varies inversely with both the number of incorrect event forecasts (false alarms) and the number of missed events (Y). However, as has been previously noted, the number of correct nonforecasts of the event (W) does not affect the CSI.

TABLE 1. Four-cell contingency table—see text for details.

		FORECASTS YES	FORECASTS NO	
EVENTS	YES	X	Y	$E = X + Y$
	NO	Z	W	$T = X + Y + Z + W$
		$P = X + Z$		

With a little algebraic manipulation, the CSI formula can be expressed in terms of the false alarm rate and the probability of detection as:

$$\text{CSI} = [(\text{POD})^{-1} + (1 - \text{FAR})^{-1} - 1]^{-1}. \quad (7)$$

While this formulation shows that the dependence of the CSI on either the FAR or the POD is highly nonlinear, it can give the false impression that the CSI is undefined if either the POD equals zero or if the FAR equals unity. This is not the case! For either a zero POD or a unit FAR, the value of X must be zero (there are no hits) and the CSI is uniquely equal to zero.

However, there is a problem with the CSI when it is used as a tool for comparative verification of forecasts. Since (3) can be written as

$$X + Y + Z = T - W,$$

the first formulation of (6) is equivalent to

$$\text{CSI} = X/(T - W).$$

The CSI is functionally related to the relative size of W as compared to T. As events become more frequent, $(T - W)$ decreases and the CSI increases. Conversely, as events become rarer, $(T - W)$ increases and the CSI decreases. The CSI is a *biased score* that is dependent upon the frequency of the event that is forecasted.

3. A skill dependent CSI

As defined in the *Glossary of Meteorology* (Huschke 1959), a skill score is "an index of the degree of skill of a set of forecasts, expressed with reference to some standard such as forecasts based upon chance, persistence, or climatology." The comparison of the forecasts to a reference is the most significant feature of a skill score. It removes the element of serendipity from verification.

If forecasts are made at random, some number (C) will verify by chance. Accordingly, if skill is to be considered, the verification statistics must be modified to account for hits due to chance. Counting only the hits *not* due to chance, the skill corrected success ratio (SR_s) is:

$$\text{SR}_s = (X - C)/(X - C + Z) = (X - C)/(P - C).$$

If only forecasts for event occurrence are considered, SR_s is the same as the Heidke Skill Score (Brier and Allen 1951). Similarly, the POD can be modified to account for chance verification. The skill corrected POD is:

$$\mathrm{POD}_s = (X - C)/(X - C + Y) = (X - C)/(E - C).$$

A skill corrected CSI, which we will name the Gilbert Skill Score, is obtained by substituting $(X - C)$ for X in (6). Its formulation is:

$$\mathrm{GS} = (X - C)/(X - C + Y + Z) = (X - C)/(P + E - X - C). \quad (8)$$

This score is simply the number of correct forecasts in excess to those that would verify by chance, divided by the number of cases when there was a threat that would not have been foreseen by chance. Examination of (8) shows that zero skill occurs when the number of correct positive forecasts (X) is the same as the number of hits by chance (C). Also, negative skill occurs when X is less than C (i.e., the forecaster actually has a negative impact). The maximum score is unity, and is obtained when events and forecasts of the event correlate perfectly. (Only positive forecasts are associated with an event, *and* all events occur with a positive forecast, i.e., $X = Z = 0$.)

The number of fortuitously correct forecasts (C) is simply the event frequency (E/T) multiplied by the number of forecasts of event occurrence:

$$C = P \cdot E/T = (X + Y) \cdot (X + Z)/(X + Y + Z + W). \quad (9)$$

Inserting this into (8) gives the Gilbert Skill Score as:

$$\mathrm{GS} = [X - P \cdot (E/T)]/[(P + E - X) - P \cdot (E/T)]. \quad (10)$$

It can also be expressed in terms of the cells of the contingency table:

$$\mathrm{GS} = (X \cdot W - Y \cdot Z)/[(Y + Z) \times (X + Y + Z + W) + (X \cdot W - Y \cdot Z)]. \quad (11)$$

From (11) it can be shown that the minimum GS occurs when (1) no correct forecasts are made ($X = W = 0$), *and* when (2) the number of missed events is equal to the number of bad forecasts ($Y = Z$). Since there is no readily apparent reason why the second of these conditions should indicate the least possible skill in a set of forecasts, the fact that the minimum GS is $-\frac{1}{3}$ rather than -1 is not particularly troubling. In interpreting negative Gilbert Skill Scores, it is sufficient to note that the more negative the score, the worse the forecaster performed relative to chance at foreseeing the event.

