13B.4 REEVALUATING PERFORMANCE IN THE STATISTICAL METRICS FOR NATIONAL WEATHER SERVICE TORNADO WARNINGS

Greg M. Schoor[1] *, J.G. Gibbs[2], J.P. Camp[3]

[1]NOAA/National Weather Service
Analyze, Forecast, and Support Office

[2]NOAA/National Weather Service
OCLO/Warning Decision Training Division

[3]NOAA/National Weather Service
Tallahassee, FL

1. INTRODUCTION

For decades, Tornado Warnings have been a fundamental part of the service-based mission of the National Weather Service (NWS). Over this period, the NWS warning program has evolved, from the tools that remotely scan thunderstorms, to the tools meteorologists use to issue the warnings. However, the evaluation and validation methodologies of these warnings have not received commensurate attention in terms of aligning with service improvements. Over the past couple of decades especially, there has been disproportionate amount of progress made between the increase in technological enhancements aimed at improving the warning service and the metrics that judge warning performance.

The Government Performance and Results Act (GPRA) of 1993 set the mandates for the NWS to track certain performance metrics for various watches and warnings. For NWS Tornado Warnings, these metrics are the Probability of Detection (POD), False Alarm Ratio (FAR), and Average Lead Time (LT). A fourth metric, Critical Success Index (CSI, Schaefer 1993) is also tracked but not required under the Act. POD, FAR, and CSI are largely based on the 2x2 forecast matrix of forecast goodness or accuracy (Murphy 1993), an analog methodology that compares the warning versus an event, of a tornado in this instance. The notion of accuracy based in the POD and FAR metrics are based upon the straightforward comparison between the occurrence of a forecast and a qualifying event. The severity of an event is not considered with any of these metrics, regardless of impact.

Since the adoption and implementation of the Storm-Based Warning (SBW) paradigm for NWS Tornado Warnings in late 2007 (Ferree et al. 2007), some modifications to this methodology were needed to better align the polygon threat areas to the occurrence of a tornadic event. However, this has been the only substantive change or modification to these performance metrics since the GPRA metrics were enacted and adopted by the NWS. The study described here introduces the concept of assimilating certain characteristics of an event or hazard, such as the intensity, into performance metrics for warning service.

* *Corresponding author address:* Greg Schoor, NOAA/NWS/AFSO, 120 David L. Boren Blvd. Ste. 2312, Norman, OK 73072 email: gregory.m.schoor@noaa.gov

2. STORM DAMAGE SURVEYS THE EF-SCALE AND THE DAMAGE ASSESSMENT TOOLKIT

After the occurrence of a tornado with the area of damage identified, the NWS assesses the damage and records the activity, becoming the official record for that event. As new technologies and capabilities have become more mainstream over time, so have certain techniques involving damage surveying and assessment. Prior to 2007, the NWS utilized the F-Scale (Fujita and Pearson 1973) to rate tornadoes by increasing classes, from F0 to F5. The F class of a particular tornado was classified with the highest degree of surveyed damage to the estimation of wind speed that may have caused that level of damage. An enhanced version of this scale was adopted by the NWS in early 2007, called the EF-scale (NWS 2008). This updated scale provides a more robust set of markers or indicators of damage per the science of wind speed estimation and building factors, such as materials and how well they are fortified.

Over the past several years, the need to produce survey results in Geographic Information Systems (GIS) formats has increased. To meet this need, the NWS Damage Assessment Toolkit (DAT). The DAT serves as a repository of information from the survey of a tornado track, including path length, width, and the intensity of damage (EF class) for individually surveyed damage points. This tool is used to create plots of tornado damage paths that can be used scientific analysis of the event and for local officials to view the impacts of a tornadic event on an area of interest. Figure 1 is an example of tornado damage points and paths on the online DAT Viewer, zoomed-out to a regional-scale, with numerous paths from different events overlaid on one display.

