Long-Term Performance Metrics for National Weather Service Tornado Warnings

HAROLD E. BROOKS

NOAA/National Severe Storms Laboratory, and School of Meteorology, University of Oklahoma, Norman, Oklahoma

JAMES CORREIA JR.

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/NWS/Storm Prediction Center, Norman, Oklahoma

(Manuscript received 18 July 2018, in final form 17 August 2018)

ABSTRACT

Tornado warnings are one of the flagship products of the National Weather Service. We update the time series of various metrics of performance in order to provide baselines over the 1986–2016 period for lead time, probability of detection, false alarm ratio, and warning duration. We have used metrics (mean lead time for tornadoes warned in advance, fraction of tornadoes warned in advance) that work in a consistent way across the official changes in policy for warning issuance, as well as across points in time when unofficial changes took place. The mean lead time for tornadoes warned in advance was relatively constant from 1986 to 2011, while the fraction of tornadoes warned in advance increased through about 2006, and the false alarm ratio slowly decreased. The largest changes in performance take place in 2012 when the default warning duration decreased, and there is an apparent increased emphasis on reducing false alarms. As a result, the lead time, probability of detection, and false alarm ratio all decrease in 2012.

Our analysis is based, in large part, on signal detection theory, which separates the quality of the warning system from the threshold for issuing warnings. Threshold changes lead to trade-offs between false alarms and missed detections. Such changes provide further evidence for changes in what the warning system as a whole considers important, as well as highlighting the limitations of measuring performance by looking at metrics independently.

1. Introduction

Tornado warnings are one of the most important products issued by the National Weather Service (NWS). They provide potentially lifesaving information in situations that involve decision-making under uncertainty with short times to make decisions and with the potential for great costs associated with errors. As such, the performance has been studied within a variety of contexts, such as error analysis (e.g., Brotzge and Erickson 2009, 2010; Brotzge et al. 2011), costs and benefits of warnings (Simmons and Sutter 2005, 2008; Sutter and Erickson 2010), and within a theoretical decision analysis framework (Brooks 2004). Changes in official definitions have taken place, most notably the change from so-called county-based warnings to storm-based warnings that took place in October 2007, which also included changes in evaluation methodology. No changes in the evaluation methodology took place after the initial changes with the beginning of the storm-based era.

The NWS reports official statistics for tornado warnings under the mandate of the Government Performance and Reporting Act (GPRA; Ralph et al. 2013), which sets goals for the probability of detection, false alarm ratio, and lead time for warnings. The issue of how those quantities are defined will be discussed later. The interrelationships between the GPRA metrics and other performance measures are complex. We will examine the performance over the period from 1986 to 2016, the full record of warning performance available from the NWS Performance Management website (https://verification.nws.noaa.gov/services/public/index.aspx).

Given the changes in the ways tornadoes have been reported and warnings have been created (county-based or storm-based warnings, software, etc.), as well as official evaluation metrics, it is challenging to look at the warnings in a consistent manner. We will focus on information that is available throughout the 1986–2016 period with the same definitions for metrics throughout the period regardless of the official definitions in use

Corresponding author: Harold E. Brooks, harold.brooks@noaa.gov

DOI: 10.1175/WAF-D-18-0120.1

at the time. This will allow us to identify changes in warning philosophy and see when significant changes occurred. It is important to note that we are looking neither at individual warning decision-making nor focusing on the transition to storm-based warnings. Instead, we will use what we call an archaeological approach to identify what the ''culture'' of the NWS implicitly values. Describing the field of archaeology, McIntosh (1986) cites A. H. Pitt-Rivers's view that ''it is the study of the ordinary everyday things that helps us to reconstruct the past, far more so than rare, valuable objects that were unusual even in their own time and place.'' The culture depends on official policy, as well as operational practice for discretionary activities within the official policy. As an example, an official policy might prescribe upper and lower limits on the duration of warnings, but operational practice could lead to the actual limits used by forecasters being smaller than the official bounds. The constraints on how tornado warnings are structured (e.g., area covered, duration) are relatively loose and allow for local offices to tailor products to the perceived needs of their local areas or the situation at hand. By looking at the warning products created over the years across the country, we can learn something about what has been considered important in different eras. Although it is beyond the scope of this paper, the methodology described could be used to look at changes in space or type of warning situation. We follow that by looking at the full collection of warnings over 31 years, rather than specific cases. As will become clear, major changes in warning structure and performance are revealed by this approach. In particular, the relationship between probability of detection and false alarms as it has evolved in practice can be seen and how these factors could be evaluated in concert, rather than separately, can be considered. As a result, in addition to our interest in the performance of the tornado warning system, we are interested in the more general problem of how to develop techniques that can allow the community to monitor and detect changes in performance over time. The tornado warning system is an ideal candidate to consider such techniques.

2. Methodological background

The NWS tornado warning verification database was obtained 23 June 2017 covering the period from 1 January 1986 to 30 September 2007 for county-based warnings and from 1 October 2007 through 2016 for storm-based warnings. For the county-based era, the beginning and ending times for each warning and tornado and the county they were valid for are available, as is the initial lead time for each warned tornado, and the F-scale damage rating for each tornado. For the storm-based era, each warning has a beginning and ending time and a verifying event identifier, corresponding to none, one, or multiple tornadoes contained within the issuance polygon. The tornadoes have beginning and ending times, beginning and ending locations, the number of 1-min segments, the number of warned segments, the initial lead time, an EF-scale damage rating, and corresponding warning identifiers. Tornado occurrences are counted on a county-by-county basis.

To make the differences between the possible definitions of probability of detection (POD) and lead time clear, we begin by considering three cases of warnings given that a tornado occurs. For a particular tornado, either 1) a warning was issued before the initial time of the tornado; 2) a warning was issued after the tornado begins, but before it ends; or 3) a warning was never issued before or during the tornado. For case 1, it seems clear that, for the purposes of evaluating the POD, the warning would be classified as a hit. Similarly, for case 3, it is clear that the tornado would be classified as a missed detection. The correct classification for case 2 is ambiguous. Officially, prior to the adoption of storm-based warnings in October 2007 (the so-called county-based warning era), the NWS classified case 2 events as hits, so that the POD would be the sum of case 1 and case 2 warnings, divided by the total number of tornadoes. Classifying case 2 events as hits leads, potentially, to issues in the interpretation of performance since cases 1 and 2 have key differences with respect to when the warning was issued. Although some users would certainly receive a warning on the tornadoes of case 2, imagine the limiting scenario in which warnings are never issued prior to tornadogenesis, but some sort of system has been developed that identifies all tornadoes after they form. The POD, with case 2 defined as a hit, would be 1, even though no advanced warning on a tornado was ever issued prior to the event beginning. For later use, we will adopt the notation that denotes POD_1 as indicating only case 1 events are considered hits and POD_2 as indicating that both cases 1 and 2 are considered hits. During the county-based warning era, the official definition of POD was POD_2.

