Obermeier, H. M. and M. R. Anderson, 2014: Verification and analysis of impact-based tornado warnings in the Central Region of the National Weather Service. *Electronic J. Severe Storms Meteor.*, **10** (1), 1–20.

Electronic Journal of
SEVERE STORMS METEOROLOGY

Verification and Analysis of Impact-Based Tornado Warnings in the Central Region of the National Weather Service

HOLLY McCARTHY OBERMEIER
Dept. of Earth & Atmospheric Sciences, University of Nebraska, Lincoln, Nebraska

MARK R. ANDERSON
Dept. of Earth & Atmospheric Sciences, University of Nebraska, Lincoln, Nebraska

(Submitted 11 September 2014; in final form 10 September 2015)

ABSTRACT

The Joplin, Missouri EF5 tornado event on 22 May 2011 prompted the Central Region of the National Weather Service (NWS) to re-evaluate the current tornado warning format and implement the impact-based tornado warning (IBTW) experiment. IBTWs consist of tiers including damage tags and impact wording that convey increasing levels of damage. The damage wording within an IBTW is shown to relate to the Enhanced Fujita (EF) scale. Wording included in non-tagged IBTWs corresponds to EF0–EF2 tornado damage, while the damage wording for tagged IBTWs corresponds to EF3–EF5 tornado damage. This study investigates the accuracy of IBTWs by examining if a tornado occurs during the warning time frame, and whether the resulting damage matches the damage wording in the IBTW. All IBTWs from 1 April 2013 through 30 November 2013 were collected, as well as tornado survey information. Using these survey data, IBTWs were verified by the intensity of the tornado, if one occurred. Probability of detection (POD) and false alarm ratio (FAR) statistics are calculated through 2×2 contingency tables for both non-tagged and tagged IBTWs. Results indicate that the majority of both non-tagged and tagged IBTWs are false alarms, and tagged IBTWs have a very low POD. Other studies have shown that limitations in current technology and scientific knowledge may contribute to false alarms and missed detections. Case studies are examined to analyze whether these limitations play role in the use of IBTWs.

1. Introduction

Recent events such as the Joplin, MO tornado on 22 May 2011, which killed 158 people, have prompted an effort to restructure the existing National Weather Service (NWS) tornado warning format. Before this event, no single tornado had resulted in more than 100 deaths since 1953 (NWS 2011). An NWS assessment (NWS 2011) conducted after this event determined that a majority of Joplin residents did not fully perceive the danger upon reception of the tornado warning, and therefore did not take protective action. To combat this behavior in the future, the report suggested the initiation of warnings which are more "impact-based rather than phenomenon-based" while "diminishing the perception of false alarms and their impacts on credibility" (NWS 2011). In addition, the assessment proposed a tornado warning structure consisting of tiers. Impact-based tornado warnings (IBTWs) were introduced in 2012 and are a tiered system of warnings which employ the use of tornado damage tags (Table 1), along with corresponding damage-related wording (NWS 2014a). Warning forecasters are to include damage tags in IBTWs as confidence in the occurrence of a tornado and damage increases. Three tiers of tags exist: non-tagged, considerable and catastrophic.

Corresponding author address: Holly M. Obermeier, Dept. of Earth and Atmospheric Sciences, University of Nebraska, 214 Bessey Hall, Lincoln, NE 68588.
E-mail: holly.obermeier@huskers.unl.edu

Table 1: The three 2013 IBTW tiers—no tag, considerable tag, and catastrophic tag. Each tier has corresponding damage-related wording.

IBTW Tier	Impact Wording
No Tag...EF0–EF2 Tornadoes	Mobile homes will be damaged or destroyed. Significant damage to roofs...windows and vehicles will occur. Flying debris will be deadly to people and animals. Extensive tree damage is likely.
Considerable Tag...EF3–EF5 Tornadoes	You are in a life threatening situation. Mobile homes will be destroyed. Considerable damage to homes...businesses and vehicles is likely and complete destruction is possible. Flying debris will be deadly to people and animals.
Catastrophic Tag...EF4–EF5 Tornadoes	You could be killed if not underground or in a tornado shelter. Complete destruction of neighborhoods...businesses and vehicles will occur. Flying debris will be deadly to people and animals.

IBTWs were initially used experimentally in 2012 by five weather forecast offices (WFOs) in the NWS Central Region. In 2013, the IBTW experiment was expanded to encompass the entire Central Region (for map, see www.crh.noaa.gov/climate/main.php?type=resources&page=local_contacts.php). The Central Region is comprised of 38 WFOs, each of which is responsible for issuing severe weather warnings, including IBTWs, for its geographic area. An intended outcome of the IBTW experiment is an evaluation of forecasters' ability to distinguish between high- and low-impact events (NWS 2014a). Although the NWS states that an IBTW is not meant to address tornado intensity, the tag and associated damage wording within an IBTW become stronger with each tier.

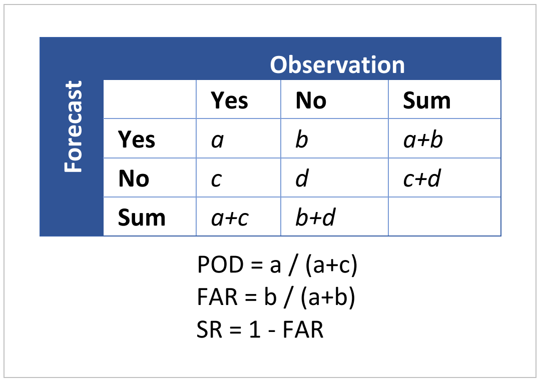
The different levels of damage wording are generally related to the Enhanced Fujita (EF) scale (WSEC 2006). Within each tier of damage wording, individual damage identifiers (DIs) and an associated degree of damage (DoD) can be identified. A DI is a type of infrastructure as outlined in the Enhanced Fujita (EF) scale, such as a single-family residence or a mobile home. The DoD is a numerical value which corresponds to a specific level of damage incurred by a tornado to a DI. As damage increases, so does the DoD number (see WSEC 2006 for specific details). For example, the sentence, "Complete destruction of neighborhoods, businesses and vehicles will occur", is included in the impact wording for the catastrophic damage tag. In this sentence, "neighborhoods" (assumed to be comprised of mainly single family homes, apartments and schools) and "businesses" (assumed to be comprised of retail buildings, professional buildings and shopping malls) are the DIs. Vehicles are not currently part of the EF scale and therefore were not considered. According to the EF scale, the DoD that best describes a state of complete destruction for each of these DIs correlates to EF4–EF5 damage. This method was performed on every sentence of non-tagged, considerable, and catastrophic impact wording. Certainly there is some subjectivity involved in defining words such as "significant" and "considerable", however, the EF scale was the best guide in attempting to lessen this subjectivity. For example, according to the EF scale, "significant" typically defines when 20% of a DI (such as windows or a roof) has experienced damage.

Analysis revealed that the impact wording in non-tagged IBTWs corresponds to EF0–EF2 damage. Impact wording included in considerable tagged IBTWs matched EF3–EF5 damage, and impact wording included in catastrophic tagged IBTWs matched EF4–EF5 damage. For this reason, non-tagged IBTWs are verified by EF0–EF2 damage. Considerable tagged IBTWs are verified by EF3–EF5 damage, and catastrophic tagged IBTWs are verified by EF4–EF5 damage. Therefore, the inclusion of a damage tag in a warning should be reserved for strong and violent tornadoes, capable of producing considerable or mass destruction.

Since tornadoes pose such a great risk to human life and property, it is vital that IBTWs perform optimally, with high probability of detection (POD) and low false alarm ratio (FAR). POD and FAR are calculated through the use of a 2×2 contingency table (Table 2).

Improving FAR while still maintaining or improving POD is difficult, considering the relationship between the two quantities. Fewer warnings could be issued in order to decrease FAR, however this would also lead to a decrease in POD (Brooks 2004a). POD does not typically improve with constant FAR unless advancements in technology or additional understanding of tornadogenesis occurs (Brooks 2004a; Wurman et al. 2012).

Table 2: A 2×2 contingency table used to calculate probability of detection (POD), false alarm ratio (FAR), and success rate (SR) (Doswell et al. 1990).



		Observation		
Forecast		Yes	No	Sum
	Yes	*a*	*b*	*a+b*
	No	*c*	*d*	*c+d*
	Sum	*a+c*	*b+d*	

POD = a / (a+c)
FAR = b / (a+b)
SR = 1 - FAR

The installation of the WSR-88D radar network in the early 1990s led to a significant improvement in POD (Polger et al. 1994; Simmons and Sutter 2009). However, the near-surface processes that lead to tornadogenesis typically occur over short time scales, which could be missed between radar scans. Even the best mesonets are not dense enough to provide the temporal or spatial information about a storm's surface environment which may lead to rapid tornadogenesis. The 22 May 2011 Joplin, MO tornado formed and moved through the city so quickly, warning forecasters were initially unaware (NWS 2011). A tornado warning was in effect 19 min before the tornado hit Joplin, however, due to the quick formation of the tornado, forecasters did not issue an updated severe weather statement (SVS) with a tornado emergency for the city. Other reasons for missed detections include limited spotter networks (and therefore limited tornado reports) and data overload on the warning forecaster (Brotzge and Donner 2013).