As an aside, it is also possible to compute a skill score considering only forecasts of no-event ($W + Y$) and observations of no-occurence ($W + Z$). It can be shown that the GS for negative forecasts is also given by (10). The Gilbert Skill Score is the same if either positive or negative forecasts are considered.

The Gilbert Skill Score has the property of having a value of zero if either an event is always forecast ($P = T$ so that $X = E$), or if an event always occurs ($E = T$ so that $X = P$). This is not the case with the CSI. Also for extremely rare events, (E/T) approaches zero,

$$\lim_{E/T \to 0} \mathrm{GS} = X/(P + E - X) = \mathrm{CSI}.$$

Thus, the Gilbert Skill Score approaches the CSI as the event forecast becomes rarer. However, the Gilbert Skill Score must always be lower than the CSI. The event frequency (E/T) determines how close the two scores are to each other. Because the fraction obtained by subtracting the same quantity from both the numerator and denomintor of a positive proper fraction is smaller than the original fraction, (10) demonstrates that *for a given CSI, the skill decreases as the forecast event becomes more frequent.*

4. A numerical example

To get an idea of the amount of the bias, let us examine Finley's 1884 forecast experiment in detail. As previously noted, Finley made 100 tornado forecasts (P), of which 28 verified (X) and 72 did not (Z). He also made 2703 no-tornado forecasts, of which 2680 were correct (W) and 23 were wrong (Y). A contingency matrix based upon these numbers is shown in Table 2. According to the basic definitions, the experiment produced the following verification statistics:

$$\mathrm{POD} = 0.549,$$
$$\mathrm{FAR} = 0.720,$$
$$\mathrm{CSI} = 0.228.$$

Using the 1.8% tornado frequency that was observed during the experiment, we see that if the 100 tornado forecasts had been made purely at random, 1.82 of them would have been correct; i.e.,

$$C = 1.82.$$

Thus, Finley's Gilbert Skill Score was

$$\mathrm{GS} = 0.216.$$

This score is only 0.012 lower than the CSI. This is not unexpected since the ratio (E/T) is only 0.018.

In contrast, let us consider what would have happened if the frequency of tornadoes during Finley's forecast experiment had been 11.8% rather than 1.8%.[1] Further, let us assume that forecast methods used by Finley would have resulted in exactly the same POD, FAR, and CSI. This amounts to maintaining the CSI statistics while artificially increasing the frequency of the event. This is done by changing T from 2803 to 432. The contingency table for this thought experiment

[1] Court (1970) notes many significant sources of underestimation in tornado enumerations. When the population and communications of the 1880s is considered, this fictitious frequency might very well be closer to reality than the "official" one.

TABLE 2. Contingency Matrix for Finley's Experiment.

		FORECASTS YES	NO	
EVENTS	YES	$X = 28$	$Y = 23$	$E = 51$
	NO	$Z = 72$	$W = 2680$	$T = 2803$
		$P = 100$		

is given in Table 3. Using the table values, we note that the number of serendipitous occurrences increases so that

$$C = 11.8,$$

and that the Gilbert Skill Score decreases to

$$GS = 0.146.$$

Remember, we held the CSI at the value Finley received:

$$CSI = 0.228.$$

The large increase in event frequency caused the skill to decrease markedly even though the CSI remained constant.

Now let us consider how the CSI would have to change in this second environment (where the event frequency is 11.8%) in order for Finley to maintain his Gilbert Skill Score (GS = 0.216) over the same number of positive forecasts and observations. The event and forecast distributions required to obtain the desired frequency and Gilbert Skill Score (GS) are shown in Table 4 (the entries in the table have been rounded to integers). From this contingency table, the statistics of the forecasts would be:

$$POD = 0.667,$$

$$FAR = 0.660,$$

$$CSI = 0.291,$$

Because of the higher frequency, a 27% increase of the CSI is required for the set of forecasts to have the same skill as Finley's.

TABLE 3. Contingency Matrix if the tornado frequency during Finley's Experiment had been 11.8% rather than 1.8%.