It is logical to define a lead time in advance (LTA) for the first case as the time between the issuance of the warning and the beginning of the tornado. The lead time for cases 2 and 3 is not so clear. Prior to the adoption of storm-based warnings, in both cases, the official NWS lead time was set to zero. The mean lead time over a number of events, say a year, would be computed as

$$\mathrm{LT}_{\mathrm{mean}} = \frac{1}{N}\sum_{i=1}^{3} N_i \mathrm{LT}_i, \tag{1}$$

where the i subscripts represent the number N_i and mean lead time (LT_i) for each of the three cases, and N is the total number of tornadoes. Now, $\mathrm{LT}_1 = \mathrm{LTA}_{\mathrm{mean}}$ (the mean LTA over all tornadoes warned in advance), and $\mathrm{LT}_i = 0$ if i is 2 or 3, so that (1) reduces to

$$\mathrm{LT}_{\mathrm{mean}} = \frac{N_1}{N}\mathrm{LT}_1. \tag{2}$$

Now, $N_1/N = \mathrm{POD}_1$, so that the official definition of lead time during the county-based warning era was equivalent to the mean lead time of events warned in advance multiplied by the fraction of events warned in advance ($\mathrm{POD}_1 \times \mathrm{LTA}_{\mathrm{mean}}$).

This definition can, again, lead to ambiguity. If we are interested in the question of how long before a tornado begins that a warning is issued (LTA for each tornado), the result from (2) says that $\mathrm{LT}_{\mathrm{mean}}$ is the product of POD_1 and $\mathrm{LTA}_{\mathrm{mean}}$. Thus, without looking at the components of the calculation, we could not tell if differences in $\mathrm{LT}_{\mathrm{mean}}$ between two populations result from tornadoes being warned longer in advance ($\mathrm{LTA}_{\mathrm{mean}}$) or more tornadoes being warned in advance (POD_1).

The situation changed as storm-based warnings began in October 2007. Each tornado was considered to be composed of a series of time segments with each segment being equal to 1 min, so that a tornado that began and ended at, say, 2210 UTC, would have one segment. A tornado beginning at 2210 UTC and ending at 2211 UTC would have two segments, and a tornado beginning at 2210 UTC and ending at 2310 UTC would have 61 segments. For each tornado, the percentage of the event warned (PEW) would be computed as the fraction of the total segments warned. If a warning began after the tornado or ended before the tornado ended, the PEW would be somewhere between 0 and 1 for that tornado. Similarly, the mean lead time averaged over each segment of the tornado was calculated with, in a similar manner to the county-based warning era, any unwarned segment being assigned a lead time of 0. The mean lead time for an individual tornado must be at least the initial lead time for a tornado. The mean performance over a number of events was calculated as the mean of the individual tornado's PEW and lead time. That is, the storm-based era POD (POD_S), is

$$\mathrm{POD}_S = \frac{1}{N}\sum_{i=1}^{N} \mathrm{PEW}_i, \tag{3}$$

where PEW_i is the percentage of the event warned for the ith tornado of the set of events. The mean lead time could be calculated in a similar way, again, with unwarned events being assigned a lead time of 0.

Although we will show the impacts on the performance metrics of the choices of definition (and other possible definitions) in the storm-based warning era, for most of the remainder of the paper, we will ignore the segmented data and focus on the initial touchdown time for tornadoes in evaluating warnings. This will allow us to use the same definitions in both the county-based and storm-based eras and, we believe, allow for insight into the changes that occurred at the change in warning definition and at other times in the record.

3. Results

a. *Long-term annual trends*

We begin consideration of the results by looking at the changes over time in tornadoes that were classified as case 1, 2, or 3 without regard to intensity (Table 1 and Fig. 1).[1] From 1986 to the early 2000s, annual POD_1 increased from ~0.25 to nearly 0.70. From 2002 to 2011 or 2012, POD_1 stayed relatively flat before dropping to ~0.50 from 2013 to 2016. Not surprisingly, the case 3 numbers move in inverse with POD_1. The remainder, case 2, provides some interesting behavior. It slowly increases from 1986 to 2011. A least squares regression yields an increase from 0.071 in 1986 to 0.088 in 2011, with a p value of the slope that is less than 0.01. The value of 0.106 in 2010 was the only year above 0.1 during that 26-yr period. Abruptly, in 2012, the case 2 value increased to 0.116 and, in 2016, it was 0.139. Over the period 2012–16, 12.4% of all tornadoes were case 2, almost 40% above what would have been expected from the regression line based on the 1986–2011 results. Crudely, we can qualitatively describe the long-term behavior of warnings prior to tornado as an increase in POD_1 over the first half of the record from ¼ to ⅔ over a 15-yr period, followed by a period of little change for a decade, followed by a drop in final few years to ½. Most of that change is associated with corresponding changes in tornadoes never warned on, but a substantial increase in the fraction of tornadoes

[1] Even though warnings changed from county based to storm based in October 2007, we treat them the same. For 2007, the last few months were relatively quiet in terms of tornado warnings, so the storm-based warning numbers are small. In addition, as will be seen in the results, there is little or no discernable change in the probability of detection or lead-time values associated with the change in warning format.

TABLE 1. Nationally averaged fractions of tornadoes by warning case: case 1, warned before tornado (POD_1); case 2, warned during tornado and the official POD definition during the county-warning era, the sum of cases 1and case 2 (POD_2); and case 3, never warned.

Year	Case 1 (before) POD_1	Case 2 (during)	POD_2	Case 3 (never)
1986	0.262	0.077	0.339	0.661
1987	0.209	0.077	0.287	0.713
1988	0.224	0.065	0.289	0.711
1989	0.271	0.086	0.357	0.643
1990	0.365	0.075	0.439	0.561
1991	0.331	0.080	0.411	0.589
1992	0.382	0.070	0.452	0.548
1993	0.346	0.083	0.429	0.571
1994	0.401	0.058	0.459	0.541
1995	0.532	0.069	0.600	0.400
1996	0.508	0.084	0.592	0.408
1997	0.512	0.077	0.590	0.410
1998	0.582	0.071	0.653	0.347
1999	0.633	0.069	0.702	0.298
2000	0.558	0.089	0.647	0.353
2001	0.615	0.085	0.700	0.300
2002	0.672	0.086	0.758	0.242
2003	0.716	0.075	0.791	0.209
2004	0.685	0.074	0.759	0.241
2005	0.666	0.095	0.761	0.239
2006	0.663	0.078	0.741	0.259
2007	0.680	0.087	0.767	0.233
2008	0.672	0.082	0.754	0.246
2009	0.614	0.088	0.702	0.298
2010	0.646	0.106	0.752	0.248
2011	0.701	0.085	0.786	0.214
2012	0.625	0.116	0.740	0.260
2013	0.528	0.119	0.647	0.353
2014	0.469	0.117	0.587	0.413
2015	0.525	0.126	0.651	0.349
2016	0.502	0.139	0.642	0.358

warned while they were in progress appeared abruptly in 2012. We will return to this result later.

