The Influence of Weather Watch Type on the Quality of Tornado Warnings and Its Implications for Future Forecasting Systems

MAKENZIE J. KROCAK[a,b,c,d] AND HAROLD E. BROOKS[e,f]

[a] *Center for Risk and Crisis Management, University of Oklahoma, Norman, Oklahoma*
[b] *National Institute for Risk and Resilience, University of Oklahoma, Norman, Oklahoma*
[c] *Cooperative Institute for Mesoscale Meteorological Studies, Norman, Oklahoma*
[d] *NOAA/Storm Prediction Center, Norman, Oklahoma*
[e] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
[f] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*

(Manuscript received 1 April 2021, in final form 12 July 2021)

ABSTRACT: While many studies have looked at the quality of forecast products, few have attempted to understand the relationship between them. We begin to consider whether or not such an influence exists by analyzing storm-based tornado warning product metrics with respect to whether they occurred within a severe weather watch and, if so, what type of watch they occurred within. The probability of detection, false alarm ratio, and lead time all show a general improvement with increasing watch severity. In fact, the probability of detection increased more as a function of watch-type severity than the change in probability of detection during the time period of analysis. False alarm ratio decreased as watch type increased in severity, but with a much smaller magnitude than the difference in probability of detection. Lead time also improved with an increase in watch-type severity. Warnings outside of any watch had a mean lead time of 5.5 min, while those inside of a particularly dangerous situation tornado watch had a mean lead time of 15.1 min. These results indicate that the existence and type of severe weather watch may have an influence on the quality of tornado warnings. However, it is impossible to separate the influence of weather watches from possible differences in warning strategy or differences in environmental characteristics that make it more or less challenging to warn for tornadoes. Future studies should attempt to disentangle these numerous influences to assess how much influence intermediate products have on downstream products.

KEYWORDS: Forecast verification/skill; Operational forecasting

1. Introduction and background

The National Weather Service (NWS) generates a set of forecast products that span a large range of spatiotemporal scales. Each one plays an important role in preparing the public for different impacts. However, little is known about the relationships between these products, and whether or not the issuance of one product influences the quality of another. Studies have attempted to define what a ''good'' forecast is (e.g., Murphy 1993), and more specifically study how well probabilistic forecasts have verified in specific products (e.g., Hitchens et al. 2013), but few have attempted to assess the influence of one forecast product on another.

There are three main product ''levels'' that make up the current NWS severe weather forecasting system. The first one is the convective outlook. This product is issued by the NOAA Storm Prediction Center (SPC) from 1 to 8 days in advance and is valid from 1200 UTC on a given day to 1200 UTC on the following day. Convective outlooks contain probabilities that indicate the forecast likelihood that a hazard (i.e., severe hail, severe wind, and tornado) will occur within 25 nautical miles (n mi; 1 n mi = 1.852 km) of a point within the 24-h convective day. Previous work has shown that these probabilities have increased in skill since the 1990s (Hitchens et al. 2013). The next product level is the severe weather watch, which is also issued by the SPC in coordination with local NWS Weather Forecast Offices (WFOs) and is valid upon issuance and usually lasts for 4–8 h from that time. There are different types of watches, including severe thunderstorm watches (where few, if any, tornadoes are expected), tornado watches, and particularly dangerous situation (PDS) tornado watches. These PDS watches are issued in the rare situation where confidence is high that multiple strong or violent tornadoes will occur within the watch area. Finally, the last level of the current NWS severe weather forecasting system is the severe weather warning, which is issued by local WFOs and is valid from issuance and usually lasts for 30–60 min. Warnings are typically much smaller than watches and are verified if a severe weather event occurs within the warning polygon.

This work attempts to study the relationship between an intermediate product (weather watches) and the associated downstream product (tornado warnings). This information is important to understand not just for the current NWS severe weather forecast system, but also for ongoing work and decisions being made about future severe weather forecasting systems. The NOAA Forecasting a Continuum of Environmental Threats (FACETs) project is attempting to create a communication infrastructure in which the end user

Corresponding author: Makenzie J. Krocak, makenzie.krocak@noaa.gov

DOI: 10.1175/WAF-D-21-0052.1

TABLE 1. The number of tornado events that occurred in each watch existence/type and year.