Tornado verification can be a challenging process. The Central Region of the NWS encompasses much of the geographical central and northern Great Plains, as well as the Midwest and parts of the Great Lakes region. Terrain and population density vary greatly. Population is quite sparse over parts of the Great Plains states and the High Plains of Colorado and Wyoming. Brotzge et al. (2011) indicates that FAR typically increases in sparsely populated areas, which are often located farther from a radar site as well. When tornadoes occur in these locations, they infrequently hit infrastructure, making verification and assignment of an EF rating difficult. Many of these tornadoes may be underrated. This may have implications for the verification of damage tags. Some tornadoes that are underrated may be considered a miss if they were covered by a tagged warning. However, the opposite may also be true; circumstances in which a tornado is underrated and not tagged in the corresponding warning actually may inflate the verification statistics.

Hypothetically, IBTWs will have similar POD and FAR as the current format of tornado warnings used by the NWS (traditional tornado warnings). Nationally, traditional tornado warnings have an FAR of 76% and a POD of 70% (NWS 2011). POD is even higher for traditional tornado warnings issued for EF3–EF5 tornado events (Table 3), typically over 90% (NWS 2011; Brotzge et al. 2013). This study, however, verifies the damage tags included in IBTWs. The results of this study cannot necessarily be directly compared to results of others, although they do provide some meaningful background statistics. This study will also explore when and where tags were most often issued across the Central Region. To gain additional understanding of the IBTW process, specific tornado events are examined along with WSR-88D radar data to hypothesize what may contribute to the successful or unsuccessful use of IBTWs.

Table 3: NWS tornado warning verification statistics from 1 October 2007 to 1 April 2011 for the United States (NWS 2011).

Event	POD	FAR
All Tornado	70%	76%
EF0-EF1	68%	NA
EF2-EF5	84%	NA
EF3-EF5	94%	NA

Table 4: A detailed description of both the non-tagged and tagged IBTW 2×2 contingency tables.

A		Observation		
Forecast		Yes	No	Sum
	Yes	*a*	*b*	*a+b*
	No	*c*	*d*	*c+d*
	Sum	*a+c*	*b+d*	

a = No tag is included in the TOR or SVS, and an EF0, EF1, or EF2 tornado does occur.

b = No tag is included and no tornado occurs, or an EF3, EF4, or EF5 tornado occurs

c = No TOR is issued, or a TOR or SVS is issued and a tag is included, and an EF0, EF1, or EF2 tornado occurs

d = Not calculated

B		Observation		
Forecast		Yes	No	Sum
	Yes	*a*	*b*	*a+b*
	No	*c*	*d*	*c+d*
	Sum	*a+c*	*b+d*	

a = A tag is included in a TOR or SVS, and EF3, EF4, or EF5 tornado does occur

b = A tag is included in a TOR or SVS, and no tornado occurs, or an EF0, EF1, or EF2 occurs

c = No TOR is issued or a TOR or SVS is issued without a tag, and EF3, EF4 or EF5 occurs

d = Not calculated

2. Data and methods

All tornado warnings with IBTW statements (TORs) and subsequent IBTW SVSs issued in the Central Region from 1 April 2013 to 30 November 2013 were gathered from the NWS Performance Management Branch (NWS 2014b). An SVS is a continuance or update to the original TOR issuance. These updates often include the most recent information about whether a tornado is radar-indicated or observed, and can include upgrades or downgrades in tornado damage tags. Data from 2012 were not included in this study, since IBTWs were only used in five WFOs, representing a small and relatively non-diverse geographic area of the Central Region. In addition, the impact wording changed several times from 2012 to 2013.

Typically an initial TOR and the following SVS(s) are grouped and verified as a single event. Yet, considering upgrades or downgrades in damage tags can occur, each TOR and SVS was verified individually in this study. Only SVSs that were continuances (CON) were kept; cancellation (CAN) and expiration (EXP) SVSs were eliminated. Also, the Omaha/Valley, NE (OAX) WFO did not participate in the 2013 IBTW experiment, choosing not to issue tags in any tornado warnings. Therefore, data from OAX were not used in this study.

The tornado dataset used for verification was gathered from NWS *Storm Data*, available from the National Climatic Data Center (NCDC 2014a). This information is also available from the NWS Performance Management website (NWS 2014b). *Storm Data* contains data about all tornadoes, including EF scale rating, path length, duration, and resulting damage. To account for all tornadoes which occurred in the Central Region during this study, each tornado was identified one-by-one in the *Storm Data* publication. If no TOR was in effect or issued during the life of the tornado, the tornado is considered unwarned. All unwarned tornadoes and their intensities are documented, as this information is necessary for calculating POD.

Statistical information such as POD, FAR and success rate (SR) are calculated through two separate contingency tables, one for non-tagged TORs and SVSs and the other for tagged TORs and SVSs (Table 4). Considerable and catastrophic tags were grouped together, mainly because a low number of catastrophic tags were issued and an additional contingency table for these tags does not lead to meaningful statistics. The method of verification in this study varies somewhat from past tornado-warning verification studies. Before the launch of IBTWs, a tornado warning would verify depending on the occurrence of a tornado, regardless of the rating. In this manner, verification is a relatively simple binary result. However, to verify IBTWs and corresponding tags correctly, the rating of any tornado that occurred must be used. This results in positively verified warnings, over-warnings and under-warnings. FAR consists of both over-warnings and under-warnings when calculated for non-tagged TORs and SVSs. FAR consists of only under-warnings when calculated for tagged TORs and SVSs. In this way, an IBTW still can be considered a false alarm if weaker or stronger tornado damage occurred than what was expected.

Consider a TOR or SVS that has been issued containing the first tier of damage wording (non-tagged). If an EF0–EF2 tornado occurred during the TOR or SVS, then the TOR or SVS verifies. If an EF3–EF5 occurred, then the TOR or SVS is considered an under-warning and does not verify. If no tornado occurred at all, the TOR or SVS is considered an over-warning and also does not verify. The only exception happens if an EF0–EF2 tornado did not occur during a TOR or SVS, yet does occur during a subsequent SVS. In this case, any proceeding TOR or SVS still verifies. This way, lead time does not penalize the overall statistics of the warning. This is true only if a tag has not been added to a subsequent SVS. If a tag is added, any proceeding TOR or SVS does not verify.

Consider a TOR or SVS containing a considerable or catastrophic tag, which would include the second or third tier of damage wording. If an EF3–EF5 tornado occurred during the TOR or SVS, then the TOR or SVS verified. If an EF0–EF2 tornado occurred, then the TOR or SVS was considered an over-warning and did not verify. If no tornado occurred at all, the TOR or SVS was considered an over-warning and did not verify. The exception happened if an EF3–EF5 did not occur during a TOR or SVS, yet occurred during a subsequent SVS. In this case, the previous TOR or SVSs still verified. This is true only if a tag has not been removed from the subsequent SVS. If a tag was removed, any proceeding TOR or SVS(s) did not verify.

In this study, statistics were also calculated as a function of the Storm Prediction Center (SPC) “day-1” convective outlook. These outlooks set the level of awareness for warning operations throughout the day. Each IBTW TOR and SVS was binned according to the 2013 categorical risk area for which it was issued: slight, moderate and high. Using the method laid out above, contingency tables were then calculated for non-tagged and tagged IBTW TORs and SVSs to determine POD and FAR. In addition, IBTWs were verified and statistics were calculated according to the maximum tag issued within the entire timeframe of the warning (grouping TORs and subsequent SVSs as a single event). The warnings were still verified by tornado rating. This method did not account for changes in tags between a TOR and later SVSs, but may help to mitigate any biases introduced by the fact that a tornado did not necessarily maintain maximum intensity during its entire lifetime.