Breaking tornadoes down by damage rating,[2] we see a pattern in POD_1 that is similar to the overall tornado record with more noise (Fig. 2). The more damaging tornadoes tend to have higher POD_1 values throughout, although the smaller sample sizes makes them extremely noisy at times. [We do not plot values for (E)F4 or (E)F5 storms for this reason.] The POD_1 for (E)F3 tornadoes was approximately 0.9 for many years after 2000, compared to (E)F0/(E)F1 at only 0.6–0.7. The change in POD_1 after 2011 appears most strongly in the lower

[2] Ratings for tornadoes changed from the Fujita scale to the enhanced Fujita scale in February 2007. Again, as with the change from county-based to storm-based warnings, differences across this boundary for warning statistics are not clearly apparent, so we will treat them equivalently.

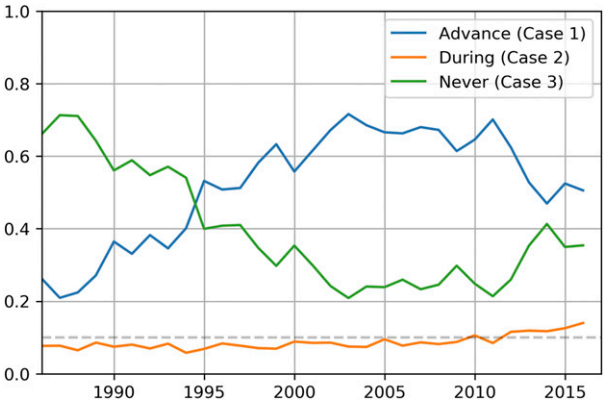

FIG. 1. Table 1 in graphical form. Relative proportion of tornadoes with warnings issued in advance (case 1, blue), during the tornado (case 2, orange), or never (case 3, green).

damage rating tornadoes: EF0 and EF1. Because there are so many more of these cases than the stronger tornadoes, this change tends to dominate the overall change seen in Fig. 1.

From 1986 to 2011, the annual mean lead time for tornadoes warned in advance (LTA_{mean}) shows no statistically significant trend (Fig. 3). The mean over those 26 years is 18.8 min. From 2012 to 2016, it is only 15.6 min. The four smallest annual LTA_{mean} values in the entire record are in 2013–16 and only 2002 (16.4 min) is shorter than 2012 (16.8 min). As with the POD_1 and case 2 warnings, large changes took place in metrics starting in 2012. The NWS official definition of lead time (LTO) during the county-based era, which is $POD_1 \times LTA_{mean}$, behaves essentially like POD_1 prior to 2012, since LTA_{mean} is nearly constant over that period. During 2012–16, the gap between LTA_{mean} and the official definition increases because of the combination of the lower POD_1 compared to earlier times and the lower LTA_{mean}. The long-term increase seen in LTO over the first half of the record is entirely an increase in the fraction of tornadoes warned in advance (POD_1) and not a result of warning longer in advance for case 1 tornadoes. This distinction is clearly important in understanding POD_1 changes in NWS practice over the years.

Unlike what was seen for POD_1, when we look at LTA_{mean} by different (E)F scales, we see no significant differences in performance (Fig. 4). For the 1986–2011 period, LTA_{mean} results by (E)F scale from 0 to 3 are 18.8, 18.7, 19.3, and 18.8 min, respectively. In effect, there's no distinguishable difference in LTA_{mean} from 1986 to 2011 for any year or any (E)F-scale value. The POD_1 changes dramatically, but LTA_{mean} is relatively constant. What differs is not how long before a tornado a warning is issued, but whether a warning is issued in

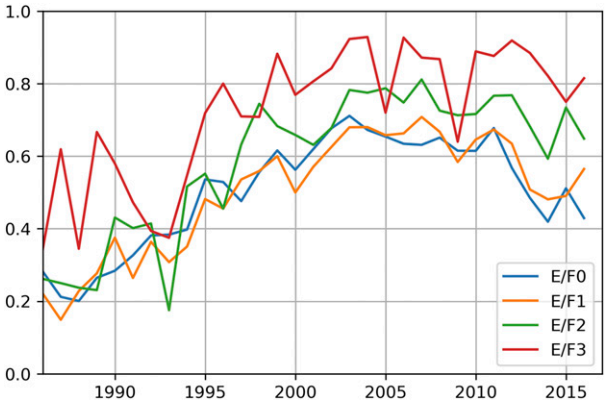

FIG. 2. Values of POD_1 by (E)F-scale. Because of small sample sizes, (E)F4 and (E)F5 are not shown.

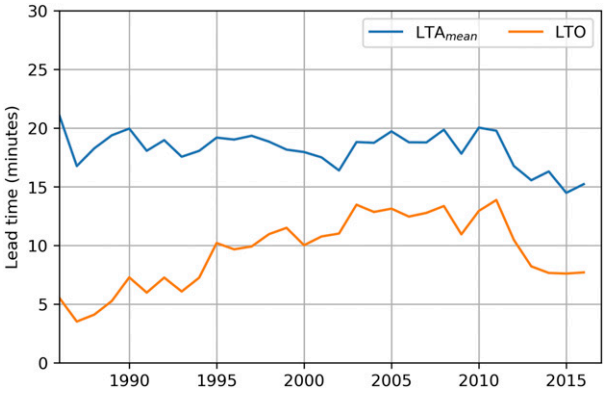

FIG. 3. Results for LTA_{mean} and LTO.

advance at all. There is a slight indication of longer lead times for stronger tornadoes compared to weaker tornadoes during the 2012–16 period with the equivalent values of 15.4, 15.4, 17.0, and 17.1 min, although the differences between EF scales in the recent period are smaller than the decrease from the early period to the later period. As would be expected, the official definition of lead time by (E)F scale shows changes that are dominated by the differences in the POD_1 for each damage rating level (Fig. 5).