Year	No watch	Severe watch	Tornado watch	PDS watch	Total
2008	435	208	1085	347	2075
2009	389	283	577	25	1274
2010	335	177	775	158	1445
2011	283	229	1057	506	2075
2012	295	123	467	170	1055
2013	277	174	486	114	1051
2014	328	206	447	61	1042
2015	404	168	738	12	1322
2016	409	181	450	33	1073
2017	457	324	757	105	1643
Total	3612	2073	6839	1531	

TABLE 2. The number of tornado warnings that occurred in each watch existence/type and year.

Year	No watch	Severe watch	Tornado watch	PDS watch	Total
2008	876	594	3033	523	5026
2009	684	692	1712	141	3229
2010	647	574	1889	334	3444
2011	536	599	2550	949	4634
2012	512	336	1232	325	2405
2013	343	351	985	147	1826
2014	458	338	953	93	1842
2015	523	369	1348	30	2270
2016	502	427	1003	38	1970
2017	600	536	1475	195	2806
Total	5681	4816	16 180	2775	

is continually updated about the hazardous weather threats and impacts (Rothfusz et al. 2014, 2018). FACETs aims to establish a system in which forecast information is provided at many spatiotemporal scales to suit the many needs of different users. However, it is likely that the issuance of previous products can have a profound impact on forecasting philosophy and communication strategies during particularly impactful events (e.g., Hales 1989). Therefore, it is critical to understand these intricacies in the current infrastructure so that best practices can be developed for a future one.

Many studies have looked at the quality of warnings as defined by Murphy 1993 (e.g., Brooks 2004; Brooks and Correia 2018; Anderson-Frey and Brooks 2021), but few have compared the quality of warnings based on the type of upstream product they exist within. For example, does warning quality improve if the warning is within a tornado watch instead of a severe thunderstorm watch? Is warning quality a function of watch type or convective outlook category? Previous work identified early on that the severe weather watch plays an important role in tornado warning procedures (Hales 1989). Not only was the probability of detection (POD) higher for warnings within a tornado watch, but the study concluded that the watch played an important role in setting the stage for warning operations within local NWS WFOs. Additionally, Keene et al. (2008) found an increase in the POD if the tornado warning occurred in a tornado watch instead of a severe thunderstorm watch, and there is an even greater increase in POD over warnings outside of any watch. These two studies indicate that the watch type is related to the quality of tornado warnings and that the interdependencies between products need to be understood to ensure any future forecasting systems also benefit from those interdependencies.

2. Data and metrics

Tornado warning and event data between October 2007 (the start of the polygon warning era) and December 2017 were obtained from the NWS verification website. Data regarding the existence of a severe weather watch and watch type for each tornado warning were provided by the SPC for the same timeframe. Warnings and events were cross-referenced to identify verified and missed warnings/events. Since we wanted to understand how the quality of warnings changed with watch type, performance metrics were calculated for the entire database and separately for each watch type. Data from 2007 (a total of 328 warnings) were combined with 2008 (a total of 4698 warnings) since few data points exist in the fall of 2007. See Table 1 for an overview of the event sample sizes by year, and Table 2 for an overview of the warning sample sizes by year. Summary metrics were calculated to identify overall patterns within different watch types. Probability of detection (POD) and success ratio [SR, which is 1 − false alarm ratio (FAR)] were calculated for each watch type (Roebber 2009), where POD was calculated as the fraction of tornadoes warned in advance, and SR was calculated as the fraction of warnings with a tornado. Additionally, the mean lead time was calculated over tornadoes warned in advance for each year and watch type.