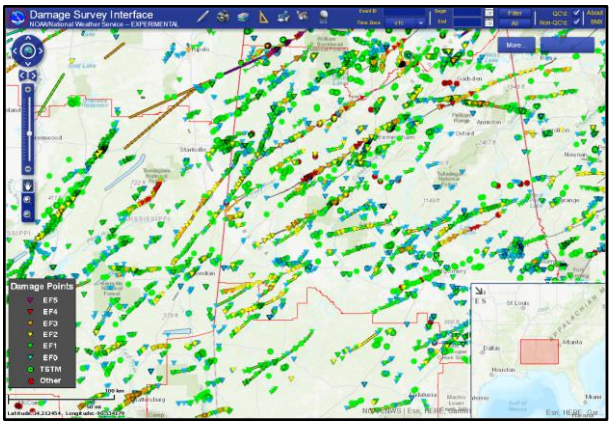

Figure 1. Example of tornado damage plots (points and paths) in the online NWS Damage Assessment Toolkit (DAT). The colored triangles are analyzed damage points and the straight lines are analyzed tornado paths. All damage points and paths correspond to the assigned EF class, shown in the legend on the lower-left side of the image. Image: NWS, 2018a

3. DATA NORMALIZATION METHOD

3.1 *Challenges in Warning Metrics*

The currently utilized methods for the determinations of NWS Tornado Warning performance, through POD, FAR, LT, and CSI have certain utility, namely within the understanding of providing warning service relative to the mere occurrence of an event. However, these metrics do not relay information on the applicable event intensity. The current metrics utilize the 2x2 forecast and event matrix, taking a "yes or no" approach to tornado events, using the following equations:

$$POD = A \div (A + C)$$

where A = no. of hits, C = no. of misses

$$FAR = B \div (A + B)$$

where A = no. of hits, B = no. of forecasts (warnings) without a qualifying event

$$LT_{avg} = (D_1 + D_2 + D_3 \ldots D_n) \div n$$

where D = lead time at one-min. intervals; n = no. one-min. intervals

$$CSI = A \div (A + B + C)$$

A rare but highly impactful EF5 tornado will typically leave visual scars on the landscape that it affects and lasting psychological scars on survivors and others impacted by the storm. These events are usually and immeasurably more than significant in their impact than a common short-lived EF0 tornado that may not produce more than minor tree and/or property damage along a relatively short path. However, each tornado is treated equitably, as one tornado event in the context of the associated warning, using the current methodology. As an example for how this may work in a hypothetical but realistic scenario, a certain NWS office that hits several significant tornadoes with substantial lead time for each event but misses a series of weak tornadoes over that period may attain a cumulative POD around 0.60 (60% accuracy). However, this value does not capture or quantify the elements of how the lead time of the warning relates to the eventual intensity of the tornado.

Measuring actual societal impact, including humanistic considerations, such as fatalities, injuries, or even lives potentially "saved" as a response to a warning is complex. These types of values per tornadic event and the related warning presents notional challenges from the perspective of their usage as a proxy for societal impact or as a measure of public service, with many nuances and questions about the benefits of this relevant to the warning service (Simmons and Sutter, 2011).

The nature of the 2x2 matrix methodology relative to a meteorological occurrence, such as a tornado, provides an operational challenge as well as a scientific one. On average, a vast majority of U.S. tornadoes are rated on the lower end of the intensity scale or spectrum. Analyzing NWS Storm Data for tornado reports, between 2008 and 2017, which is essentially the entirety of the SBW era to the present time, revealed that over 86% of all U.S. tornadoes are of EF0 or EF1 intensity. This value increases to over 96% with EF2 tornadoes included, making less than 4% of all tornadoes EF3, 4, or 5 intensity, with EF5 by far the rarest. Many of the lowest-level intensity tornadoes can be challenging to predict in real-time with substantial lead time and at times challenging to detect on radar unless they are within a relatively close proximity.

