INVESTIGATING INTRA-STORM FALSE ALARM TORNADO WARNINGS DURING TORNADO OUTBREAKS

Evelyn R. Bohlmann[1, 2], Matthew D. Flournoy[3, 4], and Makenzie J. Krocak[3, 5, 6]

[1]National Weather Center Research Experiences for Undergraduates Program
Norman, Oklahoma
[2]Department of Geography and Meteorology, Valparaiso University
Valparaiso, Indiana
[3]Cooperative Institute for Severe and High-Impact Weather Research and Operations
Norman, Oklahoma
[4]NOAA/National Severe Storms Laboratory
Norman, Oklahoma
[5]Institute for Public Policy Research and Analysis, University of Oklahoma
Norman, Oklahoma
[6]NOAA/NWS/NCEP/Storm Prediction Center
Norman, Oklahoma

ABSTRACT

Tornado outbreaks carry the potential for catastrophic, widespread damage to property and loss of life, making timely, accurate warnings a necessity. Effective forecasting in these outbreak situations has resulted in a relatively high probability of detection and adequate lead times. A consequence of this, however, is the common issuance of "false alarm" warnings (meaning, a tornado does not occur within the warning area during the valid time). The approach to assessing warning performance has typically been on a larger scale. None so far have assessed the occurrence of false alarms on the scale of an individual storm within a tornado outbreak. We were particularly interested in where "false alarm" warnings tend to occur within a storm's lifetime (e.g., whether they are more common before the first tornado that the storm produces, after the last, or between two separate events). By matching both verified and unverified warnings issued during tornado outbreaks to individual storms, we found that the majority of false alarm warnings were issued after the first tornado that a storm produces (both between tornadoes and after the last), while the first warnings on each storm had a very low false alarm rate. However, storms that remained nontornadic had fewer warnings issued on average than tornadic storms. This shows a lack of understanding of the phases before a storm produces its first tornado and after it has produced its final tornado. Future research could focus on these stages in order to improve forecasting skill and therefore warning performance.

1. INTRODUCTION AND BACKGROUND

In any given year, approximately 1200 tornadoes occur in the United States, more than any other country (NCEI). Very few of these events are considered violent tornadoes (EF-4+), but tornadoes of any rating have the potential to cause damage to property and loss of life. Timely, accurate warnings are a necessary piece of the puzzle to ensure proper preparation in advance of tornado occurrence. Therefore, disseminating tornado warnings to those who could be affected is a high priority. Weather alerts can be received via NOAA weather radio, a multitude of apps on cell phones, and the various routes that one can receive automated Emergency Alert System notifications, to name a few. The ultimate goal of the tornado warning is to give people who may be affected by these hazards time to prepare, which often means issuing warnings on storms that have not yet produced a tornado despite the possibility that the warning may never verify (i.e., a false alarm). A false alarm warning is far less risky than allowing a potentially tornadic storm to go unwarned, but a low false alarm rate (FAR) is ideal. As a consequence, the NWS's task of balancing longer lead times with fewer false alarms has proven to be extremely difficult. While

[1] *Corresponding author address:* Evelyn Bohlmann, Valparaiso University: evelyn.bohlmann@valpo.edu

120 David L. Boren Blvd, Suite 2500
Norman, OK 73072-7309, USA

false alarms are inevitable, they signal that there may be a weakness in warning issuance and short-term forecasting that should be addressed. The urgent, time-sensitive nature of these products and their direct relation to the protection of life and property make them a topic of particular interest and scrutiny.