3. Warning performance

In general, POD increases with increasing severity of the watch (i.e., severe POD < tornado POD < PDS tornado POD; Fig. 1). The POD for warnings in PDS watches remained around 0.8 between 2007 and 2014, then decreased to around 0.7 for the last few years of the study period. The POD for tornado warnings in tornado watches remained around 0.75 until 2012 and then decreased to around 0.6–0.7. Tornado warnings within severe thunderstorm watches and within no watch had a much lower POD throughout the entire period, generally between 0.4 and 0.6 until 2012, when values decreased to around or below 0.4. Most notable is the difference between warnings without a watch or in a severe watch and warnings in a tornado watch or a PDS tornado watch. The mean POD for warnings within tornado watches was 0.76 from 2008 to 2012 and 0.65 from 2013 to 2017. Contrast that to the mean POD for warnings in severe thunderstorm watches, which was 0.52 from 2008 to 2012 and 0.42 from 2013 to 2017. There is a consistent difference in POD of around 0.20 between warnings in a severe watch and warnings in a tornado watch, which is a larger difference than the change across the time period within any watch type.

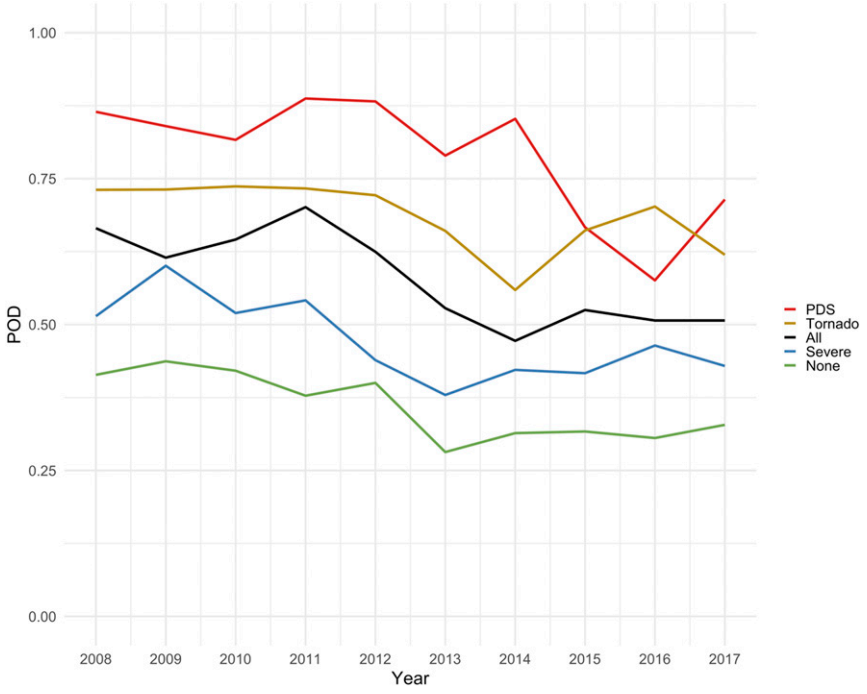

FIG. 1. Probability of detection values for tornadoes based on watch existence/type and year.

A similar pattern to POD was seen with FAR, although there is much less spread among the different watch types (Fig. 2). FAR values for warnings outside of any watch or within a severe thunderstorm watch decrease slightly over the period, while warnings within tornado watches and especially PDS tornado watches show a larger decrease in FAR over the entire period. There are a few points where PDS FAR values are not the lowest, namely 2009 and 2015.