Although 2011 was a historically active year for tornadoes nationwide, the number "false alarm" warnings that year outnumbered the warned instances by nearly a factor of three. Since then, the moving average of warnings that are false alarms is over 1,000, annually. The three-year moving average of FAR however, only decreased from 73% (2010-12) to 70% in (2015-17), demonstrating the rigidity of this metric under the current methodology. During this same period from 2011 to 2017, the POD decreased from 0.75 (2011) to 0.57 (2013) and has maintained relatively steady within this range since dating back to 2017. None of these statistics account for the intensities of the tornado events, nor demonstrate any quantification of societal impact.

3.2 *A Different Approach: Feature Scaling Methodology*

Statistical methods, such as feature scaling (Aksoy and Haralick 2001) allow for the normalization of skill scores that do not vary with fluctuations in the data. This particular method standardizes a specified range, such as the range between 0 and 1, and provides the ability to relativize a desired value for individual

occurrences for the range of independent variables. For warnings of specific tornado occurrences, feature scaling, through the sub-method of Max/Min Scaling, can elicit a supposed "value" of the performance of a specific characteristic of forecast (warning), to the qualifying event (tornado). The desired warning characteristic for this effort is lead time, in minutes.

Lead time in advance of a tornado is the most critical service that can be provided from a warning, especially lead time that is substantial enough to allow recipients of the warning to take appropriate sheltering actions for them and anyone else within the immediate physical proximity. Although the NWS lead time goal is set at 13 minutes for all tornadoes, the minimum amount of lead necessary for individuals can vary subtly or substantially, depending on a number of factors at the time a message is received. Consequently, lead time amounts that are abnormally long, such as 45 minutes to 1 hour ahead of the arrival of a tornado can cause confusion for the recipient. These long lead times may lead to delayed sheltering actions, if the warning recipient feels as if the tornado is far enough away that they can wait for an update and do not need to take immediate protective actions.

The basis for the normalization property of feature scaling involves a theoretical "perfect" score/value (Max) and a theoretical worst-possible, or lowest possible score/value (Min) and is then calculated as:

$$Scaled\ Value\ = \frac{Event\ Value\ - Min\ Value}{Max\ Value\ - Min\ Value}$$

For the purposes of NWS Tornado Warning statistical analysis:

$$Warning\ Score\ = \frac{Event\ Score\ - Min\ Score}{Max\ Score\ - Min\ Score}$$

The event score is calculated by dividing-up the path of a tornado event (track) into segments that are delineated by evenly-spaced timing differences. Each segment accounts for a 5-minute interval, starting at the initial touchdown of the tornado, through an analyzed point of damage, then assigning a unique timestamp for each subsequent segment, every 5 minutes. The SBW paradigm brought about the modification to the GPRA metric for LT, adopting the calculation of average LT (LT_{avg}). LT_{avg} is calculated through one-minute segments, as denoted in Section 3.1, correlating the time the warning was issued with the LT at each one-segment along the tornado path, then producing an average value based on that information.

Common surveillance modes for the NEXRAD WSR-88D Radar have full coverage pattern scanning updates on the order of 4 to 5 min (NWS WDTD 2018). Since the network of NEXRAD radars were deployed, common forecaster decisions on issuing warnings for predicted tornado occurrences are correlated with the update cycle times for these radar surveillance modes. Dividing up the tornado path into 5-minute segments follows this generalized paradigm of decision-making and relates the segment properties to the issuance time of the warning and therefore, the provision of lead time variances.