Tornado warning performance has been analyzed under many different circumstances, time, and space scales. In a pair of studies done by Brotzge and Erickson (2009) and (2010), lead time and probability of detection (POD) were the foci. These studies concluded that many different factors can influence warning performance, including the order of tornadoes within each convective day (1200–1159 UTC). Only 55.8% of first tornadoes of the day are warned (Brotzge and Erickson 2010), and if a warning is issued, they are more likely to have negative lead times (Brotzge and Erickson 2009). POD generally increases for first tornadoes as the background kinematic environment becomes more supportive of tornado production (e.g., in terms of shear or storm-relative helicity; Krocak et al. 2021). A more favorable, predictable background environment could increase forecaster confidence. Other human factors like forecaster experience also likely play a role; Weather Forecast Offices (WFOs) that issue tornado warnings less frequently tend to have lower probabilities of detection and a greater chance of negative lead times than WFOs that issue more warnings (Brotzge and Erickson 2009, 2010).

The general consensus between these prior studies is that many small-scale factors influence tornado warning performance during convective events. While larger-scale, annual warning performance is well understood (e.g., Brooks and Correia 2018), many unknowns exist on smaller scales. Studies on the sub-daily (Krocak et al. 2021) and daily (Brotzge and Erickson 2009, 2010) timescales begin to fill this gap in knowledge, but warning performance on the storm-scale remains much less explored. Supercell thunderstorms are, by definition, long-lived. While the average tornado warning is valid for roughly 30 minutes (NWS), the typical lifespan of a supercell is one to two hours, or 3–4 times as long as the duration of a typical warning (Bunkers et al. 2006). Many storms outlive this average, as well, and may prompt up to 10 times as many warnings as they cross WFO boundaries and state lines. Using this small scale, we are able to gain some insight into short-term forecasting skill.

One prior study by Chamberlain et al. (2022) examined storm-scale tornado warning performance in an effort to expand upon prior work in warning performance respective to tornado order. Their novel database includes 4103 tornadoes within 75 outbreaks that occurred from 2008–2014. Outbreak tornadoes were defined using the kernel-density estimation clustering method discussed in Anderson-Frey et al. (2018). For this database, each outbreak was given an outbreak ID, then, using archived radar data, tornado reports, and warning verification information, each tornado was attributed to an individual storm ID within its associated outbreak. Similar to the Krocak et al. (2021) study, the tornadoes produced by each storm were grouped by order into four categories: first, middle, last, and only tornadoes. Tornadoes on a storm that only produced one tornado were designated as “only” tornadoes. For storms that produced two or more tornadoes, the first and last were designated appropriately, while all tornadoes that occurred between those designated as “middle” tornadoes.

Lead time and POD within those categories were the most significant results; it was found that, consistent with prior studies on tornado order, the first and only tornadoes produced by a given storm tended to have shorter lead times and a lower POD than their middle and last tornado counterparts. Geographical region and diurnal cycle were also taken into consideration. Similar trends in POD by tornado order were evident in all 4 regions, but regional differences did exist within a given category (for example, only 65% of first tornadoes were warned in the Central region, compared to 75% in the Eastern region). Additionally, the broader trend in POD by tornado order was still evident when considering only daytime or nighttime tornadoes, but nighttime tornadoes consistently had a lower POD than daytime tornadoes. Despite the inclusion of NWS warning verification information, the Chamberlain et al. (2022) study relied on tornado reports. Our complementary study also relies on the tornado warnings issued on each storm in order to investigate the occurrence of false alarms under outbreak conditions.

2. DATA AND METHODS

Three main data sources were used for the storm-scale warning performance analysis. The first is warning data, obtained through the NWS Verification website (verification.nws.noaa.gov). This included the issuing WFO, valid time and date, Event Tracking Number (ETNs), and any verified tornado event IDs (different from the manually assigned storm

and outbreak IDs) for each warning. The second dataset consisted of archived NEXRAD Level-II radar data with all products available, downloaded from NOAA/NCEI. We used data from every radar site that sampled at least one tornado-warned storm, regardless of whether that warning verified. Approximately one hour of radar data from before the first warning of the outbreak within that CWA was issued and after the final warning expired was included in the dataset. Additional data from certain radar sites were added as needed to improve analysis in regions with poor radar coverage, either due to distance from the radar or lack of availability for a certain time frame. We did this to reliably view each storm's entire lifetime from multiple perspectives, which was especially helpful when considering storms that crossed CWA borders.