		FORECASTS YES	NO	
EVENTS	YES	$X = 28$	$Y = 23$	$E = 51$
	NO	$Z = 72$	$W = 309$	$T = 432$
		$P = 100$		

TABLE 4. Contingency Matrix for a tornado frequency of 11.8% with the number of positive forecasts (100) and events (51) reported by Finley.

		FORECASTS YES	NO	
EVENTS	YES	$X = 34$	$Y = 17$	$E = 51$
	NO	$Z = 66$	$W = 315$	$T = 432$
		$P = 100$		

5. Estimation of local severe thunderstorm potential

In order to compute an inclusive verification score, a count of no-event forecasts which verify (W) is needed (Flueck 1987). However, this quantity is known only for routinely issued yes/no forecasts at specific points. In contrast, many products, such as severe thunderstorm and tornado warnings (truly yes/no forecasts), are issued only when and where they are needed. However, if some assumptions are made, it is possible to estimate how often the non-issuance of a warning requires a conscious decision. This count can then be used to obtain an approximation to the Gilbert Skill Score.

Radar observations play an important role in the warning process. In a study of New England storms, Donaldson (1958) noted that most damaging windstorms and tornadoes are associated with parent echoes that extend above 40 000 feet. Further, Darrah (1978) found that only about 1% of thunderstorms with radar tops lower than 40 000 feet were severe (including storms producing hail of at least ¾ inch diameter). However, a problem with using this height as a threshold for severe weather potential is that non-severe thunderstorms frequently grow to great heights in regions south of the mean track of the subtropical jet stream where a high tropical tropopause is the norm (Lee and Galway 1956). Thus, if Florida and the immediate coastal areas of the southeast United States are excluded, the existence of radar echoes at heights greater than 40 000 feet can be used as a primitive discriminator of conditions during which severe convective storms are possible.

Because of the necessity to continually determine the exact geographic location of storms, check their motion, forecast their short term trajectories, compose the warning messages, etc., it is not unreasonable to postulate that warnings are issued at a maximum rate of once every 10 min. Thus, for all but the extreme southeastern United States, a crude approximation is that about six warning decisions are required for every hour that a radar echo above 40 000 feet is observed within an office's area of responsibility. While one can argue the details of this estimate (i.e., there is often more than one severe storm in existence at the same time; individual warnings often last for an hour), it

does provide a basis for gauging the frequency of warning decisions at various offices around the country.

6. Severe convective storm warning Gilbert Skill Scores

The NWS routinely computes verification statistics (i.e., Grenier et al. 1989) for combined severe convective storm warnings (a combination of severe thunderstorm and tornado warnings). Among the statistics are the number of severe thunderstorm events (E in Eq. 2), the number of events during which a warning was in effect (X in Eq. 2), the POD, and modified versions of the FAR and CSI. The modifications are necessary because warnings are issued for areas rather than individual points. In computing the FAR, the NWS assumes that the basic unit of area is the county, and the number of warnings or positive forecasts (P in Eq. 1) is equivalent to the total number of counties warned during the period of consideration (Pearson and David 1979). The modified CSI reported in the verification statistics is computed via (7).

From these statistics and the local radar climatology it is possible to estimate the values needed to construct a four-cell contingency table for individual offices. As noted, X is given. Y is obtained from (2). An estimate of Z comes from inverting (5), using X and the modified FAR. Finally, the climatological frequency distribution of the existence of radar echo at 40 000 feet and above, within a 100-mile radius of various stations, (Grantham and Kantor 1967) gives an estimate of the total number of warning decisions made at an office (T). Equation 3 will then yield W, and the Gilbert Skill Score can be evaluated using (8) and (9). These operations are summarized in Table 5.

As an example, consider Minneapolis, Minnesota. This is a typical noncoastal station that is subject to occasional severe convective storms, but is not in the climatological area where severe thunderstorms are most frequent (Kelly et al. 1985). At Minneapolis, radar echo at 40 000 feet is observed on 3.3% of all hourly observations. Thus, to a first approximation, 1734 warning decisions are implicitly, or explicitly, made per year (T in Eq. 3). During 1988, there were 35 severe convective storms within the Minneapolis area of responsibility. Warnings were issued for 21 of these storms. The office had a FAR of 0.702, a POD of 0.600, and a CSI of 0.250. Table 6 gives the resulting contingency matrix for Minneapolis' 1988 severe thunderstorm/tornado warnings. Because of the large size of the no forecast/no event category, 1.4 correct warnings could have been issued by random selection alone. The GS for the warnings is 0.237. Since T is much larger than P and E, the CSI only inflates skill by 5.5%.