For the sake of completeness, we show the annual values of possible definitions for POD (Fig. 6). We have already defined the POD_1, POD_2, and POD_S. The value of POD_S can only be calculated during the storm-based warning era. A fourth alternative for the storm-based era is to calculate the total number of segments warned divided by the total number of segments during the year (SEG). Qualitatively, SEG produces values that are relatively close to POD_2. Note that SEG $>$ POD_S by about 0.05 because it weights each segment equally, rather than giving more weight to segments from short-lived tornadoes. We find that POD_S implicitly values performance on short-lived tornadoes more strongly than long-lived tornadoes because missing a one-segment tornado gives a PEW of 0 for that tornado, but missing 10 segments out of a 20 segment tornado gives a PEW of 0.5 for that tornado. Because all tornadoes are equally weighted in POD_S, the 10 warned (or missed) segments in the 20-segment tornado contribute less to POD_S than the single-warned (or missed) one-segment tornado. The gap between POD_1 and POD_2 is the ''warned during'' line from Fig. 1.

b. *The 2012 discontinuity*

Values of POD_1 and LTA_{mean} show declines after 2011 that are large compared to the long-term trend. In addition, the fraction of case 2 tornadoes jumps in 2012. All three of those quantities remain at values not seen during the previous decade (POD_1) or entire record for the other two (LTA_{mean} and case 2 tornadoes) through 2016. We will examine the differences in performance during the period just before 2012 to the following period. To facilitate discussion, we will look only at performance during the storm-based warning era and compare warnings from October 2007 through the end of 2011 to those from 2012 to 2016. Even though little difference in metrics is seen in the transition from county-based to storm-based warnings, this eliminates some possible points of confusion.

First, let us consider the characteristics of tornadoes during the two periods. Although the early period had two historically large years (2008 and 2011) and the later period was characterized by years with fewer tornadoes, so that there are more tornadoes in the first period than the second, the relative distribution by intensity is similar, except at EF5, which represents a very small number (Fig. 7). Similarly, the distribution of tornadoes by time on the ground is the same during both periods (Fig. 8). It is clear that even though there were more tornadoes in the early part of the storm-based warning era, there are no large differences in the distribution of tornadoes by damage rating or lifetime. The effects of the differences on warning performance are small. Assuming POD_1 remained constant by EF scale between the two periods, the changes in overall POD_1 explained by the change in damage distribution is less than 10% of the total change.

Since the distribution of characteristics of the tornadoes was similar during the two periods, we will consider changes in some characteristics of the warnings. To put these characteristics into context, we start by considering the mean duration of the warnings. For the ending time of a warning, we use the ending time initially specified in the warning and disregard any cancellation messages that might have ended the warning earlier than

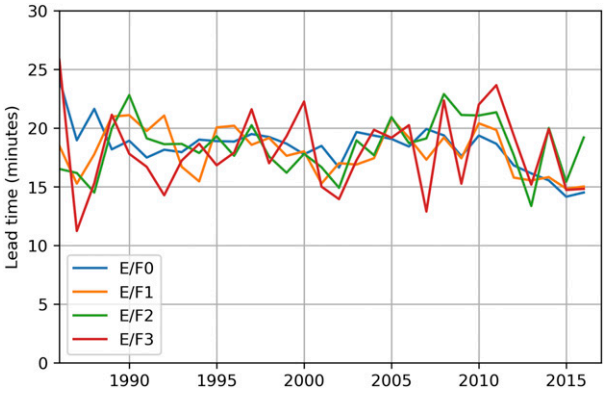

FIG. 4. As in Fig. 2, but for LTA_{mean}.

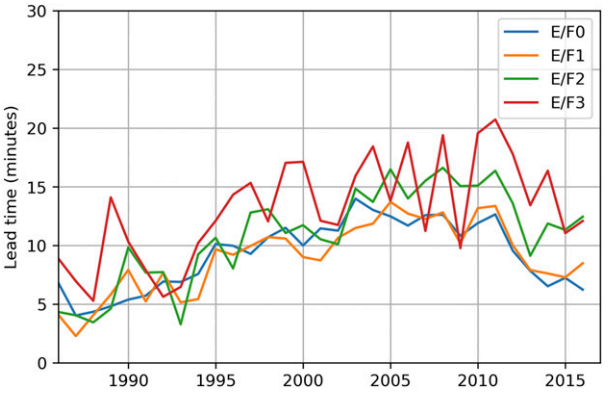

FIG. 5. LTO by F scale.

stated when the warning was issued. Although there was a substantial trend toward shorter warnings from 1986 to 2011 (-0.17 min yr^{-1}), there was a large drop during 2012–16 (Fig. 9). Warnings were approximately 4 min shorter in duration than would have been expected based on the 1986–2011 trend. Adding 4 min to the 2012–16 values would move the points onto the long-term trend, yielding the same regression line to within 0.02 min at all points along the line. The 4-min value is equivalent to 23.3 years along the regression line, so the change between 2011 and 2012 could be interpreted as equivalent to 20–25 years of long-term change. The distribution of warning duration further emphasizes the difference in warning characteristics between the two periods (Fig. 10). Warning durations tend to be clustered near 30, 45, and 60 min. The number of warnings in the clusters near 45 and 60 min drops dramatically starting in 2012 with the warnings clustered near 30 min in duration increasing.

The changes in POD_1 and LTA_{mean} have been shown earlier, but it is instructive to break the two periods of the storm-based warning era down by EF scale. During both periods, POD_1 tends to increase with EF scale (Fig. 11), as seen before. However, POD_1 drops below 0.60 for EF0 tornadoes. Since the majority of tornadoes are weak, this drop in POD_1 dominates the overall POD_1 trend seen in Fig. 6. On the other hand, LTA_{mean} shows approximately the same decrease across all EF-scale values (Fig. 12). The clear message is that, whatever happened in 2012, the lead time change was independent of EF scale, but the probability a warning would be issued prior to the tornado was a strong function of the EF scale.

c. *Multiple performance metrics*

The interrelationship between errors of various kinds (missed detections, false alarms) can be described using graphics that visualize multiple performance metrics in a single diagram. Here, we will consider two: the relative (or receiver) operating characteristics (ROC) diagram (Mason 1982) and the performance diagram (Roebber 2009). To provide some background, we review the discussion of Brooks (2004) on signal detection theory related to tornado warnings. We posit the existence of some variable that describes the weight of evidence needed for a forecaster to make a decision to issue a warning or not. We assume there are distributions of this variable associated with nontornadoes and with tornadoes such that, for simplicity, the value for a tornado tends to be higher than for nontornadoes. A forecaster who has to make a binary yes/no decision on warning will have some threshold of the variable above which they will issue a warning and below which they will not. Since there is uncertainty, the existence of this threshold leads to misclassifications, with some tornadoes having low values of the variable and some nontornadoes having high values. By increasing the threshold (requiring more evidence to warn), the forecaster will lower the number of false alarms but increase the number of missed detections. The selection of the appropriate threshold could be based upon the costs associated with the various errors and then some attempt to minimize; although, at this time, we know of no work to estimate those costs. The decision threshold could have significant impacts on the *value* of weather forecasts, the costs or benefits associated with decisions made using the forecasts, even when the *quality* of the forecasts, the relationship between the forecasts and observations, does not change (Murphy 1993). Note that the costs may be expressed in monetary units (loss of business during a false alarm) or in some other utility (loss of time, loss of trust in the warning system, increased injuries or deaths; Ripberger et al. (2015).