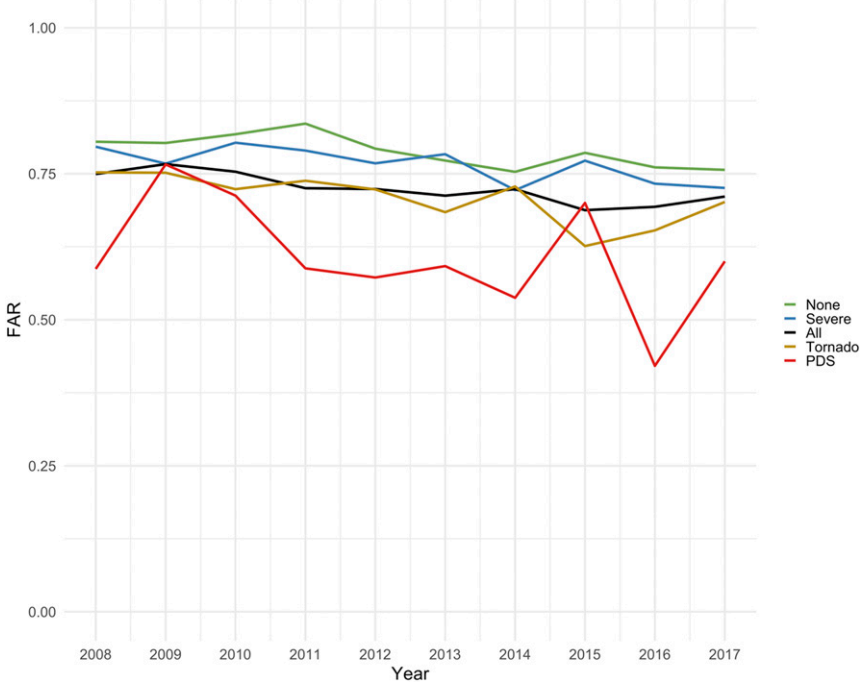

FIG. 2. False alarm ratio values based on watch existence/type and year.