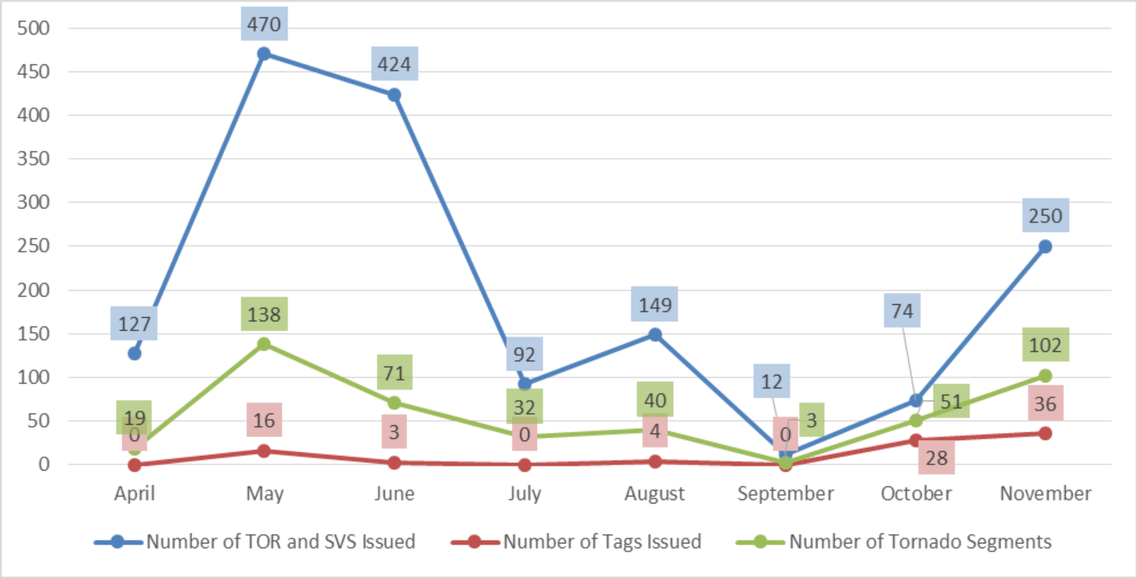


Figure 1: Total number of TORs and SVSs (blue), total number of tags, both considerable and catastrophic (red), as well as total number of tornado segments for each month during the study (green).

3. Results

a. Use of damage tags

The IBTW dataset for the dates of 1 April 2013 through 30 November 2013 totaled 702 TORs and 896 SVSs, for a total of 1598 statements issued by the NWS Central Region WFOs (Fig. 1). The majority of TORs and related SVSs were issued during the spring months; however, 2013 also featured two autumn severe-weather events. Of the 1598 TORs and SVSs issued during this study, 84 contained the considerable tag and three contained the catastrophic tag. These 87 tagged TORs and SVSs account for approximately 5% of the total. The two autumn tornado events (4–5 October 2013 and 17 November 2013) resulted in 74% of the tags during this study. The remaining 26% of the tags were issued during May, June and August. In May and June, 209 tornado segments occurred, with 19 tags issued (Fig.1). In October and November, 153 tornado segments occurred, with 64 tags. Tags were issued on eight dates (Fig. 2), each having an EF3 or greater tornado, although a tagged warning did not always correspond to an EF3 tornado. The 1200 UTC SPC day-1 convective outlook included a slight risk on four of the eight days, a moderate risk on three, and a high risk on one. The high-risk day corresponded to the Midwest tornadoes of 17 November 2013, with the greatest one-day total number of tags. An EF3 or greater tornado occurred in the Central Region on only one day (2 June 2013), and no tagged warning was issued.

Every participating office in the Central Region issued at least one IBTW during this study, with the exception of the Grand Junction, CO (GJT) WFO. Of these 37 offices, 11 issued at least one TOR and SVS that included a tag (Fig. 3). Again, data from OAX were not included in the results of this study. The Paducah, KY (PAH) and Sioux Falls, SD (FSD) WFOs issued the majority of the tags, for a combined 50 out of 87 (59.5%). All tags issued by the FSD WFO occurred in northeastern Nebraska, southeastern South Dakota and western Iowa on 4–5 Oct 2013. The PAH WFO issued all tags during the 17 November 2013 tornadoes. Tornadoes with EF3–EF4 damage occurred during this study in the coverage areas of three WFOs that did not issue tags [Des Moines, IA (DMX), Dodge City, KS (DDC) and North Webster, IN (IWX)]. No EF5 tornadoes occurred in the Central Region in 2013.

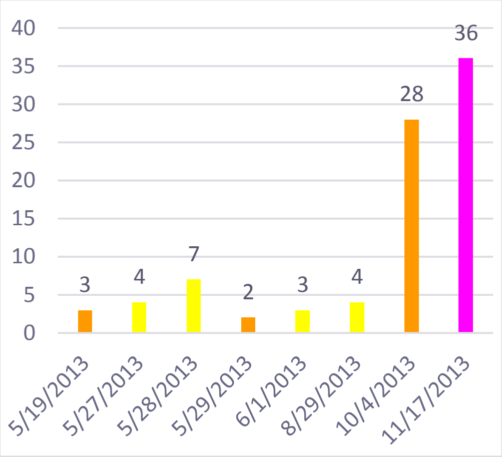


Figure 2: Dates on which tags were issued (x-axis), and the total number of tags issued (y-axis). The color of the bar corresponds to the SPC day-1 convective outlook risk; yellow for slight, orange for moderate and pink for high.

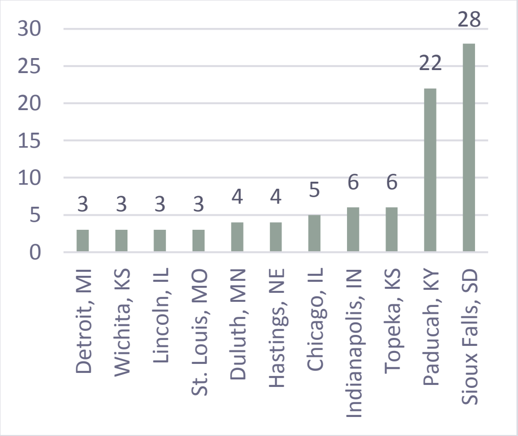


Figure 3: NWS WFOs that issued tags during this study, and the number of tagged TORs and SVSs (y-axis) issued by each.

Tags were generally issued in an SVS and not in the initial TOR (Fig. 4a). Detection of a tornado through reports of damage or visual observation after the initial TOR could prompt a tag in an SVS. Of the 22 tags issued in TOR statements, seven verified for the occurrence of an EF3–EF4 tornado, a 32% success rate (Table 5a). Interestingly, the 65 tags issued in an SVS statement verified less often, with a 25% success rate. When analyzed in combination, the majority of tags were issued in SVS statements with visual observation of a tornado (Fig. 4b). A smaller number of tags were included in SVS statements with radar indication of a tornado.

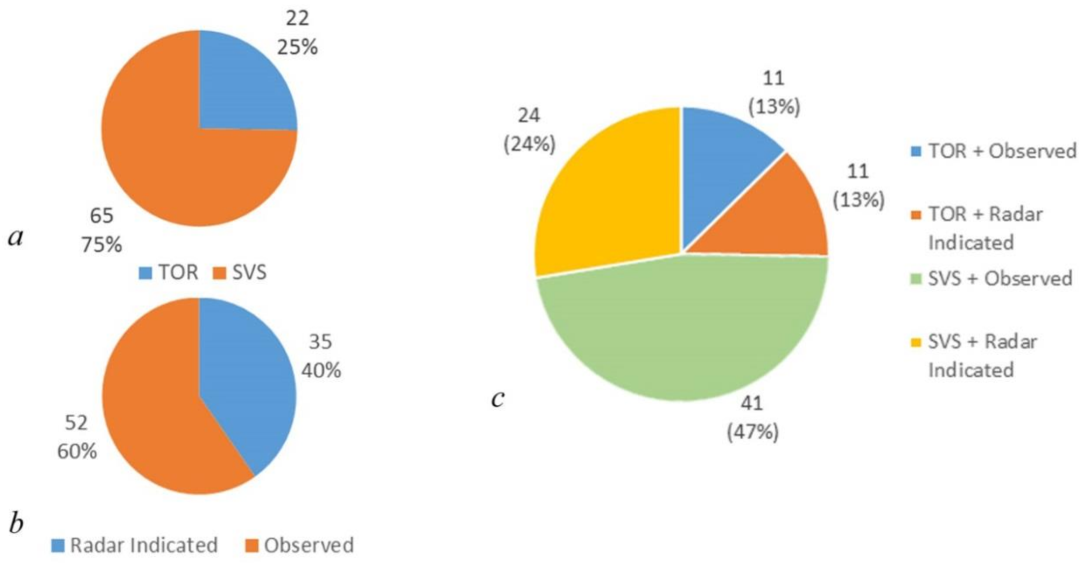


Figure 4: a) The number of tags issued in a TOR versus an SVS. b) The number of tags issued in a TOR or SVS in which a tornado was visually observed versus radar-indicated. c) The number of tags which were included in the combinations of TOR and observed tornado, TOR and radar-indicated tornado, SVS and observed tornado, and SVS and radar indicated tornado.