The ROC diagram consists of plots of the POD versus the probability of false detection (POFD), the fraction of nonevents that are warned on. We will use POD_1 as

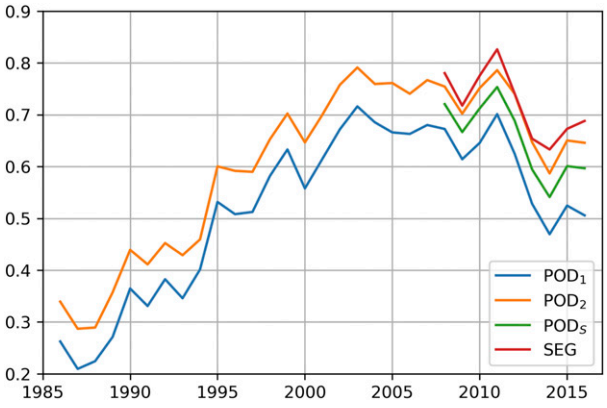

FIG. 6. Time series of different definitions of POD. POD_1 (blue), POD_2 (orange), and POD_S (green), as well as SEG (red).

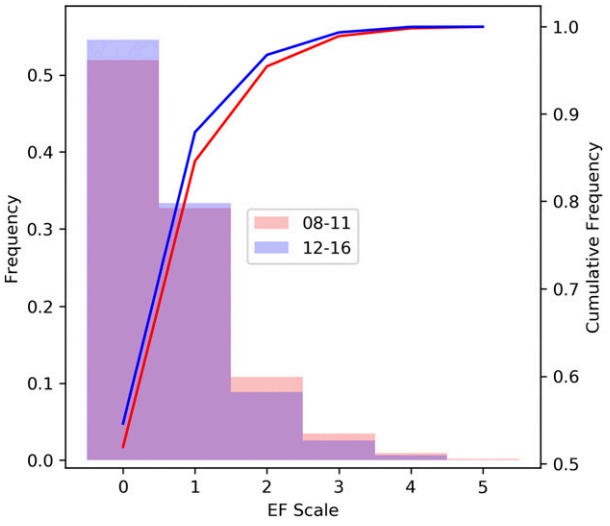

FIG. 7. Relative proportion of tornadoes by EF scale for 2007–11 (red) and 2012–16 (blue). Solid lines represent the cumulative distribution.

the value for the former. As with Brooks (2004), a challenge exists in defining the correct forecasts of nonevents that are considered in the decision model. We choose to follow Brooks (2004), in setting that value such that the fraction of tornadoes to the total number of events being considered in warning is 0.1. This value is arbitrary and affects the results quantitatively, but not qualitatively. Based on fitting curves to periods of several years, it is unlikely the value is as low as 0.05 or as high as 0.2 (Brooks 2004). A simple model underlying the ROC is that the distributions of the variable in question associated with tornadoes and nontornadoes can be fit to a normal distribution. For ease, we assume that the standard deviations associated with the two normal distributions are the same and that the discrimination between the two distributions can be described by their separation as the number of standard deviations the means of the two distributions are apart (Mason 1982; Brooks 2004). This separation is usually designated as D', with larger values indicating the distributions are farther apart and implying easier discrimination. Brooks (2004) associated it with the quality of the science and technology in the warning process. Another descriptor of the quality of the discrimination is the area under the curve (AUC) described by the ROC (Mason 1982). The AUC is equal to the Mann–Whitney U-test statistic, which is the probability that a value of the variable describing the weight of evidence to warn drawn randomly from one population (say the tornado population) is greater than a randomly drawn value from the other population (nontornado) (Mason and Graham 2002).

We update the time series from Brooks (2004) (Fig. 13). In the early years of the record, performance moves toward larger values of D', indicating better discrimination starting in the early 1990s, most likely associated with the implementation of the WSR-88D radar system. In the late 1990s, annual points move up and to the right along a constant D' line, associated with a lower threshold for warning. As discussed in Brooks (2004), because of the larger number of nontornadic events, this threshold led to large increases in the POD_1 with small changes in the FAR. In the early 2000s, there was another move toward higher D' values. There is little evidence for a change associated with the beginning of the storm-based warning program, but a drop in POD_1 takes place in 2012 (see Fig. 1) with later points clustered along the same D' line, but closer to the bottom left, consistent with an increase in the decision threshold. As with the earlier moves toward the top right, the changes in POD_1 are greater than the changes in POFD and associated FAR. This suggests no change in skill of the forecasts between the early and late periods of the storm-based warning eras, but a change in the threshold for warning. The change in the decision threshold at 2012 is consistent with responding to calls for reductions in false alarms for tornado warnings in the aftermath of the tornadoes of spring 2011, particularly the Joplin, Missouri, tornado (NWS 2011, 2013; Kuligowski et al. 2013). The changes are not consistent with differences in the difficulty of the warning situations or the nature of the storms. Brotzge et al. (2011) showed fewer missed detections and fewer false alarms on tornado outbreak days. In that case, we would expect to see lower POD_1 and higher POFD and FAR in years with fewer outbreaks, such as during the late storm-based period. Instead, we see lower POD_1, POFD, and FAR in recent years, consistent with a change in threshold rather than a large change in quality.

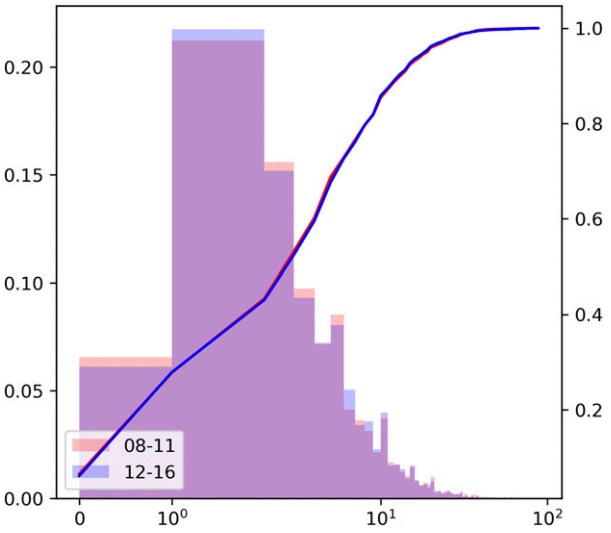

FIG. 8. As in Fig. 7, but for tornado duration (min).