The third data source is the database of outbreak tornadoes (Chamberlain et al. 2022). We were primarily concerned with the outbreak and storm IDs; these became the reference points for our analysis. The outbreak ID is a shorthand identifier for which outbreak the tornado is associated with, which eliminates the need to adjust for any outbreaks that span multiple days, yet allows for individual outbreaks within outbreak sequences to be considered separately. The storm ID connects each tornado to a particular storm; in other words, if a storm were to produce multiple tornadoes, those tornadoes would all be connected to the same storm ID. The event ID, event start time, start and end locations, EF scale rating, lead time, and percentage of the event that was warned are included for each tornado in addition to the storm and outbreak IDs.

In order to pair storm IDs with warnings (either verified or false-alarm), radar data was viewed using GR2Analyst and assessed manually using the variety of radar products available. We were able to use the event IDs, locations, and start times given in the Chamberlain et al. (2022) database to attribute a verified warning to a given storm. The outbreaks analyzed represented various storm modes, times of day, and geographical locations, and as a result, radar analysis of storm objects inherently included some subjectivity. The Chamberlain et al. (2022) database was also compiled through manual analysis, and while each analysis was performed with similar tools, there were a few discrepancies between our analysis and the Chamberlain et al (2022) database. These discrepancies were rare, and when they were present, it was clear that they were simply a result of different interpretations and were therefore consistent within each dataset (i.e., two separate storms could be recorded as one, where all tornadoes produced by both storms were labeled as that storm, or one storm could be recorded as two with distinct storms labeled accordingly). In these cases, we generally accepted the Chamberlain et al. interpretation for consistency. Base reflectivity, base velocity, vertically integrated liquid, and spectrum width were particularly helpful in locating the characteristic persistent updrafts as storms evolved, especially in ambiguous linear modes and merging or splitting cells. GR2Analyst contains a mesocyclone detection algorithm, which was not reliable enough to be applied in the majority of cases. Valid warning polygons and associated warning text are included within the radar files, allowing us to associate a given storm and associated series of warnings with the information in both the tornado warning database and the Chamberlain et al. (2022) database.

After each warning is attributed to a storm ID, warning information (valid time, ETN, and verification) and outbreak/storm information (outbreak ID, storm ID, event start times, event ratings, lead time, and the percentage of the event that was warned) were compiled into a single database that can then be analyzed using the Python language. We chose to assess the number of tornadic versus nontornadic storms, the numbers of false alarms on each of those categories' storms as well as their position in that specific storm's lifetime (before the first warned tornado, between the first and last warned tornado, and after the last warned tornado), and the warned lifespan of tornadic and nontornadic storms in order to get a comprehensive picture of the warning process.

3. RESULTS

Within our dataset, there were 724 warnings issued on 253 individual storms, both tornadic and nontornadic. Seven outbreaks were included in our dataset, outlined in Table 1.

Across all seven outbreaks, 131 storms were tornadic (meaning they produced at least one tornado during the analysis period), and 122 were nontornadic for the analysis period. The majority of tornadic storms underwent nontornadic phases, in which the warnings issued on them were false alarms, but not all. As mentioned earlier, supercell thunderstorms are long-lived and often prompt multiple warnings, which was well-represented in our dataset. The maximum number

Outbreak ID	Date	Primary Region	Tornadic Storms	Nontornadic storms
Outbreak 1	1/7–1/8/2008	Southern	22	16
Outbreak 2	2/5–2/6/2008	Southern	31	34
Outbreak 5	4/9-4/11/2008	Central	21	17
Outbreak 9	5/24/2008	Central	12	9
Outbreak 11	5/30/2008	Central	14	14
Outbreak 14	6/7/2008	Central	16	17
Outbreak 15	6/11-6/12/2008	Southern and Central	23	13

Table 1.

of warnings on a single storm was 40 (on a supercell in Dodge City on May 23rd, 2008); only 5 of those warnings did not verify. Additionally, 20 of the tornadic storms were associated with 10 or more warnings, and 7 storms with 20 or more warnings. This is also a testament to their potential to produce long-track and/or cyclic tornadoes during tornado outbreaks. There was no evident correlation between the total number of warnings issued on a storm and the number of false alarms on that storm, as seen in Figure 1.