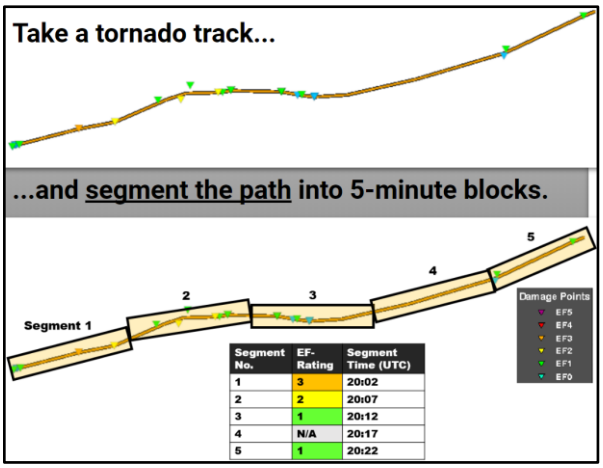

Figure 2. A diagram of a tornado track from the NWS DAT (top) with the damage points (small colored triangles) and track centerline. The track is transposed (bottom) with annotations and fictitious times for possible segmentation of the path.

A common tornado path will elicit varying degrees of damage, even within a relatively short distance. Each segment through this version of the Max/Min feature scaling methods is assigned a representative maximum EF-rating or EF class value, regardless of the number of damage points, provided there is at least one point in the segment. Once a warning is issued for a tornado, that specific issuance/valid time is the marker for the lead time provided ahead of the specific maximum EF class values along the path.

3.3 *Max/Min Feature Scaling Procedure*

The maximum (Max) and minimum (Min) values through this procedure are determined through separate means. The Max scores are determined by multiplying the maximum rating of a tornado's EF-scale with a maximum amount of lead time desired for warnings, a lead time limiter value, further defined in Section 3.5. Specifically, an EF5 tornado is the highest possible rating for a tornado and an EF0 is the lowest possible rating. By adding in a component of event intensity, through the tornado EF ratings, this information can be further used to compare with other elements of the Max/Min procedure to build a more complete narrative of the warning service.

In Table 1, EF0 is assigned a value of 0.5, to prevent a maximum associated score of zero. This particular table uses 25 minutes to compare with the EF-rating. Other times, such as 20 or 30 minutes, can be used with varying results which can be explored in future research.

EF-Rating	Maximum Score
0	12.5
1	25
2	50
3	75
4	100
5	125

Table 1. The values for maximum (Max) used for the Max/Min Scaling algorithm which correspond to the EF class for tornadoes, using a lead time limiter of 25 minutes.

NWS Instruction 10-511, WFO Severe Weather Products Specification, recommends that NWS Tornado Warnings should be valid between 15 and 45 minutes, given that tornadoes have variable lifespans. Since tornadoes are highly variable, not only from event to event but often within the same event. Most tornadoes that last more than a few minutes will strengthen from initial touchdown to a certain mature stage in their lifespan which is usually associated with the maximum damage rating and then dissipating. Varying levels of intensity throughout a tornado's lifespan provide challenges in providing a practically perfect lead time for every tornado. A proxy for a societally-averaged maximum lead time was set at 25 minutes for usage in the Max/Min Scaling algorithm. The 25 minute choice also correlates to the optimal lead times suggested by Simmons and Sutter (2011) and Hoekstra et al. (2011). A limit on increasing score based on lead time also prevents a few extremely high lead time events, e.g. 52 minutes, from skewing average lead time numbers significantly. Additionally, a maximum warning lead time prevents the notion of substantially-long valid times, e.g. 75 or 90 minutes, from becoming a preferred application for common warning service, when long track and incredibly destructive tornadoes are rare and challenging to predict that far in advance with individual thunderstorms in their infancy stage. More information on this element is described in Section 3.5.

EF-Rating	Minimum Score
0	0
1	-10
2	-50
3	-100
4	-150
5	-200

Table 2. The values for minimum (Min) used for the Max/Min Scaling algorithm which correspond to the EF class for tornadoes, using a lead time limiter of 25 minutes.

Conversely, the Min score in the Max/Min feature scaling method, shown in Table 2, is derived through the quantification of the asymmetric penalty function (Fine 2004). This notion is of an incongruent penalization for a "miss" of an event, relative to the "hit" of an event, of the same intensity or magnitude. Tornado event demonstrate this well, regardless of the lack of prior quantification for this notion. Whether a person is directly impacted or indirectly impacted by a tornado event, an unwarned tornado that produces EF5 damage is viewed as being significantly worse than in terms of warning service, than if the tornado was warned for in advance.