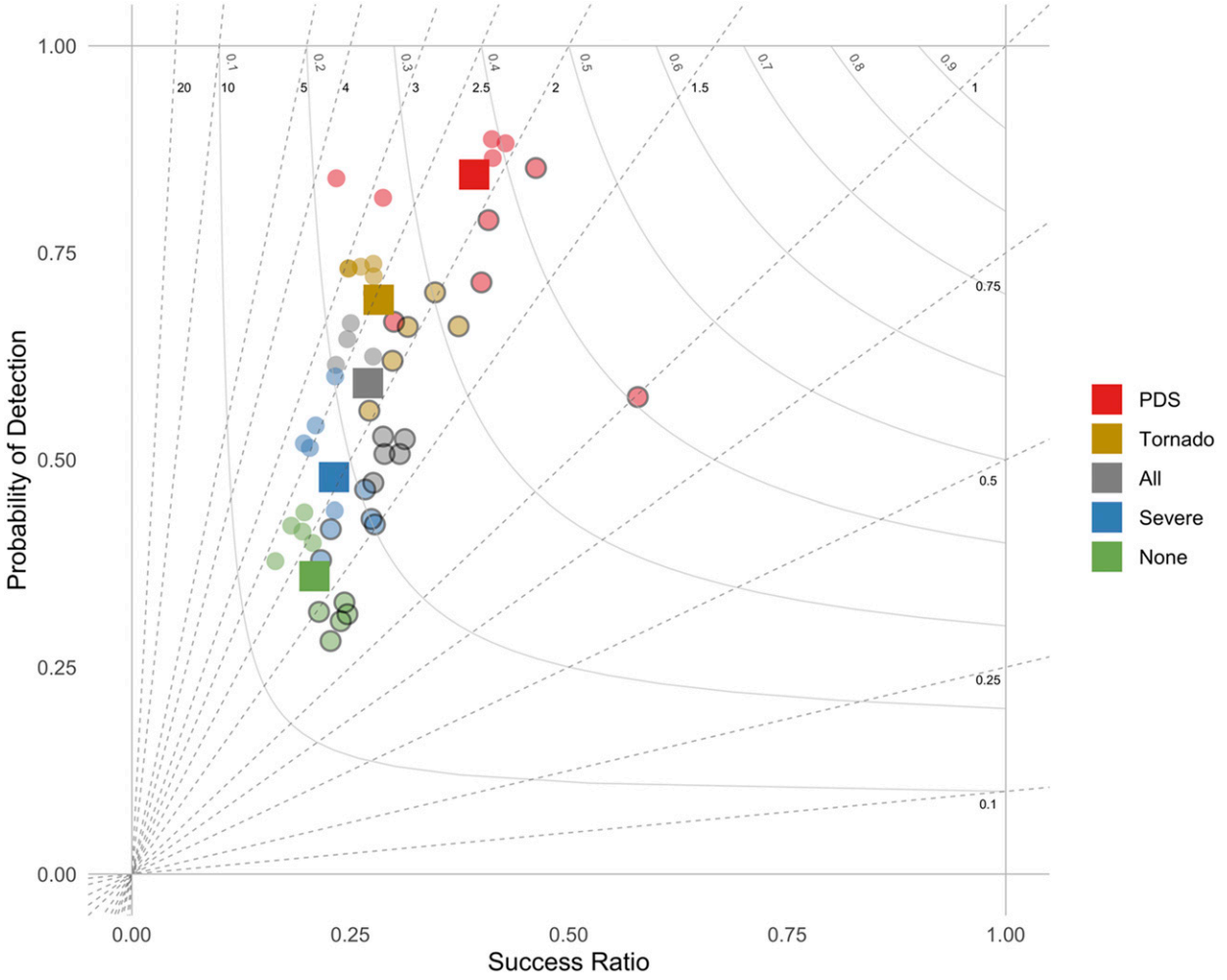

FIG. 3. Performance diagram showing tornado warning probability of detection (y axis) and success ratio (x axis) by watch existence/type. Dots without borders indicate values from 2008 to 2012, while dots with black borders are 2013–17. Square markers are the overall POD and SR (1 − FAR) for that category.

PDS FAR values are more variable likely because of the smaller sample size of warnings. Similarly, severe thunderstorm FAR is generally lower than the no watch FAR, with the exception of 2008 and 2013. The most notable difference between the POD values and the FAR values is the much smaller range between watch types. Not only are the differences between watch type much smaller (for FAR versus POD), but the change in FAR values over time (with the exception of PDS watches) is also much smaller. These differences between POD and FAR range may indicate that forecasters are still warning with similar thresholds (therefore still allowing for a relatively high FAR), but more of the tornadoes are being correctly identified and warned (allowing for a higher POD).

This information was then combined onto a performance diagram (Roebber 2009, modified from precision-recall diagrams described in Raghavan et al. 1989) to show the impact of both POD and SR (Fig. 3). The separation between watch type is evident across all types, but especially between severe thunderstorm watches and tornado watches. In fact, it is clear that the change in POD among watch types is similar or greater than the overall change in POD over the 10-yr period within any single watch type. Additionally, the warnings not in a watch and those in severe thunderstorm and tornado watches all show a similar pattern. The earlier years of the record (2008–12, shown as dots without borders) have a higher POD and slightly lower SR. Beginning around 2013, the POD lowers and the SR marginally improves. However, the PDS watch category does not follow this pattern as closely, indicating these situations are somehow different (potentially due to a smaller sample size in the PDS category). The overall pattern between the categories shows a marginal increase in SR and around a 0.1 increase in POD with each increase in watch severity.

Finally, the mean lead time for each year and each category was calculated (Fig. 4). For example, for all warned tornadoes within tornado watches, we calculated the mean lead time for each year in the dataset. Although the lead