Table 5: a) The number of considerable or catastrophic tags issued in an IBTW TOR or SVS statement, along with the success rate for the tags according to the statement type, which are verified by the occurrence of an EF3–EF4 tornado. b) The number of considerable or catastrophic tags issued in an IBTW for which the tornado status was radar-indicates or observed, along with the success rate for the tags according to the tornado status, which are verified by the occurrence of an EF3–EF4 tornado.

a) IBTW Statement	Considerable Tag	Catastrophic Tag	Total	Percent of Total	Verify for EF3–EF4 Tornado?	Success Rate
TOR	22	0	22	25%	7	32%
SVS	62	3	65	75%	16	25%
Total	84	3	87	100%	23	26%
b) Tornado Status	Considerable Tag	Catastrophic Tag	Total	Percent of Total	Verify for EF3–EF4 Tornado?	Success Rate
Radar-Indicated	35	0	35	40%	10	29%
Observed	49	3	52	60%	13	25%
Total	84	3	87	100%	23	26%

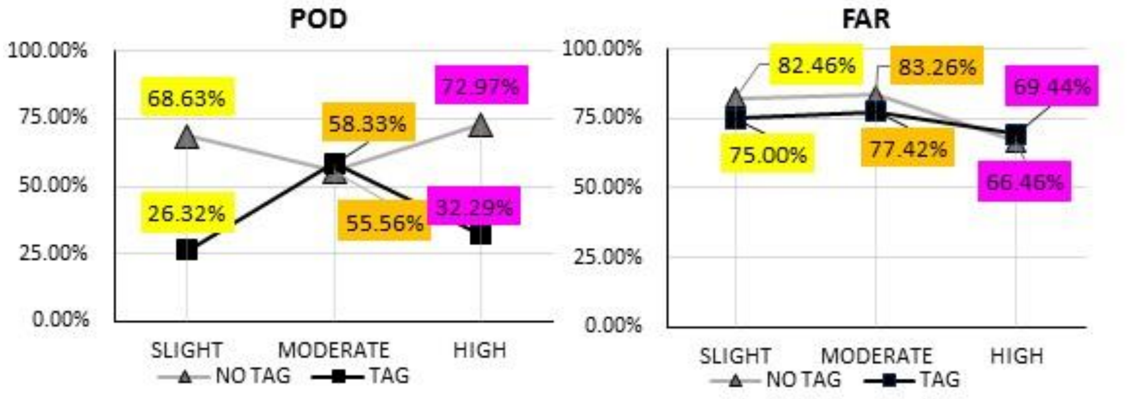


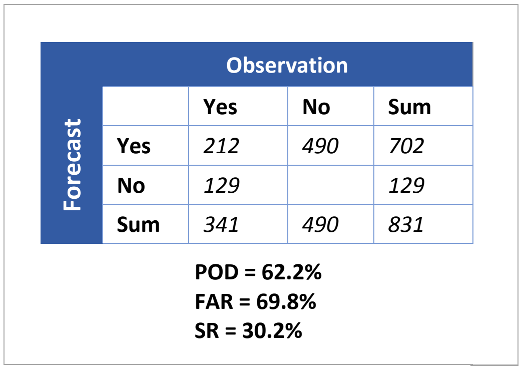
Figure 5: POD (left) and FAR (right) calculated as a function of SPC day one convective outlook. Values for slight risk days are in yellow, moderate in orange and high in pink.

Tornado segments that occurred during a tagged IBTW in this study had a longer average duration and path length when compared to tornado segments which occurred during a non-tagged IBTW. The average lifespan of a tornado segment which was covered by a tagged warning was 10.66 min, and the average path length was 9.4 km (5.82 mi). The average lifespan of a tornado which was covered by a non-tagged warning was 6.56 min, and the average path length was 6.0 km (3.74 mi). Since path length is positively related to damage rating (Brooks 2004b), forecasters may have been more likely to detect a tornado and then tag the warning. Also, tags were used most often when a tornado was observed, rather than radar-indicated (Fig. 4c). This finding seems logical, considering the report of a confirmed tornado likely increased confidence for issuing a tag. However, tornado-observed tagged TORs and SVSs verified slightly less often than radar-indicated tagged TORs and SVSs (Table 5b). This suggests that the tornadic evidence (observation versus radar-indicated) by which a tag was issued does not necessarily facilitate the ability to estimate tornado intensity. Tornado observations come from a variety of sources, ranging from the public, to law enforcement, to trained weather spotters. Some of these sources are deemed more credible than others, and a report of a tornado did not always result in tagged IBTWs which verified. In some cases, this was perhaps due to erroneous reports; other times the tornado dissipated before causing any damage. Sometimes, a tornado occurred, but was less damaging than anticipated.

b. POD and FAR calculations

Initially, IBTWs in this study were verified in the same manner as traditional tornado warnings. Using this method, a TOR and the following SVS(s) were grouped and verified as a single event. Each IBTW was verified by the occurrence of a tornado, regardless of tornado intensity or IBTW tags. This method yielded statistics by which to directly compare to the national averages (Table 3). For the Central Region during this study, FAR was found to be nearly 70%, which is about 6% lower than the national average (Table 6). POD was 62%, about 8% lower than the national average. POD also was calculated with regard to tornado rating (Table 7). Similar to the national statistics, POD increases with increasing tornado-damage strength. In fact, no EF3 or EF4 tornado occurred without warning in the Central Region during this study, leading to a POD of 100% for EF3 or greater tornadoes. This increase in POD with increasing tornado rating is important, because it indicates that more substantial tornadoes were more typically accompanied by a warning at some point during their lifetime, regardless of whether the warning was issued with lead time. Considering the traditional POD for EF3–EF5 tornadoes is higher than the traditional POD for weaker tornadoes, it might be expected that the POD concerning tagged TORs and SVSs would be higher than non-tagged TORs and SVSs.

Table 6: Contingency table for IBTWs using the traditional method.



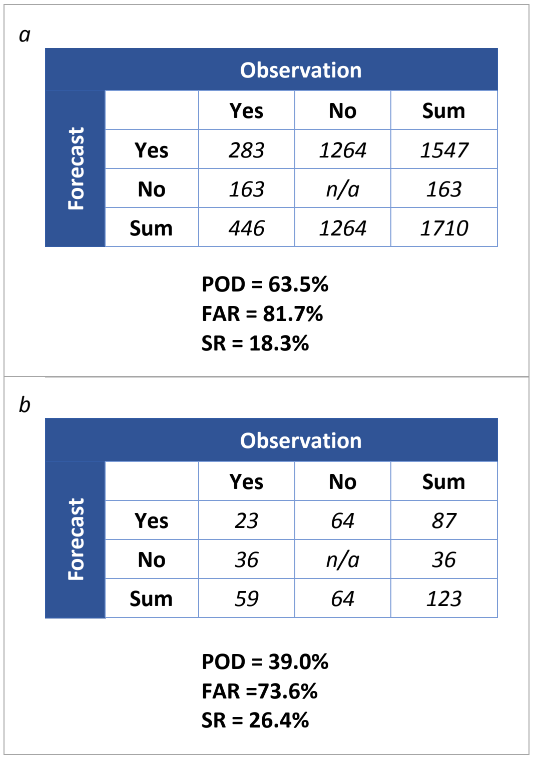
		Observation		
		Yes	No	Sum
Forecast	Yes	*212*	*490*	*702*
	No	*129*		*129*
	Sum	*341*	*490*	*831*

POD = 62.2%
FAR = 69.8%
SR = 30.2%

Table 7: Central Region IBTW verification statistics found using the traditional method.

Event	POD	FAR
All Tornado	62%	70%
EF0-EF1	55%	NA
EF2-EF5	92%	NA
EF3-EF5	100%	NA

Table 8: a) Contingency table for non-tagged IBTW. b) Contingency table for tagged IBTW.



a

		Observation		
		Yes	No	Sum
Forecast	Yes	*283*	*1264*	*1547*
	No	*163*	*n/a*	*163*
	Sum	*446*	*1264*	*1710*

POD = 63.5%
FAR = 81.7%
SR = 18.3%

b

		Observation		
		Yes	No	Sum
Forecast	Yes	*23*	*64*	*87*
	No	*36*	*n/a*	*36*
	Sum	*59*	*64*	*123*

POD = 39.0%
FAR =73.6%
SR = 26.4%

Further statistical analysis through contingency tables evaluated the performance of non-tagged and tagged IBTWs (Table 8) as verified by tornado rating, and by treating TOR and SVSs individually. POD for both non-tagged and tagged TORs and SVSs was lower than the traditional numbers, and POD for tagged TORs and SVSs was much lower than POD for non-tagged TORs and SVSs. Non-tagged TORs and SVSs (corresponding to EF0–EF2 tornadoes) had a POD of 64%, while tagged TORs and SVSs (corresponding to EF3–EF5 tornadoes) had a POD of 39%.