Utilizing the feature scaling Max/Min algorithm to determine a scaled value for a particular event, in this instance, a warning, the total event score is based on an accumulation of scores from the individual segments. Each 5-minute segment is processed through the same algorithm, where the score for each segment is determined by comparing the integer value from the EF class of the segment (e.g. 1, 2, 3, etc.) to the lead time, in minutes, between the valid/issuance time of the warning and timestamp of the individual segment.

3.4 *Warning-related Scenarios*

As with the current paradigm for statistical performance measures, the four warning-related scenarios described in this section are, fully-warned, un-warned (missed), false alarm, or partially-warned. These are not official terms used by NWS, but are used here for demonstration purposes only.

A fully-warned event will be treated through the procedure in section 3.3, by determining the segment event scores and eventually the total warning score, if there is more than one segment involved. All values for the event score will be positive integers as each segment will have a positive value for the advance lead time, even if the first segment has a lead time of one minute.

Conversely, an unwarned or missed event, where a tornado track occurred entirely without a coincident valid Tornado Warning, temporally or spatially, will then contain negative values for lead time, as they relate to each 5-minute segment. Initial touchdown of the tornado counts as the start of the unwarned event and then each segment begins every 5-minutes thereafter, through the entirety of the tornado track/path. Feature scaling within the Max/Min produce, prevents individual segment event scores or the overall warning score from being a negative value. All unwarned or missed events, since there was no qualifying warning, receive an overall warning score of zero.

False alarm warnings, having no qualifying tornadic event, are tracked separately and similarly to the current methodology. These warnings would still be placed through the legacy FAR metric, as a separately tracked ratio of fully-warned and partially warned events versus warnings with no events.

Partially-warned events are considered to be warnings that were issued with a negative lead time but eventually the warning did capture a future portion of the tornado path, both within the valid warning time and spatial area. These scenarios follow the same procedure and methodology as fully-warned events, with the difference of a negative lead time being used in the event score for any segment that has a

representative time that was before the issuance of the warning. Ultimately, the total warning score will be impacted negatively but it would depend on the degree of the damage (i.e. tornado intensity) relative to the negative lead time and not just one or the other. Instances where a warning was issued one or two minutes after touchdown. If the tornado damage at those initial points was minimal, that segment may not substantially degrade the overall warning score. Segment scoring is also dependent on the characteristics along the remainder of the tornado track.

3.5 *Lead Time Limiter*

Lead time is the most sought after element related to the issuance of a Tornado Warning. However the question of "too much" lead time is one to explore and is also a foundational part of the feature scaling method. In order to properly assess the value of an individual occurrence (warning) within a dataset, there must be a highest attainable or maximum amount of the desired outcome. When considering lead times for tornadoes, the immediate consideration is about how to link lead times to tornado intensities. The current legacy GPRA metric for average lead time does institutionalize one specific lead time amount for all tornadoes, regardless of intensity, and without delineation or ranges for the varying needs of the public. Since FY2012, the NWS goal is 13 minutes for lead time on tornadoes.

As mentioned in Section 3.3, applying lessons in societal response to warnings is necessary to understand and inform the services provided from warnings. A vast majority of tornadoes are on the weaker end of the intensity scale as well as lower end of the duration and path length/width range. All tornadoes, regardless of intensity, are assumed to be dangerous to human life, especially if flying debris is involved, unsecured or poorly objects, and the like. However, with only a small percentage of tornadoes (EF3-5; 3.7% based on 2008-2017 nationwide values) producing the vast majority of destruction, loss of life and injuries, focusing performance across the spectrum of tornadic threats may provide opportunities for improved service.