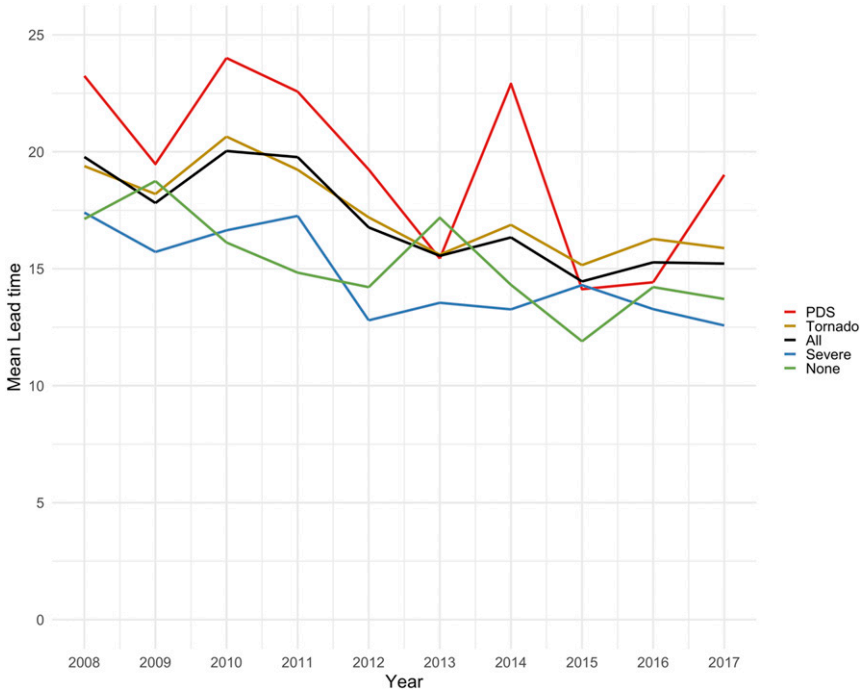

FIG. 4. Mean lead time values based on watch existence/type between October 2007 and 2017.

time in the PDS watch category is inconsistent likely due to a small sample size, there is a notable increase in tornado warning lead time between warnings that occur in no watch or a severe thunderstorm watch compared to those that occur in a tornado or PDS watch. This increase in lead time between warnings in severe thunderstorm watches and tornado watches is often around 5 min, which is significant given the mean lead time for all warnings is between 15 and 20 min for almost all years. Once again there is also a decrease in the mean overall lead time between 2011 and 2012. The overall lead time drops from around 20 min in 2011 to 17 min in 2012 and under 15 min by 2015. This drop is evident for all watch types, although the PDS category is inconsistent.

While the changes in warning skill as a function of watch types is the focus of this paper, our results support the findings of Brooks and Correia (2018). There is evidence that a change in the warning threshold occurred in 2012, resulting in generally lower POD and slightly lower FAR. This is likely due to a change in the default warning length from 45 to 30 min and an increased emphasis on reducing false alarm occurrences (Brooks and Correia 2018).

4. Discussion

The critical component of this work was to identify if intermediate forecast products impact the quality of downstream products, and if so, how they impacted downstream products. Separating tornado warnings based on watch existence and type shows that there is a difference in verification metrics based on the watch type, with warnings not in a watch generally being the least successful and warnings in PDS tornado watches generally being the most successful.

The current NWS system for severe weather forecasts and communication relies on multiple different products from different offices (i.e., SPC and local WFOs) telling a story from days (sometimes up to 8 days out) down to minutes before the event occurs. What we do not know is how these different products influence future products. In this work, we attempt to investigate the performance metrics of tornado warnings based on what type of watch (if any) they occurred within. Results showed that POD increases, FAR decreases, and lead time generally increases with increasing watch severity.

These results indicate the intermediate products (i.e., those on the ''watch'' scale) are important and are related to the quality of downstream products. However, what we still do not know is why or how the downstream products are influenced. Is it because NWS Weather Forecast Office forecasters are operating under the knowledge that other forecasters (like those in the SPC) believe something will happen, which impacts their warning decision process? Or is it because the environment within more severe watch types makes warning decisions more obvious? Work by Alsheimer et al. (2018) indicates that at least some forecasters change their warning decision process when a PDS tornado watch is in effect for their area. Alternatively, Anderson-Frey and Brooks (2021) show that warning skill is (and is expected to be) different for different environments, which ultimately means that baseline skills should be different as well. The NWS has recently increased emphasis on environmental analysis during warning operations, even having a separate meteorologist assessing the mesoscale environment

for the warning forecaster. In addition to environmental factors, radar presentation, previous storm behavior, and improvements in technology (like the introduction of dual-polarization capabilities) all influence warning decisions. This process is complex, and while this paper shows the increase in warning quality by watch type, there are many other factors that play a role in warning decisions, which cannot be summarized in a single study. Future work should continue to evaluate forecaster decision making, specifically what products, strategies, and cues are most helpful to the warning decision process.

Ultimately, this work begins to show that intermediate products likely have an influence on downstream products, pointing to the need for quality intermediate products in future severe weather forecast paradigms. We have shown that in the current system, a static product (the type of convective watch) is related to the quality of a downstream static product (tornado warnings). The FACETs project has a goal of creating evolving products, which are all related to each other. Therefore, early decisions and products produced by one forecaster could have huge impacts on what another forecaster can output. Additional work should focus on how a watch-like product could be incorporated into a continuum of always-evolving products. Could such a product be initiated 8–10 h before the event and continuously updated throughout the hours leading up to the event (similar to a ''long-lead-time'' watch)? How could this product fit into the FACETs paradigm and how would it influence the ''warning'' product performance?