During this study 129 unwarned EF0–EF2 tornado events occurred. In addition, 32 over-warned TORs and SVSs with an EF0–EF2 tornado occurred, meaning a tag was issued when it was not warranted. There were 36 under-warned TORs and SVSs in which an EF3–EF4 tornado occurred (again, no EF5 tornado occurred in this study), and no considerable or catastrophic tag was included.

FAR values were slightly lower for tagged TORs and SVSs and higher for non-tagged TORs and SVSs. Non-tagged TORs and SVSs had a FAR of 82%, 6% higher than the traditional FAR. The total number of false alarms in the non-tagged IBTWs category was almost entirely comprised of TORs and SVSs in which no EF0–EF2 tornado event occurred (over-warnings). The rest of the false alarms accounted for non-tagged TORs and SVSs during which EF3–EF4 tornado events occurred (under-warnings).

Tagged IBTW TORs and SVSs had a FAR of 74%, which is slightly lower than the traditional tornado-warning FAR. All false alarms were over-warned events with either no tornado or a one rated less than EF3. A tornado of any strength occurred during 60 of the 87 tagged TOR or SVS. If the tags were verified by the occurrence of a tornado regardless of damage, the FAR would be 31%. The low FAR suggests that tagged TORs and SVSs were issued most often with evidence (for example, strong rotational radar velocity signature) that increased forecaster confidence that a tornado would occur or was occurring. The fact that tornadoes often were occurring during tagged TORs and SVSs is encouraging, because it indicates the warning forecaster's ability to realize situations in which a tornado was likely. However, there seemed to be less ability to distinguish whether a very strong or violent tornado was likely or having occurred, considering the high FAR of tagged TORs and SVSs verified by tornado rating.

As mentioned, the majority of damage tags were issued in IBTWs during two separate tornado outbreaks, on 4–5 October and 17 November 2013. To determine if statistics were improved during these outbreak situations, the IBTWs from these events were grouped and considered separately. POD and FAR were calculated and compared. For non-tagged TORs and SVSs, POD was similar to that for the entire dataset, at 65%. FAR was approximately 74%, 7% more than the entire dataset. For tagged TORs and SVSs, POD is near 49%, while FAR was 72%. These numbers indicate that POD was about 10% better on the "outbreak" days, while the FAR did not change greatly.

The improvement in POD could be due to a number of reasons. Perhaps heightened forecaster situational awareness played a role. However, outbreak days are typically associated with longer tornado path length (Doswell et al. 2006) and longer tornado path length is positively correlated with higher rating (Brooks 2004b). The POD improvement may have been more attributable to higher statistical likelihood that a forecaster will warn and tag a tornado occurring on an outbreak day.

Verification statistics were also calculated as a function of the 12Z SPC day-1 convective outlook. Figure 5 indicates POD was lower for tagged IBTWs on high-risk days, but greater for non-tagged IBTWs. FAR improved for both tagged and non-tagged IBTWs on high-risk days as compared to slight- and moderate-risk days.

Two additional contingency tables were calculated after verifying IBTWs according to the highest tag issued within the entire warning, as described in the section 2. As compared to the initial method (treating TORs and SVSs separately), non-tagged IBTWs had a POD around 53% (10% lower) and a FAR around 75% (7% lower). Tagged IBTWs had a POD around 45% (6% higher) and a FAR around 68% (3% higher). These PODs were still much lower than the national statistics for traditional tornado warnings.

c. Case studies

Collection of WSR-88D radar imagery and construction of event timelines led to additional insight about the success of damage tags. A successful use of damage tags occurred on 17 November 2013 in two IBTWs issued for a supercell in southeastern Missouri. In this case, warning forecasters at the PAH WFO issued four considerable damage tags, all of which verified by the occurrence of an EF3 tornado. Warning forecasters correctly indicated the potential for EF3 or stronger impacts, and correctly made the choice to issue the considerable damage tags. A factor that led to this decision likely included the WSR-88D radial velocity imagery which indicated robust, well-organized rotation within the storm (NCDC 2014b). Also, severe weather was anticipated on this day, per the moderate risk for this region in the SPC day-1 convective outlook (SPC 2013). Given the elevated situational awareness, warning forecasters were aware of the potential for strong tornadoes. However, more in-depth analysis of the warnings issued by PAH on this date showed that the FAR of tagged IBTWs (TORs and SVSs verified by EF3 or greater tornadoes) was 64%. Forecasters used tags correctly in this particular example, but tag use was not consistently successful during the tornado outbreak.

Also examined were the situations containing the catastrophic damage tag. This third tier of damage wording should be used only on rare occasions when EF4–EF5 impacts are expected. The catastrophic damage tag was used in three SVSs during this study, each of which resulted in false alarm for the occurrence of a violent tornado. A catastrophic damage tag was included in an SVS issued by the Wichita, KS WFO on 19 May 2013 as a supercell was approaching the city of Wichita. WSR-88D reflectivity and radial velocity imagery indicated a well-defined hook echo and robust circulation within a supercell southwest of Wichita (NCDC 2014b). The storm had a history of producing a few tornadoes that storm spotters reported just outside of the city limits. Considering the impressive radar signature, it seemed probable the storm would produce a violent tornado potentially tracking into Wichita, leading to the issuance of the catastrophic damage tag. However, the storm lost its well-defined circulation in a matter of one volume scan, and did not produce another tornado. In this example, the catastrophic tag was a false alarm, but also illustrates the uncertainty involved with issuing tags. Clearly, this could have been a very serious situation for the city of Wichita. This example showed how quickly a storm can change in intensity.

This study contains 36 incidents of EF3 or greater tornadoes with no IBTW tag. One such case occurred on 17 November 2013 in southern Illinois, in an SPC moderate risk. The first TOR was issued for the storm of interest, although other tornado warnings had been issued for nearby storms. This particular storm had no tornadic history. WSR-88D radial velocity imagery indicated robust circulation within the supercell, which produced an EF4 tornado that tracked into the town of New Minden, IL (NCDC 2014b). The tornado was warned, but not tagged. It is difficult to assess why no damage tag was issued at any time during the span of the IBTW. A survey is needed to understand the warning forecast decision-making process during this event, and is beyond the scope of this study.

Analysis of several IBTW examples revealed more details about the scenarios in which tags were or were not used during this study. Tags were occasionally used well, such as seen in the case study of the 17 November 2013 IBTW damage tags issued by the PAH WFO. However, many situations had damage tags with false alarms, specifically those with catastrophic tags. Swift changes in tornado character and/or thunderstorm intensity made successfully issuing tags more difficult. While radar evidence and spotter reports were critical tools used to make IBTW damage tag decisions, limitations in these tools could have an unfavorable impact on the POD and FAR of tagged IBTWs.

4. Conclusion

IBTWs are meant to convey expected impacts of tornadoes in a tiered structure through the use of damage tags. This study revealed that the majority of IBTWs are false alarms, and tagged IBTWs have a very low POD. Examination of specific events indicates that IBTWs occasionally can be used with success, although more often the tags result in false alarms for the occurrence of EF3 or greater tornadoes. The case study of the catastrophic tag issued in the IBTW for Wichita, KS on 19 May 2013 revealed a tornadic circulation that rapidly weakened and dissipated as the parent supercell passed over the city. In addition, when many EF3 or greater events occurred, no damage tag was included in the IBTW. However, every EF3–EF4 tornado during this study was warned. Despite the ability of warning forecasters to detect strong and violent tornadoes (based on their rating), forecasters did not often use tags. More study, including surveys of forecaster warning behavior, would have to be conducted to explore reasons behind the low IBTW POD. As the NWS expands the IBTW project into other NWS WFOs and regions over the coming years, additional studies should assess whether IBTWs continue to perform with similar results.

Advancements in radar technology or increased knowledge in regard to tornadogenesis may lead to improved IBTW verification statistics in the future. The introduction of dual-polarization to the WSR-88D radars may provide an avenue by which tornado intensity sometimes be estimated. Use of the tornadic debris signature, which incorporates dual polarimetric variables such as correlation coefficient, has been shown to relate to changes and trends in damage intensity during a tornado event (Bodine et al. 2013; Van Den Broeke and Jauernic 2014). This information, used operationally, may allow warning forecasters to issue tags with more success. The operational introduction of SAILS (Supplemental Adaptive Intra-Volume Low-Level Scan), a new volumetric scanning strategy for the WSR-88D, allows warning forecasters additional low-level scans with less elapsed time in between (NWS 2014c). The higher temporal resolution will provide forecasters more information that may be valuable in the IBTW process. Increases in the spatial resolution of radar data and greater coverage near the ground may only be achieved by the installation of more radars.