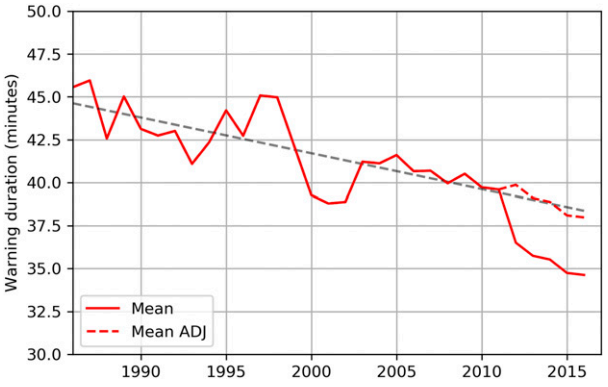

FIG. 9. Annual mean warning duration. The black dashed line is a linear regression fit to 1986–2011. The red dashed line is 2012–16 with difference of means between 2008–11 and 2012–16 added.

A similar story is illustrated by the use of a performance diagram, where POD_1 is plotted against the success ratio (SR), which is 1 − FAR (Fig. 14). The additional insight provided by the performance diagram, compared to plotting the component time series separately, is the inclusion of lines of constant critical success index (CSI). CSI is the number of correct forecasts of events divided by the union of the forecasts and events. Graphically, it can be thought of as the intersection of the forecasts and events on a Venn diagram. Constant CSI lines are qualitatively similar to the constant D' curves on the ROC, in that they reflect constant quality, with movement along the curve associated with a change in the decision threshold. Again, large increases in POD_1 with small changes in SR take place from 1986 to 2011. After that, the cluster of the most recent years moves along constant CSI values toward lower POD_1, higher SR, and lower bias compared to the early storm-warning era cluster. This suggests no change in overall quality but an increase in the threshold for issuing a warning.

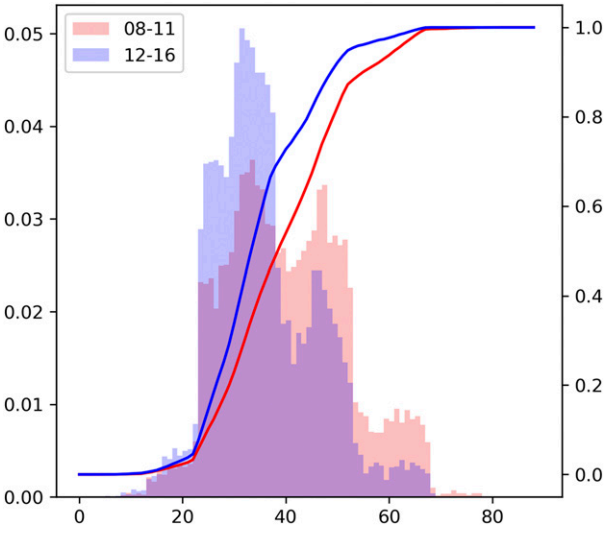

FIG. 10. As in Fig. 7, but for tornado warning duration.

Using the model formulation from Brooks (2004), we can estimate the magnitude of the derivative of the POD with respect to FAR as a function of FAR for different values of D', assuming, as before, that the fraction of events that are thought about in the warning process that are tornadic is 0.1 (Fig. 15). Except at low and very high values of FAR, the changes in POD are greater than a change in FAR. In particular, for values of FAR near 0.75 and D' between 1 and 1.35 (approximately the state of the system in 2008–11), for a decrease in FAR, we would expect the POD to change twice as much. This is consistent with the change observed between the early and late periods of the storm-based warning era.

4. Summary and discussion

We have considered multiple, interrelated aspects of tornado warning performance over time, looking at both individual measures as well as those that are interdependent. The methods used here can be applied to any dichotomous forecasts of dichotomous events. Those forecasts could also include probabilistic forecasts that have thresholds applied to them to help identify the optimal threshold for users. In addition, by monitoring the multiple aspects, it may be possible to determine when changes (planned or inadvertent) in the forecasting system have taken place. It also should

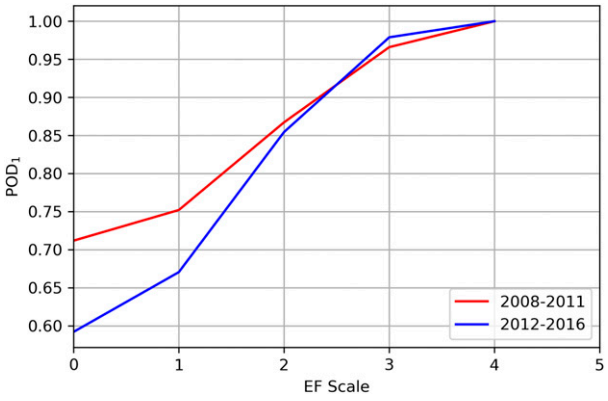

FIG. 11. Values of POD_1 by EF scale for the early and late eras of storm-based warnings.

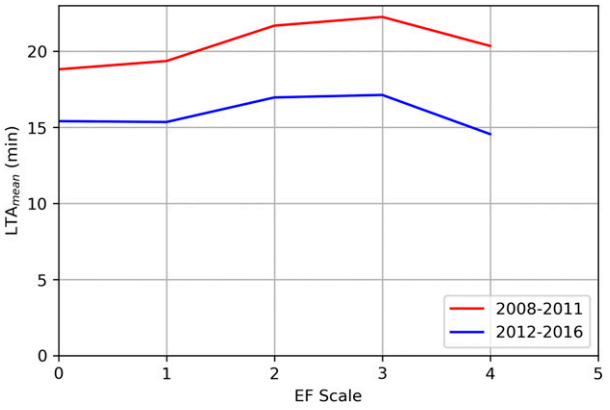

FIG. 12. As in Fig. 11, but for LTA_{mean}.

be possible to identify changes in the quality separately from changes in the decision threshold, which could affect the value of the forecasts (Murphy 1993). In the future, we hope to be able to apply variability in warning metrics as a function of environmental conditions (Anderson-Frey et al. 2016), which would provide baselines for expected performance to provide evidence for decisions in the design of possible changes in the warning system.