Our first significant result is that a slight majority of warnings issued on tornadic storms verified. Roughly 42% of warnings issued on our tornadic storms, regardless of order within each storm, were false alarms. This is far from perfect, but smaller than the national average of 70%. This is likely heavily influenced by the fact that we only examined tornado *outbreaks*, which are associated with more supportive background environments and increased warning performance (Krocak and Brooks 2021). Interestingly, the number of false alarms issued on tornadic storms is similar to that of nontornadic storms (214 to 213, respectively), but 510 warnings in total were issued on tornadic storms. This was our first indication that nontornadic storms typically have a shorter warned lifespan than tornadic storms. In other words, fewer warnings are issued on average for a nontornadic storm than a tornadic storm. We found that an average of 1.75 warnings were issued on individual nontornadic storms, as opposed to 3.89 on tornadic storms. This could mean that forecasters are relatively good at identifying when a previously warned but nontornadic storm is unlikely to become tornadic, even if there were a few erroneous warnings.

Additionally, we examined the verification on only the first warning issued on each storm. Recalling the results from Chamberlain et al. (2022) and similar studies on other scales, first tornadoes have a lower POD, which is typically

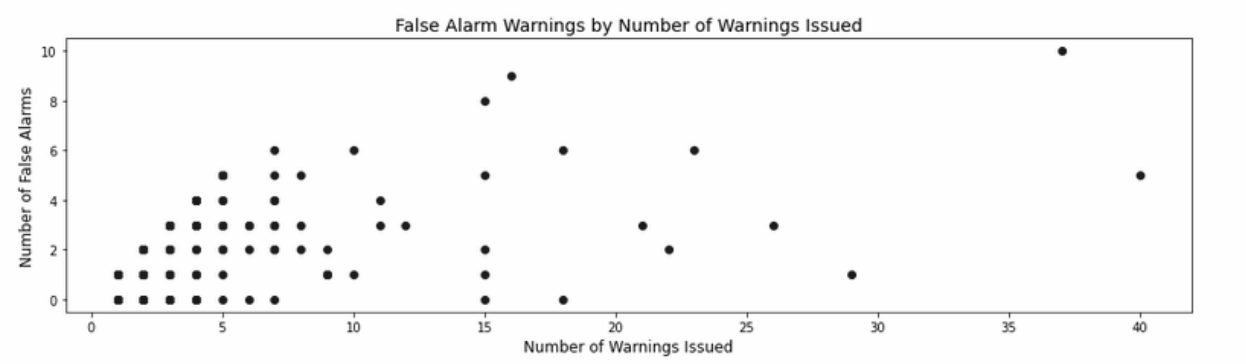

Fig. 1. Each point within this scatterplot represents one storm. All points on the line x=y represent a nontornadic storm, in which each warning issued was a false alarm.

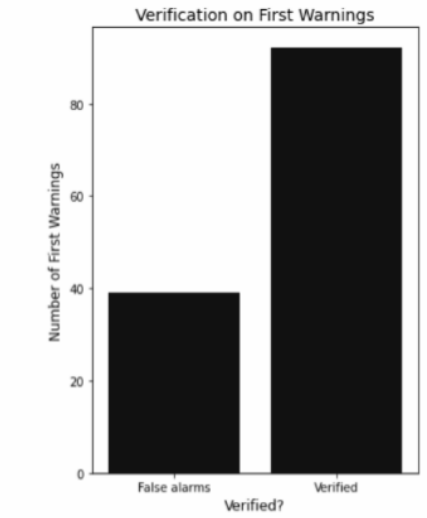

Fig. 2.