In the future, a higher-density radar network available for operational use by NWS forecasters may provide critical data leading to more successful use of tags. Furthermore, research resulting from the second installment of the VORTEX project, which operated from 2009–2010, will likely lead to even more understanding of tornadogenesis and tornado structure (Wurman et al. 2012). These increases in the knowledge of tornadogenesis, including why some supercells produce tornadoes and other do not, eventually could improve IBTW statistics.

ACKNOWLEDGMENTS

The authors wish to thank the reviewers who provided many helpful comments and suggestions, which streamlined this paper and added needed clarity.

REFERENCES

Bodine, D. J, M. R. Kumjian, R. D. Palmer, P. L. Heinselman and A. V. Ryzhkov, 2013: Tornado damage estimation using polarimetric radar. *Wea. Forecasting*, **28**, 139–158.

Brooks, H. E., 2004a: Tornado-warning performance in the past and future: A perspective from signal detection theory. *Bull. Amer. Meteor. Soc.*, **85**, 837–843.

———, 2004b: On the relationship of tornado path length and width to intensity. *Wea. Forecasting*, **19**, 310–319.

Brotzge, J. A., and W. Donner, 2013: The tornado warning process. *Bull. Amer. Meteor. Soc.*, **94**, 1715–1733.

———, S. Erickson, and H. E. Brooks, 2011: A 5-yr climatology of tornado false alarms. *Wea. Forecasting*, **26**, 534–544.

———, S. E. Nelson, R. L. Thompson, and 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.

Doswell, C. A. III, R. P. Davies-Jones and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576–585.

———, R. Edwards, R. L Thompson, J. A. Hart and K. C. Crosbie, 2006: A simple and flexible method for ranking severe weather events. *Wea. Forecasting*, **21**, 939–951.

NCDC, 2014a: *Storm Data*, Vol. 56. [Available online at http://www.ncdc.noaa.gov/IPS/sd/sd.html.]

———, 2014b: WSR-88D radar data archive. [Available online at http://www.ncdc.noaa.gov/nexradinv/.]

NWS, 2011. Central Region service assessment: Joplin, Missouri, tornado—May 22, 2011. [Available online at http://www.nws.noaa.gov/om/assessments/pdfs/Joplin_tornado.pdf.]

———, 2014a: Impact based warnings. [Available online at http://www.weather.gov/impacts.]

———, 2014b: Performance Management Branch. [Available online at https://verification.nws.noaa.gov/.]

———, 2014c: Multiple elevation scan option for SAILS: The next step in the continuing evolution of dynamic scanning. [Available at http://www.roc.noaa.gov/wsr88d/NewRadarTechnology/NewTechDefault.aspx.]

Polger, P. D., B. S. Goldsmith, R. C. Przywarty and J. R. Bocchieri, 1994: National Weather Service warning performance based on the WSR-88D. *Bull. Amer. Meteor. Soc.*, **75**, 203–214.

Simmons, K. M. and D. Sutter, 2005: WSR-88D Radar, Tornado Warnings and Tornado Casualties. *Wea. Forecasting*, **20**, 301–310.

———, and ———, 2009: False alarms, tornado warnings, and tornado casualties. *Wea. Climate Soc.*, **1**, 38–53.

Storm Prediction Center (SPC), 2013: Product archive. [Available online at http://www.spc.noaa.gov/archive/.]

Texas Tech University Wind Science and Engineering Center (WSEC), 2006: A recommendation for an enhanced Fujita scale. [Available online at www.depts.ttu.edu/nwi/Pubs/FScale/EFScale.pdf.]

Van Den Broeke, M. S., and S. T. Jauernic, 2014: Spatial and temporal characteristics of polarimetric tornadic debris signatures. *J. Appl. Meteor. Climatol.*, **53**, 2217–2231

Wurman, J., D. Dowell, Y. Richardson, P. Markowski, E. Rasmussen, D. Burgess, L. Wicker and H. B. Bluestein, 2012: The second Verification of the Origins of Rotation in Tornadoes Experiment (VORTEX 2). *Bull. Amer. Meteor. Soc.*, **93**, 1147–1170.

REVIEWER COMMENTS

[Authors' responses in *blue italics.*]

REVIEWER A (Patrick M. Marsh):

Initial Review:

Recommendation: Accept with major revisions.

Paper summary: This paper provides a cursory overview of the National Weather Service's Central Region's Impact Based Tornado Warning (IBTW) program. It starts with a brief overview of the IBTW program, including its origin. The paper then provides an overview of its data and methods, followed by the results. The results are broken down into the use of damage tags (e.g., when the damage tags were included, which offices issued them, and whether they were included in the initial warning or the subsequent severe weather statements), probability of detection and false alarm ratio, and case studies. The paper then ends by stating conclusions.

General overview: In general, the main component of this paper is the verification work on the NWS' CR IBTW program. Other than referring to FAR as false alarm rate, rather than the correct name of false alarm ratio, I do not have any major issues with this verification. My main issue with the paper is the overwhelming amount of speculation offered to explain/justify the results. For every line of speculation offered by the authors', I can offer counter-speculation that seems just as likely. It is my opinion that the authors need to remove most, if not all, of the speculation. In doing so, however, I believe the paper would be rather short to be considered as a full article. Thus, my recommendation is for the authors to remove the speculation and have the paper considered as a note.

[Editor's Note: Manuscripts can be reclassified in the EJSSM Online Journal System's metadata process at any point in the submission, review or editing process, at the discretion of the Editor, with consultation of the author(s).]

Substantive (major) comments: Abstract: "Limitations in current technology and scientific knowledge may contribute to false alarms and missed detections. These findings suggest that more advanced in technology and the understanding of tornadogenesis are necessary for more successful implementation of IBTWs."

This is overly speculative. There is nothing presented in this paper that justifies this statement. Furthermore, maybe combining environmental information with radar presentation could improve performance?

It's at least as plausible as what the authors have posited.

These statements are restructured in the abstract and within the paper to mitigate speculation, however, the authors don't consider the original statement to be complete speculation. Past studies (such as Brooks 2004, Polger et al 1994) have shown that improvements in POD are linked to improvements in technology (such as the installation of WSR-88Ds), conceptual models and forecasting methods...and that limitations in these, as well as limitations in mesonet and spotter networks, lead to missed detections (Brotzge & Donner). In addition, a primary reason for a false alarm is the rapid dissipation of a tornado (Brotzge et al. 2011), which can occur in a matter of a radar scan. A much more developed discussion of these limitations was added to the Introduction and the above references were added. Again, there were many missed EF3–EF4 events during this study, leading to very low POD for tagged IBTWs...and there were many tags issued which resulted in false alarms. The Wichita catastrophic tag case study was an example of current temporal radar limitations. From the case studies, it is a blend of the items listed above (in addition other items that this paper cannot address, such as forecaster behavior) that led to these missed events and false alarms...leading to the authors overall conclusion.

Section 2: "Yet, considering upgrades of downgrades in damage tags can occur, each TOR and SVS was verified individually in this study."

I understand why you wanted to do this (otherwise you have a very small sample), however, by doing this you bias your statistics. Follow-up SVSs are not independent of the initial warning; neither are subsequent warnings on the same thunderstorm. Using your approach, a long-track tornado is going to have multiple products issued for it over the life of the tornado. All of these products will be verified by a single tornado rating, despite the fact the tornado will not have maintained the same intensity for the tornado's duration. Thus some products will be scored as misses, despite being hits in reality.

Alternatively, if a small proportion of tornadoes result in a large proportion of damage tag products, the resulting verification will not be representative of the overall skill with discerning weak vs strong/violent tornadoes. There are several additional ways that this approach could bias your statistics, but I won't go into all of them. In fact, in section 3, the authors mention how nearly 60 percent of the damage tags used came from two offices, with one of those offices issuing all of their products during a single tornado outbreak.

At the very least, I would suggest that the authors also conduct the verification work utilizing only the highest tiered product each WFO issued for a given tornado. This would cut down on a lot of the biases alluded to. I would also suggest that the authors also present information regarding the number of tornadoes that occurred on each day.

The reason the authors chose to treat each TOR and SVS separately is because of the intentional changes in damage tags can occur from a TOR to the following SVS(s). The goal is verify the tags, not just the warning. A tornado could occur during an IBTW TOR/SVS, thus verifying the warning, but not necessarily the tag. Sometimes, a tag was included in a TOR, then dropped in the following SVS. There were also instances in which a tag was not included in the TOR, included in an SVS, then dropped in the following SVS, and then added again in an additional SVS. While we cannot say why this occurred or why the warning forecaster may have [chosen] to add or drop a tag, it does indicate an intentional change. This method accounts for these changes, and that is why the TORs/SVSs are treated separately. To clarify how the verification method, additional explanation for each of the contingency tables was added to the Data and Methods section.