In particular, we have examined 31 years of NWS tornado warning performance metrics in order to clarify how warning characteristics, operational procedures, and resultant warning performance have changed over the period. By focusing on whether and how long warnings were issued prior to tornado occurrence and not including warnings issued after tornadoes have occurred, we believe we have a clearer picture of what has changed and what has been consistent over the years. In particular, this approach allows us to consider county-based and storm-based warnings with the same methodology, which shows no significant changes occurring in performance when storm-based warnings were introduced. From 1986 to the mid-2000s, mean lead time and FAR were consistent, but POD_1 increased dramatically.

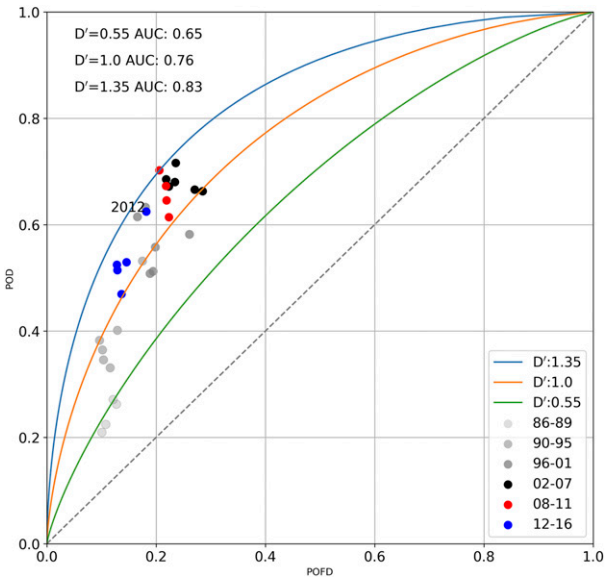

FIG. 13. ROC for tornado warnings by year. The county-based era is in grayscale, with the most recent period (2002–07) in black, early storm-based era in red, and late era in blue. The year 2012 is highlighted. The dashed line represents no skill, and the solid lines represent various theoretical values of constant skill indicated by D'. Perfect forecasts would be in the top-left-hand corner.

The largest changes in performance took place in 2012 when POD_1, FAR, and lead time all decreased. These decreases are associated with two factors. First, and most important, there was an apparent increased emphasis on reducing FAR that led to a change in the threshold for issuing warnings, such that fewer warnings were issued and lead time was reduced. Second, there was a reduction in duration of tornado warnings with a larger fraction of warnings being issued that were approximately 30 min in duration, rather than 45 min. The change in default warning length (from 45 to 30 min) led to an overall decrease in warning duration equivalent to more than 20 years of the previous trend in warning duration and had impacts on the performance metrics, particularly in reducing POD_1 and LTA. The warning decision threshold change appears to be associated with approximately constant overall skill, as may be seen in ROC and performance diagrams. For comparison, a performance diagram for severe thunderstorm warnings, which saw none of the increased emphasis on false alarms, compared to the emphasis for tornado warnings and, thus, no implied changes in their perceived cost, shows little to no change in performance in 2012

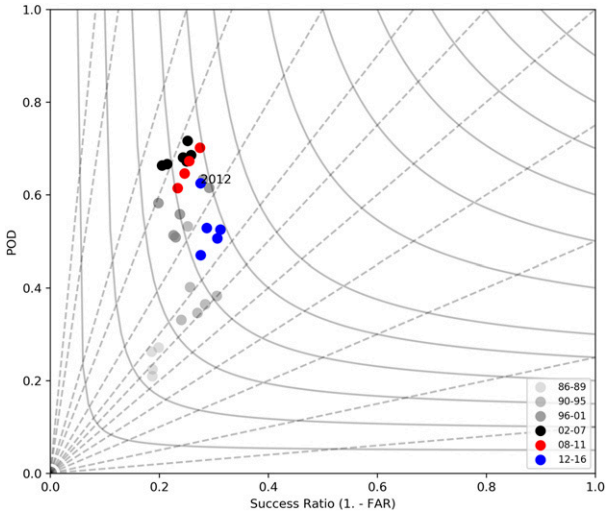

FIG. 14. As in Fig. 13, but for a performance diagram. Solid lines are constant CSI, increasing to the top right, and dashed lines are constant bias, increasing to the top left.

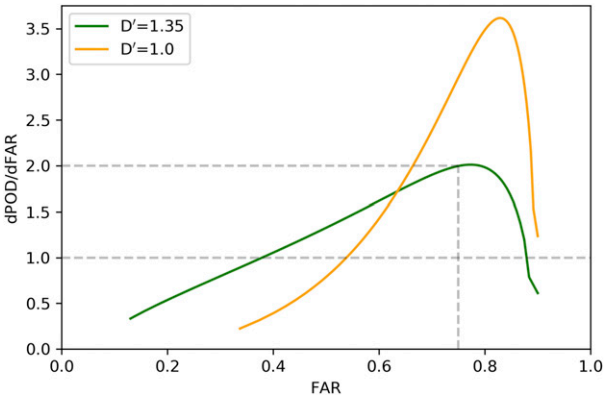

FIG. 15. Change of POD with FAR as a function of FAR for $D' = 1$ (orange) and $D' = 1.35$ (green). Horizontal dashed lines show values of 1 and 2. The vertical dashed line shows FAR = 0.75.

(Fig. 16). The trade-off between missed detections and false alarms for tornado warnings, with larger changes in the FAR than POD_1, is consistent with expectations from the signal detection model from Brooks (2004). The impacts of reduced LTA on recipients of the warnings are unclear. Simmons and Sutter (2008) suggest a reduction in fatalities associated with lead times up to 15 min, with little or no reduction for longer lead times. Hoekstra et al. (2011) collected survey information on preferred lead time by the public (a mean of little more than half an hour), but an unanswered question is how that preference is tied to current warning performance. Perceptions of performance are important in assessing the impacts of warnings, but are complex and not necessarily related to official definitions of warning performance (Ripberger et al. 2015).

The implication of our results is that care should be taken in constructing metrics to evaluate the warning performance. Defining tornadoes that are not warned on in advance as having a lead time of zero can lead to confusion about whether performance changes have resulted from how often warnings occur in advance or how long in advance those warnings are issued. In addition, the relationship between missed detections and false alarms is such that any performance goals associated with POD_1 and FAR should be constructed with both taken into consideration. There are periods in the record where the two of them change in predictable ways associated with similar overall skill and a change in the decision threshold to issue a warning. Metrics such as D' or, to a lesser extent, CSI, could be used to see how skill changes, while developing methods to estimate the costs of errors to develop criteria as to where the threshold for issuing warnings should be set.