Given the evidence presented in this paper, it is reasonable to surmise that the existence of a rapidly updating intermediate product would influence the quality of probabilistic warnings. This could be due to a number of factors, some of which are directly related to the product itself. Should forecasters in local NWS offices know that forecasters at a national center (SPC) believe that tornadoes will happen and continue to believe they will happen throughout the event (communicated by the updating of the intermediate product), it is reasonable to believe that the local forecasters would be primed to issue local probabilities (or warnings, or whatever downstream product exists in a FACETs paradigm). The continuous updating nature of the products would mean that warning forecasters are constantly being updated, reassured, or reoriented to the changing weather situation, potentially allowing for a more rapid shift in strategy. In a slightly different paradigm, forecasters could be managing a shared database of weather information and warning strategy, creating an even more interconnected system. Further analysis of these possibilities and others will help researchers understand the strengths and weaknesses of the current infrastructure, and should identify the characteristics that are important to maintain should a FACETs-like system be adopted by the NWS.

Acknowledgments. The authors thank Andy Dean with the Storm Prediction Center for providing watch and warning data for this work. Funding was provided in part by NOAA's Office of Weather and Air Quality through the U.S. Weather Research Program and by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce.

Data availability statement. The tornado warning and event data are available on the NWS Performance Management Web Portal (https://verification.nws.noaa.gov/services/public/index.aspx). The watch data are available from the NOAA Storm Prediction Center.

REFERENCES

Alsheimer, F., T. Johnstone, D. Sharp, V. Brown, and L. Myers, 2018: Human factors affecting tornado warning decisions in National Weather Service Forecast Offices. *13th Symp. on Societal Applications: Policy, Research and Practice*, Austin, TX, Amer. Meteor. Soc., 3A.8, https://ams.confex.com/ams/98Annual/webprogram/Paper326524.html.

Anderson-Frey, A. K., and H. Brooks, 2021: Compared to what? Establishing environmental baselines for tornado warning skill. *Bull. Amer. Meteor. Soc.*, **102**, E738–E747, https://doi.org/10.1175/BAMS-D-19-0310.1.

Brooks, H. E., 2004: Tornado-warning performance in the past and future: A perspective from signal detection theory. *Bull. Amer. Meteor. Soc.*, **85**, 837–844, https://doi.org/10.1175/BAMS-85-6-837.

——, and J. Correia Jr., 2018: Long-term performance metrics for National Weather Service tornado warnings. *Wea. Forecasting*, **33**, 1501–1511, https://doi.org/10.1175/WAF-D-18-0120.1.

Hales, J. E., Jr., 1989: The crucial role of tornado watches in the issuance of warnings for significant tornadoes. *Natl. Wea. Dig.*, **15**, 30–36.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534, https://doi.org/10.1175/WAF-D-12-00113.1.

Keene, K. M., P. T. Schlatter, J. E. Hales, and H. Brooks, 2008: Evaluation of NWS watch and warning performance related to tornadic events. *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., 3.19, https://ams.confex.com/ams/pdfpapers/142183.pdf.

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.

Raghavan, V., P. Bollmann, and G. S. Jung, 1989: A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Info. Syst.*, **7**, 205–229, https://doi.org/10.1145/65943.65945.

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

Rothfusz, L., C. Karstens, and D. Hilderband, 2014: Next-generation severe weather forecasting and communication. *Eos, Trans. Amer. Geophys. Union*,, **95**, 325–326, https://doi.org/10.1002/2014EO360001

——, R. Schneider, D. Novak, K. Klockow-McClain, A. E. Gerard, C. Karstens, G. J. Stumpf, and T. M. Smith, 2018: FACETs: A proposed next-generation paradigm for high-impact weather forecasting. *Bull. Amer. Meteor. Soc.*, **99**, 2025–2043, https://doi.org/10.1175/BAMS-D-16-0100.1.