As per your suggestion, the verification was also performed using the highest tag in each IBTW. Two contingency tables were still calculated, and results were added to the paper in the Results section.

Also, the number of tornado segments per month was added to a figure, in addition to the number of TORs/SVSs and tags per month.

[Minor comments omitted…]

Second Review:

Recommendation: Accept with major revision.

General overview: I still have issues with the portrayal of the verification statistics (related to the non-independent nature of the data). I am of the opinion that a single year of tornado warning data is insufficient to draw meaningful conclusions related to the success of the IBW "experiment". However, I recognize that if one were to try and draw meaningful conclusions from the single year of the experiment, that one would have to approach it in a manner similar to how the authors have done.

In response to reviewers' concerns in round one, the authors have added quite a bit of new text. Unfortunately, it is my opinion that a lot of the new text reduces the clarity of the paper. I found myself frequently having to re-read paragraphs in an attempt to grasp what the authors were trying to convey. I would strongly urge the authors to take another pass through the paper in an attempt to reduce the number of pronouns used, make sure that words have not been left out of sentences, and ensure a proper explanation of all terms is used. The need for a rewrite to improve readability and clarity is the main basis for my major revisions still needed recommendation.

The authors wish to thank the reviewer for the additional comments/corrections. As the reviewer noticed, the first round of reviews resulted in a good deal of additional text. A thorough evaluation of the paper resulted in restructuring of some sentences/paragraphs in order to improve clarity and readability. To reduce confusion, pronouns were eliminated or clarification was added in order to make sure any reader could easily follow the authors' meaning.

I am not sure that I agree with the authors' conclusions [on outbreak-day warnings]. Are forecasters actually better on outbreak days due to heightened situational awareness, or is there a tendency to over-warn? [The following] figure plots all the tornado warnings from 17 November 2013. It becomes readily apparent that the number of false alarms (no tornado reports at all) for tagged warnings is quite high for Paducah.

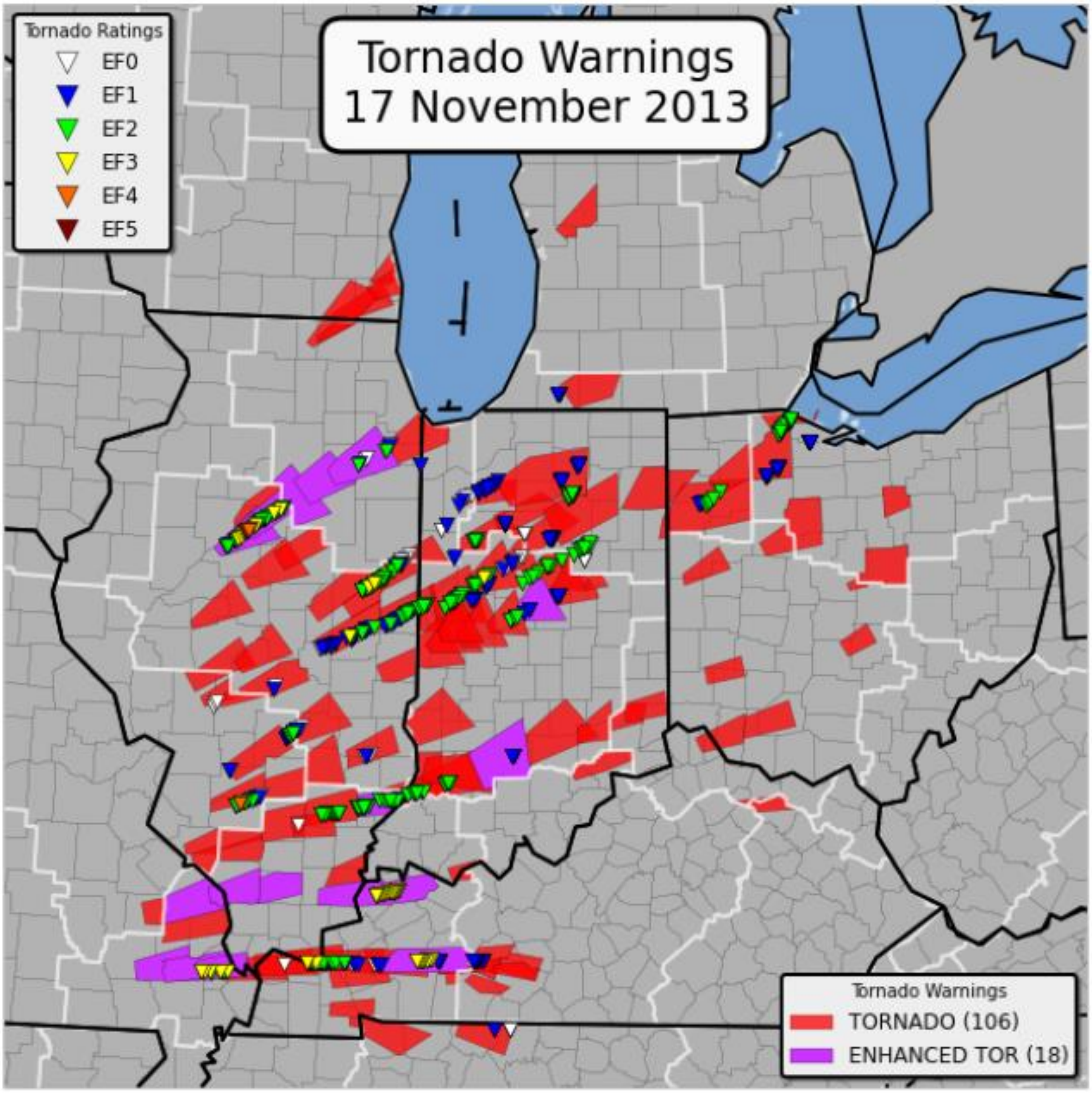


Figure R1: Tornado warnings for 17 November 2013. Report information was taken from the NWS Damage Assessment Tool in the days following the tornado outbreak.

From the figure that the reviewer provided, it's certainly obvious that some IBTWs with damage tags did not verify for a tornado of any intensity. However, there were several which did verify. The PAH WFO issued 12 IBTWs which contained a damage tag at some point. Out of these 12 warnings, a tornado of any intensity occurred during 7 of them (FAR 42%). Warnings issued on 4–5 Oct (another outbreak day) from WFO FSD were investigated also, and similar results were found (FAR 11%). The warnings themselves are obviously warranted...but POD and FAR statistics become far less favorable once the warnings are verified by intensity. So perhaps, over-warning isn't a problem in the sense that there was a warning, but that there was a tag. The authors added some text explaining that the successful use of tags in this case study wasn't necessarily the typical outcome for the day.

[Minor comments omitted...]

Third Review:

Recommendation: Accept.

General comment: I have no additional comments for this paper. I feel the authors have scaled back some of their claims and hedged their conclusions where appropriate.

There are a couple of places where the reading is tough to slog through, but I'm not really sure of a means of addressing that given the rote nature of the topic.

REVIEWER B (Jerald A. Brotzge):

Initial Review:

Recommendation: Accept with major revisions.

Overview: This manuscript analyzes results from a year-long experiment run by Central Region exploring the accuracy of issuing impact-based tornado warnings. Such an independent analysis is needed, and this paper does that effectively. While the general public may indeed react more urgently to the tagged warnings (as intended), this work addresses the more pertinent question of whether or not those warnings are accurate, and thus warranted. This paper is sorely needed, and I'm delighted to see it addressed.

The authors wish to thank the reviewer for the helpful comments, which have added more depth to the text. There are many ways to analyze the data, and the reviewer suggested many interesting points which are both relevant to this paper as well as to future research, such as tag use in regard to radar velocity signature (which will be important as the IBTW project is continued and expanded).

Major (substantive) comments: Tornado verification is notoriously difficult to do, and I would think especially difficult across the Central Region due to the relatively sparse population in many areas. How well verification is done also varies by WFO. Please include some very brief discussion on the difficulties with tornado verification and how that might impact the results.

A brief discussion was added to the Introduction section addressing these issues. An additional reference was added (Brotzge et al 2011) to highlight the increase in FAR observed in sparsely populated areas, which are typically located farther from a radar site.

Related to the above, it may be likely that at least some of the EF-scale ratings were underestimated, simply because those tornadoes remained in open fields and failed to hit much infrastructure. However, these tornadoes may have been tagged w/a higher IBTW warning, but classified as a miss due to its lower EF-rating. Please comment on this.