In closing, by examining the bulk structure of all of the warnings, we have gained insight into how the tornado warning system works and how it has changed over time. The large performance changes that took place in 2012, in contrast to the lack of performance changes when storm-based warnings were introduced in 2007, stand out. Importantly, the trade-off between false alarms and missed detections associated with an apparent change in decision threshold for issuing warnings can be seen. The decision threshold change is consistent with responding to calls for reduced false alarms in the warning process. It is, perhaps, remarkable how quickly the widespread performance change occurred, indicating the capability of a rapid response within NWS to perceived changes in user needs and desires.

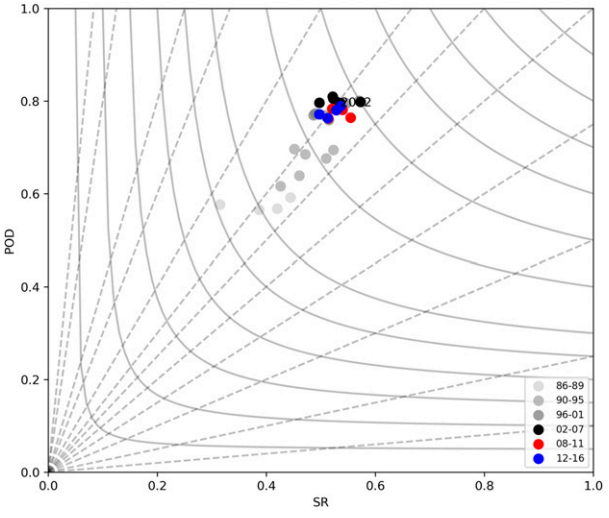

FIG. 16. As in Fig. 14, but for severe thunderstorm warnings.

Acknowledgments. We were inspired by Hamming's motto: "The purpose of computing is insight, not numbers" (Hamming 1962). The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect the views of NOAA or the Department of Commerce. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA16OAR4320115, U.S. Department of Commerce. Somer Erickson provided insight into the county-warning era data. We thank Jerald Brotzge, Pieter Groenemeijer, and Daryl Herzmann for their constructive comments in the review process. Conversations with Stephen Mullens, Chris Karstens, Dale Morris, and Makenzie Krocak helped us to formulate our ideas. Earlier versions of the manuscript were read by Pam Heinselman, Alan Gerard, Michael Coniglio, and Steven Koch of NOAA/NSSL; Steven Weiss and Russell Schneider of the National Weather Service Storm Prediction Center; Greg Schoor, John Murphy, Mike Hudson, Richard Wagenmaker, Greg Mann, and Mary Erickson of the National Weather Service; and Randy Peppler of the University of Oklahoma Cooperative Institute for Mesoscale Meteorological Studies. We cannot guarantee that anyone we discussed the paper with or who reviewed it informally agrees or disagrees with any of the conclusions.

REFERENCES

Anderson-Frey, A., Y. Richardson, A. Dean, R. Thompson, and B. Smith, 2016: Investigation of near-storm environments for tornado events and warnings. *Wea. Forecasting*, **31**, 1771–1790, https://doi.org/10.1175/WAF-D-16-0046.1.

Brooks, H. E., 2004: Tornado warning performance in the past and future: A perspective from signal detection theory. *Bull. Amer. Meteor. Soc.*, **85**, 837–843, https://doi.org/10.1175/BAMS-85-6-837.

Brotzge, J., and S. Erickson, 2009: NWS tornado warnings with zero or negative lead times. *Wea. Forecasting*, **24**, 140–154, https://doi.org/10.1175/2008WAF2007076.1.

——, and ——, 2010: Tornadoes without NWS warning. *Wea. Forecasting*, **25**, 159–172, https://doi.org/10.1175/2009WAF2222270.1.

——, ——, and H. Brooks, 2011: A 5-yr climatology of tornado false alarms. *Wea. Forecasting*, **26**, 534–544, https://doi.org/10.1175/WAF-D-10-05004.1.

Hamming, R. W., 1962: *Numerical Methods for Scientists and Engineers*. McGraw-Hill. 752 pp.

Hoekstra, S., K. Klockow, R. Riley, J. Brotzge, H. Brooks, and S. Erickson, 2011: A preliminary look at the social perspective of Warn-on-Forecast: Preferred tornado warning lead time and the general public's perceptions of weather risks. *Wea. Climate Soc.*, **3**, 128–140, https://doi.org/10.1175/2011WCAS1076.1.

Kuligowski, E. D., F. T. Lombardo, L. T. Phan, M. L. Levitan, and D. P. Jorgensen, 2013: Technical investigation of the May 22, 2011 tornado in Joplin, Missouri. National Construction Safety Team Act Rep. 3, 428 pp., https://nvlpubs.nist.gov/nistpubs/NCSTAR/NIST.NCSTAR.3.pdf.

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

Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.*, **128**, 2145–2166, https://doi.org/10.1256/003590002320603584.

McIntosh, J., 1986: *The Practical Archaeologist: How We Know What We Know about the Past*. Facts on File, 192 pp.

Murphy, A. H., 1993: What is a "good" forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

NWS, 2011: The historic tornadoes of April 2011. NOAA National Disaster Survey Rep., 76 pp., http://www.weather.gov/media/publications/assessments/historic_tornadoes.pdf.

——, 2013: National Weather Service Weather-Ready Nation Roadmap. WRN Roadmap, version 2.0, 75 pp., http://www.weather.gov/media/wrn/nws_wrn_roadmap_final_april17.pdf.

Ralph, F. M., and Coauthors, 2013: The emergence of weather-related testbeds linking research and forecasting operations. *Bull. Amer. Meteor. Soc.*, **94**, 1187–1211, https://doi.org/10.1175/BAMS-D-12-00080.1.

Ripberger, J. T., C. L. Silva, H. C. Jenkins-Smith, D. E. Carlson, M. James, and K. G. Herron, 2015: False alarms and missed events: The impact and origins of perceived inaccuracy in tornado warning systems. *Risk Anal.*, **35**, 44–56, https://doi.org/10.1111/risa.12262.

Roebber, P., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Simmons, K. M., and D. Sutter, 2005: WSR-88D radar, tornado warnings, and tornado casualties. *Wea. Forecasting*, **20**, 301–310, https://doi.org/10.1175/WAF857.1.

——, and ——, 2008: Tornado warnings, lead times, and tornado casualties: An empirical investigation. *Wea. Forecasting*, **23**, 246–258, https://doi.org/10.1175/2007WAF2006027.1.

Sutter, D., and S. Erickson, 2010: The time cost of tornado warnings and the savings with storm-based warnings. *Wea. Climate Soc.*, **2**, 103–112, https://doi.org/10.1175/2009WCAS1011.1.