The potential magnitude of the difference between the CSI and the Gilbert Skill Score is shown by examining the data from the office that encountered the most severe convective storm activity during 1988. Oklahoma City, Oklahoma's area for warning responsibility recorded 405 severe weather reports. The office had a POD of 0.810 and a FAR of 0.347. These statistics yield a CSI of 0.566 (Table 7). On the other hand, the radar climatology indicates the presence of a 40 000 foot echo on 5.3% of hourly observations. This implies that the frequency of events (E/T) is 14.5% and the ratio of warnings to warning situations (P/T) is 18.0%. Thus T is much less than an order of magnitude larger than both E and P, and the Gilbert Skill Score should be significantly lower than the CSI. Indeed, working through the mathematics shows the GS equals 0.504 which is 0.062 smaller than the CSI. For this high-frequency severe weather region, CSI inflates warning skill by 12.3%.

TABLE 5. Source of Contingency Table Elements for estimation of Gilbert Skill Score for NWS warnings.

FAR is given in Grenier et al. (1989).
E is the number of events given in Grenier et al. (1989).
T comes from radar echo climatology (see section 5).
X is the number of verified warnings given in Grenier et al. (1989).
$Y = E - X \cdots$ (from Equation 2).
$Z = [\text{FAR}/(1 - \text{FAR})] \cdot X \cdots$ (from equation 5).
$W = T - (X + Y + Z)$

TABLE 6. Contingency Matrix for the severe thunderstorm/tornado warnings issued by WSFO Minneapolis, Minnesota during 1988 (numbers have been rounded to integers).

		FORECASTS: YES	FORECASTS: NO	
EVENTS	YES	$X = 21$	$Y = 14$	$E = 35$
	NO	$Z = 49$	$W = 1650$	$T = 1734$
		$P = 70$		

7. Discussion

It has been demonstrated that the CSI is not an unbiased measure of forecast skill. The CSI is an overestimate of the skill, and the magnitude of the over-

TABLE 7. Contingency Matrix for the severe thunderstorm/tornado warnings issued by WSFO Oklahoma City, Oklahoma during 1988 (numbers have been rounded to integers).

		FORECASTS: YES	FORECASTS: NO	
EVENTS	YES	$X = 328$	$Y = 77$	$E = 405$
	NO	$Z = 174$	$W = 2207$	$T = 2786$
		$P = 502$		

estimation increases as the frequency of the event being forecast increases. Because of this, it is not appropriate to rate the skill of various offices by their CSI.

The Gilbert Skill Score presented in this paper is only one of many skill-indicating verification scores available. In Doswell et al. (1990), which appears in this issue of *Weather and Forecasting,* two such scores (the True Skill Score and the Heidke Skill Score) are examined in detail and compared to the CSI. The Gilbert Skill Score (11) and their formulation of the total Heidke Skill Score (S) are related by the equation:

$$GS = S/(2 - S).$$

The difference in scores results from a stronger dependence upon the no-forecast/no-event category (W) in the Heidke score than in the Gilbert Skill Score. Woodcock (1976) presented a review of eight other standard verification measures; and McCoy (1986) discussed verification using signal detection theory. These measures are also dependent upon the value for W.

For warnings, as they are now issued by the National Weather Service, a true skill score cannot be obtained. Because of its simplicity, the Gilbert Skill Score can be estimated if the relative frequency of the forecast event is available. This paper has illustrated how climatological data can be used to give a rough estimate of the skill of severe convective storm warnings. Similar techniques should be developed for all other warning products issued. However, it must be stressed that *no single statistic can adequately depict all the attributes of an office's warning program.* The POD, the FAR, the lead time, and even the degree that individual storms reflect "text book" conditions must be considered when attempting to compare the quality of warnings issued by different offices.

Finally, a word in the defense of the CSI is in order. This score is a valid indicator of the relative worth of different forecast techniques when they are applied to the same environment. With the operational implementation of the WSR-88D (NEXRAD) radar network, individual stations will have to evaluate the appropriateness of the various algorithms to their particular locale. The CSI is an appropriate tool to do this. However, because of the dependence of the score on event frequency, problems arise when the CSI is used to gauge the values of different offices.