In order to approximate an appropriate statistical measure for the feature scaling method of Tornado Warnings, a lead time limiter must be applied which acts as the most ideal or most desired amount of lead time to aim for in any warning. As mentioned in Section 3.3 from NWSI 10-511, these warnings should be valid between 15 and 45 minutes which places the lead time limiter most appropriately in the 20 to 30 minute range. Utilizing all available data in the current state of the science, from high-resolution fields in convective-allowing models (CAM), to remotely sensed fields to satellite and radar, and even prediction-based algorithms lead times remain in the 10 to 15 minute range for supercell tornadoes. Linear-driven tornadoes, such as Quasi-Linear Convective System (QLCS) mesovortex tornadoes and other less organized modes may be on the order of 5 to 10 minutes (Brotzge et al. 2013).

Utilizing the 20 to 30 minute lead time limiter, graphs for 20, 25, and 30 minutes, relative to the amount tornado EF class are visualized graphically in Figure 3.

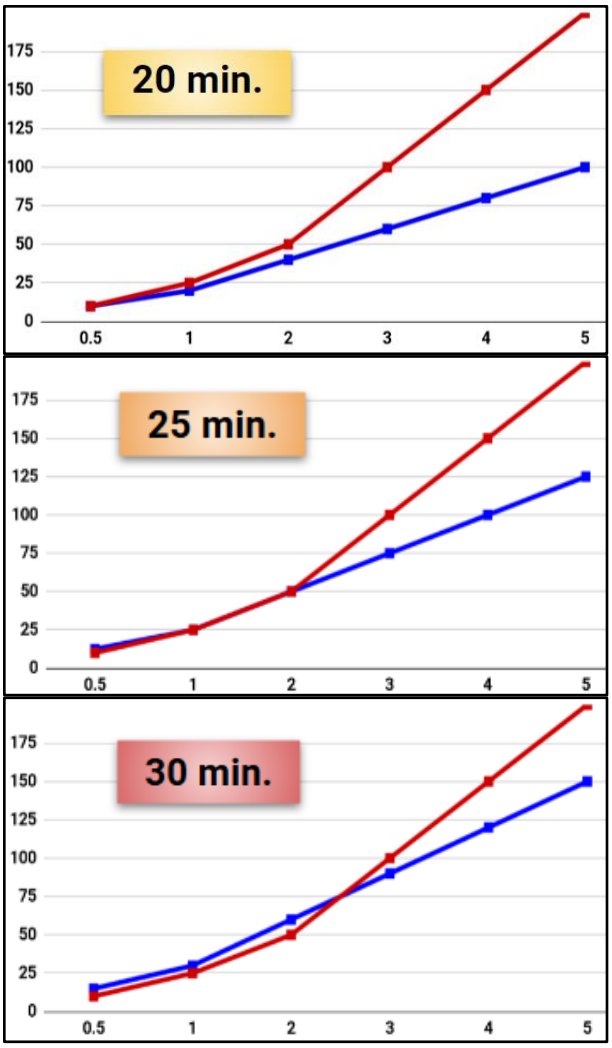

Figure 3. Event minimum scores (red) and maximum scores (blue) at 20, 25, and 30 minutes lead time, showing the increasing asymmetric penalty of missed significant events.

For each of the above graphs, the blue line represents the lead time multiplied by the EF class (integer) with EF0 = 0.5 instead of 0. The red line is a currently subjectively derived value that links to the placement of a value of the minimum amount, or penalty-like value, for missing that same EF class. This linkage puts a quantitative value on the asymmetric penalty function.

4. CONCLUSIONS

Incorporating a system and methodology such as this, through the feature scaling/data normalization methodologies, can result in a more representative

measure of performance. Results from this work may elicit a representation of high levels of performance, particularly for significant events and high-impact events, through appropriate statistical and analytical weighting. Ultimately, this methodology better represents the types of lead times and varying predictive skill of tornadoes that form from the different convective modes. By association and proper categorization of tornadic threats, more emphasis is placed on warning in advance for tornadoes that will eventually become more significant and/or longer track events. Since false alarms are not factored into the scoring system, using this scaling methodology in concert with FAR should still reveal useful information about the overall level of service provided.