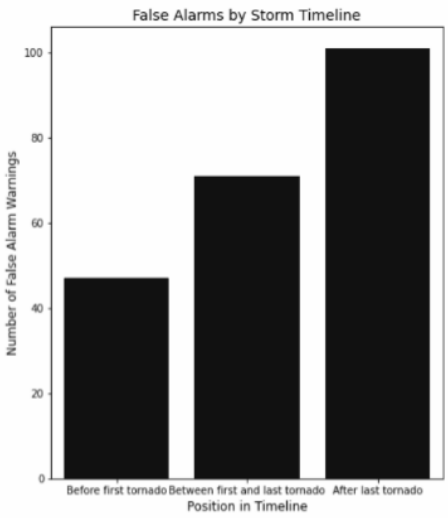

Fig. 3.

related to a lower FAR as discussed in Brooks and Correia (2018). Our results were consistent with these findings. Fewer than one-third of the first warnings on each tornadic storm (n = 39) were false alarms, as shown in Figure 2. Considering that first tornadoes tend to have lower lead times (if they are warned in advance at all) and that their POD is low even on the storm scale, this low FAR is unsurprising.

Finally, we analyzed the position of each false alarm on each tornadic storm[2]. False alarm warnings on tornadic storms were sorted into three timeline categories: before the first warned tornado, between the first and last warned tornado, and after the last warned tornado. False alarms issued prior to the first warned tornado accounted for only 21.46% (n = 47) of all false alarms issued on tornadic storms. First tornadoes have been shown to have shorter lead times on multiple time scales (Chamberlain et al. 2022, Brotzge and Erickson 2009). Consistent with the discussion in Brooks and Correia (2018), this is related to lower POD and lower FAR. The majority of false alarms on tornadic storms occur either between the first and last tornado (n = 71, 32.42%) or after the last tornado (n = 101, 46.12%).

This is the most striking finding from our study; it suggests that there is either a lack of skill in knowing when a supercell has entered a nontornadic phase, forecaster hesitancy to miss a tornado if a given storm has a history of producing them, or both. Warnings issued on long-lived storms were almost always continuous, meaning new warnings were typically issued at the edge of the expiring warning in order to provide continuous warning coverage as the storms progressed. This could indicate that distinguishing between nontornadic and tornadic phases of a storm is difficult in real-time operations, resulting in the increase of false alarms after the first tornado.

Lastly, the greatest number (n = 101, 46.11%) of false alarms issued after the final warned tornado had dissipated highlights another unique challenge: identifying the end of a storm's ability to produce a tornado. Ideally, an increase in predictive skill during this phase would reduce FAR downstream of the last tornado without decreasing POD of the final tornadoes that each storm produces.

[2] The number of unverified false alarms in these calculations is greater than the number of unverified warnings on tornadic storms due to three verified warnings that were explicitly stated in the warning text to have been issued on multiple storms (including locations of these storms), but only verified for one storm. These were each counted as verified only for the storm that produced a tornado, and a false alarm for the other storms that were not producing tornadoes at the time. Five instances of tornadic storms in nontornadic phases were covered. One of these "verified false alarms" occurred before the first tornado, two occurred between tornadoes, and two occurred after the last tornado for the individual nontornadic phases that they covered.

4. DISCUSSION

As mentioned previously, warning issuance is not a perfect process and warrants investigation on multiple spatiotemporal scales to improve our understanding of weaknesses in the current tornado forecasting paradigm. This study is one of the first to do so on the storm-scale (i.e., pairing tornado warnings with individual storm objects), and our hope is that it will provide a novel framework for assessing false alarm tornado warnings on the intra-storm level. No one method of evaluating warning performance provides a complete perspective on its own, including the one introduced here. The knowledge that roughly 70% of all tornado warnings are unverified prompted the studies on smaller scales, whether those be regional, sub-daily, or by individual storm. Larger scales are relatively well-studied and understood, necessitating the introduction of this storm-scale work to the discussion on warning performance as a whole. We hope to inspire a continual assessment of warning performance on smaller scales as more warning data is produced and available.