Acknowledgments. I would like to express my appreciation to the many colleagues with whom I have discussed this topic during the extended development period of the manuscript. Special mention must go to Wayne Sangster (NWS Central Region-retired), Richard McNulty (WSFO Topeka), Paul Polger (NWS Headquarters), John Hughes and Richard Livingston (NWS Central Region, SSD), and Fred Mosher, Preston Leftwich, and Pete Browning (NSSFC). Particular thanks is due to Ralph Donaldson (ST Systems Corp.) whose constructive criticism of the manuscript improved it markedly. Beverly Lambert's many editorial and word processing skills were invaluable.

REFERENCES

Brier, G. W., and R. A. Allen, 1951: Verification of weather forecasts. *Compendium of Meteorology.* Amer. Meteor. Soc., 841–848.

Court, A., 1970: *Tornado Incidence Maps.* ESSA Technical Memorandum ERLTM-NSSL 49, 76 pp.

Darrah, R. P., 1978: On the relationship of severe weather to radar tops. *Mon. Wea. Rev.,* **106,** 1332–1339.

Donaldson, R. J., Jr., 1958: Analysis of severe convective storms observed by radar. *J. Meteor.,* **15,** 44–50.

——, R. M. Dyer and M. J. Kraus, 1975: An objective evaluator of techniques for predicting severe weather events. Preprints: *Ninth conference on severe local storms (Norman, OK),* Amer. Meteor. Soc., Boston, 321–326.

Doswell, C. A. III, R. P. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. and Forecasting,* **5,** in press.

Finley, J. P., 1884: Tornado Predictions. *American Meteorological Journal,* **1,** 85–88.

Flueck, J. A., 1987: A study of some measures of forecast verification. Preprints: *Tenth conference on probability and statistics in the atmospheric sciences (Edmonton, Alta., Canada),* Amer. Meteor. Soc., Boston, 69–73.

Gilbert, G. F., 1884: Finley's Tornado Predictions. *American Meteorological Journal,* **1,** 166–172.

Grantham, D. D., and A. J. Kantor, 1967: *Distributions of radar echoes over the United States.* Air Force Surveys in Geophysics, No. 191, Air Force Cambridge Research Laboratories, L. G. Hanscom Field, Bedford, MA, 375 pp.

Grenier, L. A., J. T. Halmstad and P. W. Leftwich, Jr., 1989: *Severe Local Storm Warning Verification: 1988.* NOAA Technical Memorandum NWS NSSFC-23, 19 pp.

Huschke, R. E., 1959: *Glossary of Meteorology.* Amer. Meteor. Soc., Boston, 638 pp.

Kelly, D. L., J. T. Schaefer and C. A. Doswell III, 1985: Climatology of non-tornadic severe thunderstorm events in the United States. *Mon. Wea. Rev.,* **113,** 1997–2014.

Lee, J. T., and J. G. Galway, 1956: Preliminary report on the relationship between the jet at 200-mb level and tornado occurrence. *Bull. Amer. Meteor. Soc.,* **37,** 327–332.

Lemon, L. R., 1977: *New severe thunderstorm radar identification techniques and warning criteria: A preliminary report.* NOAA Technical Memorandum NWS NSSFC-1, 60 pp.

McCoy, M. C., 1986: Severe-storm-forecast results from the PROFS 1983 forecast experiment. *Bull. Amer. Meteor. Soc.,* **67,** 155–164.

Mason, I., 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Meteor. Mag.,* **37,** 75–81.

National Weather Service, 1982: *National Verification Plan.* U.S. Department of Commerce, NOAA, 81 pp.

Palmer, W. C., and R. A. Allen, 1949: Note on the accuracy of forecasts concerning the rain problem. Weather Bureau Manuscript, Washington, D.C., 2 pp.

Panofsky, H. A., and G. W. Brier, 1965: *Some applications of statistics to meteorology.* The Pennsylvania State University, University Park, PA, 224 pp.

Pearson, A. D., and C. L. David, 1979: Tornado and severe thunderstorm warning verification. Preprints: *Eleventh conference on severe local storms (Kansas City, MO),* Amer. Meteor. Soc., Boston, 567–568.

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.,* **104,** 1209–1214.