Although the focus of this work is to explore the nuances of statistically representing the service or value of an individual warning, the ultimate goal to demonstrate the totality of service provision is in the comprehensive methodology. Warnings that are processed through feature-scaling methods are not averages to attain one primary number or value. These methods provide similar characterization, whether it is for one segment, one warning, multiple segments or multiple warnings. Scaled values show benefits primarily in a couple of different ways. First, if all of the scaled warnings in a sampling are compiled for an individual WFO for a calendar year, as is commonly done, the number of events in that year, for the office becomes irrelevant. The key element to achieving a high value for the ultimate comprehensive value is the performance for the most intense events within that calendar year. Similarly, these methods can provide similar results for groups of WFOs or nationally or for individual tornado outbreaks or episodes. Over time, feature-scaled values for warnings could provide a supplementary representation of warning skill and service provision that will fluctuate little from season to season, regardless of the relative tornado activity within each season or year.

NWS meteorologists are charged with providing the public advance notice for tornadoes. Forecasters are becoming more cognizant of variability of intensity in tornadoes, as well as other weather threats. The first step to informative analytical involvement in such metrics is to define newly considered parameters, whether it is through the feature scaling methodologies or some statistical method. Specific terminologies and the numerical standards for them can change with more information, testing, and evaluation. Ultimately, enhanced analytical methods may assist in understanding and demonstrating a more representative measure of performance in the warning service.

Disclaimer: The scientific results and conclusions, as well as any view or opinions expressed herein, are those of the author(s) and do not necessarily reflect the views of NWS, NOAA, or the Department of Commerce.

5. REFERENCES

Aksoy, S., and R.M. Haralick, 2011: Feature normalization and likelihood-based similarity measures for image retrieval. Pattern Recognition Letters, 22(5), p 563-582.

Brotzge, J.A., S.E. Nelson, R.L. Thompson, B.T. Smith, 2013: Tornado Probability of Detection and Lead Time as a Function of Convective Mode and Environmental Parameters. Wea. Forecasting. 28. 1261-1276.

Ferree, J.T., Freeman, D., Looney, J.M., 2007: Storm-based Warnings: Changes to NWS Warnings For the Digital Age. 35th Conference on Broadcast Meteorology, http://ams.confex.com/ams/pdfpapers/120818.pdf

Fine, G.A., 2004: Authors of the Storm: Meteorologists and the Culture of Prediction. University of Chicago Press. p. 63.

Fujita, T. T., and A. D. Pearson, 1973: Results of FPP classification of 1971 and 1972 tornadoes. Preprints, Eighth Conf. on Severe Local Storms, Denver, CO, Amer. Meteor. Soc., 142–145.

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.

Murphy, A.H., 1993: What is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. Wea. Forecasting., 8, 281-293.

NWS, 2008, Enhanced Fujita Scale for Tornado Damage, Available online at: https://www.spc.noaa.gov/efscale/

NWS, 2018a: Damage Assessment Toolkit. Available online at: https://apps.dat.noaa.gov/StormDamage/DamageViewer/

NWS, 2018b: Performance Management Database. Available online at: https://verification.nws.noaa.gov/services/public/login.aspx

NWS Warning Decision Training Division, 2018: RDA/RPG Build 18.0 Training, https://training.weather.gov/wdtd/buildTraining/build18/presentation/presentation.html

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. Wea. Forecasting, 5, 570–575.

Simmons, K.M. and D. Sutter, 2009: False Alarms, Tornado Warnings, and Tornado Casualties. Wea. Climate Soc., 1, 38–53.

_____ and _____, 2011: Economic and Societal Impacts of Tornadoes. American Meteorological Society.