Aside from its use as a template for other studies to follow or adjust as needed, our assessment provides the first look into warning performance on the storm level, which opens up a topic of discussion on what these (and future) results could potentially mean. False alarms are often viewed as “failures”. A tornado warning is meant to communicate that a tornado is either already occurring or imminent, so an unverified warning would mean that the warning was ineffective. This study shows that nontornadic storms rarely prompt more than a few warnings. In fact, the maximum number of warnings issued on a single non-tornadic storm in our data set was 5, while the maximum issued on a tornadic storm was 40. Warnings on nontornadic storms accounted for roughly half of the unverified warnings in our dataset. False alarms on tornadic storms, however, yield a different perspective. When considering that the vast majority (over 80%) of false alarms on tornadic storms occurred after the first warned tornado, it is reasonable to assume that forecasters may be hesitant to miss an event. After all, POD is still a concern; it might be “safer” to issue a false alarm warning than leave a tornado unwarned and put others at risk by not giving them time to prepare for impact. This ultimately is a result of forecaster philosophy and whether preference is given to prioritizing high POD *or* low FAR with our current understanding and imperfect predictive ability. This is the first study to show the impacts of this philosophy (i.e., higher FAR after the first warned tornado) on the storm scale.

The greatest question going forward, aside from those regarding more detailed analyses with regards to diurnal cycle, geographical location, and other factors, is: why do nearly half of all false alarms occur after the last warned tornado? First warnings get quite a bit of attention for their poorer performance in metrics such as POD and lead time, but the high FAR after the last tornado that a storm produces indicates that not only are forecasters unsure of when exactly tornado production will start, they are unsure of when it will stop. This prompts questions about not only tornadogenesis but tornado demise. What signals the start of a nontornadic phase on a storm that has a history of producing tornadoes is this knowledge useful for real-time prediction? Additionally, what indicates the ultimate end of a storm’s tornado production? While our findings may not be immediately useful to forecasters, it gives guidance on what information we may need in the future.

Further down the line, this study and any other work that it prompts in the future could be of great significance to emergency management personnel and related roles. A reduction in false alarms between tornadoes would necessitate emergency management to understand that a non-warned storm is not necessarily a non-hazardous storm and incorporate that understanding into their response to hazardous weather. To do so, a strong relationship and communication between emergency management and local WFOs, especially during tornado outbreaks, is crucial. However, the knowledge that the warnings issued under these circumstances are less likely to be false alarms could lead to better-informed decisions on where to allocate resources and in what quantities. Authority figures (i.e. school officials, church leadership, etc.), similarly, would be able to make better-informed decisions on when and how to take action, whether it involves canceling a gathering in advance or taking shelter.

All things considered, though, the perception of FAR as a performance metric for tornado warnings might warrant reevaluation. The “cry wolf effect” does not appear to influence the reception of or confidence in tornado warnings (Lim et al. 2019), yet a low FAR is generally desirable despite its correlation with lower POD. Therefore, should forecasters prioritize increasing warning skill to improve low POD in certain

situations, at the expense of potentially increasing FAR? An example in this particular study is increasing POD of the first tornadoes of each storm. Should researchers focus on better understanding the processes that lead to the first tornado that a given storm produces? Our findings suggest that, in terms of impact on real-time operations, both the pre- and post-tornadic phases (meaning, the phases before the first and after the last tornado, respectively) would be ideal foci for future research on tornado processes.

5. SUMMARY

Tornado warning performance on every spatiotemporal scale lends a different, significant perspective to current forecasting skill. Many factors can influence warning performance, both human and environmental, but one significant circumstance that tends to improve performance from the national average is occurence as part of a tornado outbreak. Even with this improvement, however, tornado warning performance remains imperfect, and if forecasting skill is to be improved, we must understand where the improvements need to be made. This study is the first to investigate false alarm tornado warnings on the storm-scale to identify where storm-scale forecasting is falling short. To do so, we paired 724 warnings with 253 storms from 7 tornado outbreaks that occurred throughout 2008. Both tornadic and nontornadic storms were analyzed. Each was connected to a unique storm ID using archived NEXRAD Level-II radar and NWS warning verification data. Warnings on tornadic storms were given a storm ID corresponding to the database created by Chamberlain et al. (2022), while nontornadic storms were given a “negative” storm ID in a similar manner.

The most significant result of this study is that only 21.46% of false alarms on tornadic storms occur before the first warned tornado, while false alarms between tornadoes (32.42%) and after the final warned tornado (46.12%) are more frequent. Additionally, among the first warnings on each tornadic storm, only 29.77% were false alarms, and tornadic storms had more warnings on average than nontornadic storms. This is indicative of two separate forecasting weaknesses. Firstly, in terms of false alarms, first warnings perform better than subsequent warnings, but as previous literature shows, this is not the case with other metrics, meaning that first tornadoes may sometimes be warned when already in progress. Because of this, research should continue into the processes leading to the development of the first tornado that a storm produces. On the other hand, the majority of false alarms occurring after the first tornado shows a lack of understanding of what leads supercells to have nontornadic phases and what indicates a storm’s end of tornado development. Our conclusions also open up a discussion about maintaining a balance between a high POD and a low FAR, and whether the current view of a low FAR as more ideal should be reconsidered.

6. ACKNOWLEDGMENTS

We would like to thank Dr. Daphne LaDue and Alex Marmo for directing the National Weather Center Research Experience for Undergraduates program.

This work was prepared by the authors with funding provided by National Science Foundation Grant No. AGS-2050267, and NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, NOAA, or the U.S. Department of Commerce.

7. REFERENCES

Anderson-Frey, A. K., Y. P. Richardson, A. R. Dean, R. L. Thompson, and B. T. Smith, 2018: Near-storm environments of outbreak and isolated tornadoes. Wea. Forecasting, 33, 1397–1412, doi:10.1175/waf-d-18-0057.1.

Brooks, H. E. and J. Correia, 2018: Long-Term Performance Metrics for National Weather Service Tornado Warnings. Wea. Forecasting, 33, 1501–1511, doi:10.1175/WAF-D-18-0120.1.

Brotzge, J. and S. Erickson, 2009: NWS Tornado Warnings with Zero or Negative Lead Times. Wea. Forecasting, 24, 140–154, doi:10.1175/2008WAF2007076.1

Brotzge, J. and S. Erickson, 2010: Tornadoes without NWS Warning. Wea. Forecasting, 25, 159–172, doi:10.1175/2009WAF2222270.1

Bunkers, M. J., M. R. Hjelmfelt, and P. L. Smith, 2006: An Observational Examination of Long-Lived Supercells. Part I: Characteristics, Evolution, and Demise. Wea. Forecasting, 21, 673–688, doi:10.1175/WAF949.1

Krocak, M. J and H. Brooks, 2021: The Influence of Weather Watch Type on the Quality of Tornado Warnings and Its Implications for Future Forecasting Systems. Wea. Forecasting, 36, 1675–1680, doi:10.1175/WAF-D-21-0052.1

Krocak, M. J., M. D. Flournoy, and H. E. Brooks, 2021: Examining Subdaily Tornado Warning Performance and Associated Environmental Characteristics. Wea. Forecasting, 36, 1779–1784, doi:10.1175/WAF-D-21-0097.1

Lim, J. R., B. F. Liu, and M. Egnoto, 2019: Cry Wolf Effect? Evaluating the Impact of False Alarms on Public Responses to Tornado Alerts in the Southeastern United States. Wea. Climate Soc., 11, 549–563, doi:10.1175/WCAS-D-18-0080.1