This issue was also included in the brief discussion added to the Introduction (also addressed in comment 1). There certainly may have been cases in which a tornado was underrated. For example, the Wichita WFO issued damage tags in the IBTW for tornado over rural Kansas on 19 May 2013. This tornado was rated an EF2 based on damage to a few farmsteads (silos, barns), however, the tornado remained over *rural country for much of its lifetime. The damage survey notes that the Doppler on Wheels measured winds of 155 mph approximately 70 m AGL. Using the DOW data, the tornado would be classified as an EF4. (Of course, DOW data are not currently incorporated into damage assessments). Also, there are many tornadoes rated EF0 that hit no infrastructure whatsoever (mainly in the domains of offices such as BOU, CYS, and PUB) that may have been underrated. On the other hand, any underrated tornado which was covered by a non-tagged IBTW may actually be inflating the statistics.*

As an alternative to comparing tags with their EF rating, you may want to consider comparing the IBTW tags against each tornado's radar velocity signature. The magnitude of the max velocity shear may provide a better correlation to/the assigned IBTW, or at least in those cases where the EF scale was just one category off. *[Editor's note: This is a valid point but wouldn't represent a truly independent verification approach given that the same radar signatures likely factored into decisions as to what IBTW tags to use. Hence, some possibility of circularity in results would need to be acknowledged should the authors choose to test this.]*

Considering radar velocity signatures according to tags would be a very interesting approach, but it was not within the scope of this research. In the future, the authors would like to revisit the 2013 data, as well as investigate the 2014 data, to see what the typical rotational velocity looked like for tagged versus non-tagged tornadoes (although the editor does point out the possibility of circularity in this exercise). The authors' understanding is that some yet-to-be published research has been done by the NSSL using rotational velocity to assess tornado intensity and that warning forecasters are to use rotational velocity criteria when issuing damage tags heading forward with the project.

The overall statistics are largely driven by a few large-scale outbreak events (4 Oct and 17 Nov). However, it's been shown that warning statistics are generally much better in outbreak situations (though worse during non-spring months). Consider evaluating the statistics with and without the two outbreak events included, or consider them separately.

Verification statistics were evaluated for the two outbreak events (4 Oct and 17 Nov), and the findings were added to the results section.

It would be interesting to explore the potential impact that population density (in the paths of the warned areas) may have had on the forecaster's decision on whether to issue a tag. The sample size may be too small to explore this quantitatively, but it did seem to have an impact at least in the Wichita case study.

The authors agree, this would be an interesting question to explore. It's difficult to tell with this single year of data, but it certainly seems that population density could have played a role (indeed, Wichita being the prime example). To keep from speculating, the authors did not address this in the text, since the reasons for tag decisions by the warning forecaster are unknown. Tag decisions (with respect to population and otherwise) probably varied between WFOs also, and perhaps even between individual warning forecasters. During the 17 Nov outbreak, tags were included in many IBTWs, even when a town was not necessarily in the direct path of the anticipated tornado. However, population density is generally higher in the eastern portions of the Central Region where this outbreak occurred. Reviewer C also mentioned the importance of population density (and potential population bias) in regard to the usage of tags, noting the implications for forecasting practices.

One wonders how much of the forecaster's decision on tornado intensity was influenced by the prevailing SPC outlook of the day. It would be interesting to break out the statistics as a function of SPC outlook category.

Additional contingency tables were constructed and FAR and POD were calculated according to day one SPC outlooks (slight, moderate, and high). A discussion of these findings and additional figures were added to the results section.

Throughout the paper, be careful to note that the statistics for TORs/SVSs are not the same as statistics for TORs only. These are two different categories of sampling, and should not be directly compared.

Additional clarification was added to the text.

The Conclusions section is largely redundant and unnecessary. Consider summarizing your conclusions in a short, bulleted list.

The conclusion section was condensed and the redundancy was eliminated.

[Minor comments omitted...]

REVIEWER C (Kimberly E. Klockow):

Initial Review:

Recommendation: Accept with minor revisions.

Summary: The authors present an evaluation of forecast verification statistics for one of the first years of the impact-based warning experiment (popularly known as IBW). The authors find that while forecasters demonstrate an ability to discern situations that are likely to produce a tornado from those that are less likely, they do not seem to be able to distinguish those tornadoes that are likely to produce severe impacts from the rest. In fact, forecasters frequently used tags when they were not warranted, and failed to use them when they were warranted, leading to both poor POD and FAR numbers. The authors then described some of the factors that potentially contributed to this performance, ultimately leaving those questions to future researchers.

This is an important article, and I'm pleased to help prepare it for publication. The article is generally well-written, and it reads as an unbiased and even-handed evaluation for what has become a contentious topic. As I went through the text, I made note of topics that I hoped the authors would cover in the discussion, and all of those were met.

The key weakness of the article lies in its clarity and organization of concepts. Important points could find themselves buried in text or disconnected from existing graphs/tables. I have a number of comments below that are designed to help the article read more clearly and to become a more handy reference for those wishing to cite it.

The authors wish to thank the reviewer for the helpful comments, which hopefully have resulted in needed clarity to the methods used in the paper. The reviewer also makes a number of interesting remarks, especially in regards to the social constructs of "impacts". There is definitely a need for continued study of IBTWs.

General substantive comments: Defining the verification method & outlining results: Somewhere in the introduction, it would be useful for the authors to clearly re-define POD and FAR in light of the severity-based verification they're doing. They describe it in pieces throughout the text, but it'd be easier to follow if this was done concisely up-front.

An example of a 2×2 contingency table along with POD and FAR definitions was added to the Data and Methods section.

Perhaps the authors could construct a standard 2×2 contingency table, and list the contingencies that would be counted as a "hit" or "miss" within it (so it's clear, for example, that a low POD for tagged warnings doesn't just mean there wasn't a tornado—it also could mean forecasters failed to tag when it was warranted). It may also be helpful if the authors could construct a graph or table (maybe a filled-in version of the previous table?) that clarifies the source of false alarms or missed events as counted in the study (over/under-forecasted warning altogether vs. inappropriate severity tag use). This would make it much easier to follow the results and refer to them later.

A discussion about over/under warnings and how these were classified in the contingency tables was added to the Data and Methods section. Additional tables were added to clarify how the tables for both tagged and non-tagged IBTW TORs/SVSs were constructed.

Relating IBW tags to EF-scale damage: Please define DIs and DoD—are these EF-scale quantities you used to relate IBW tiers to EF-scale damage? Also, please provide a brief description of the method used to pair IBW tiers to EF-scale damage. For example, it's not clear why EF-scale damage verifying the considerable and catastrophic tags was not mutually exclusive. If this is a somewhat subjective exercise, please make note of that. Central Region has informally done some preliminary forecast evaluation work on their IBW products, and they said that anything EF2 or above verified their tagged warnings. Unsurprisingly, this made their numbers look a little better.

Damage Indicators (DIs) and Degree of Damage (DoD) are now defined more clearly in the introduction. In addition, a discussion about how IBTW tiers were paired to the EF Scale (as well as an example of this method using some of the impact wording included in the catastrophic tag) was added. As you note, there is some level of subjectivity in this exercise, especially when it comes to defining somewhat vague words such as "considerable". The authors were careful to relate the impact wording as closely as possible to what is found in the EF Scale. The authors were also aware that the internal Central Region verification was performed using EF2 or higher to verify tagged warnings...however after close investigation, the wording is more appropriate for EF3 and higher.

Dataset used: The authors do not explain why they did not include 2012 in their dataset. IBW in the Central Region began in 2012, though a different tag system was used ("significant" instead of "considerable"), and, as the authors note, the experiment was limited to 5 WFOs at the time. Still, this information could have been included. A very brief explanation of the choice would be helpful.

A brief explanation for this was added to the Data and Methods section.

False alarms with confirmed tornadoes: The authors note that tags were used much more frequently in SVSs than TORs, and were most frequently associated with reported tornadoes. This satisfies one condition of the guidance for using tags—heightened forecaster certainty that there will be impacts. However, the authors also note these tornado-observed SVS/TORs verified less often than their radar-indicated counterparts.

This may lie beyond the bounds of the present study, but I'm curious: Do the authors know if there is there a population bias to the false alarms of tagged warnings with tornadoes reported? It would be great to see this mentioned in the text, if the authors happen to know, or if they would be willing to include it. If true, this could demonstrate an urban bias to warnings—one that is ultimately unhelpful (potentially because of the variable nature of tornadoes, as the authors later describe). In other words, forecasters may hedge their bets toward greater impacts in urban areas, when that may be misleading.

This could have huge implications for practice, as forecasters seem much more willing to use enhanced language near population centers. This practice has seen quite a lot of dispute, especially in academic circles, where social scientists argue that this marginalizes rural populations (they don't get the chance to receive equal warnings). Ultimately, "impacts" are a social construct that goes well beyond population density. As I said, this is certainly not required, but it would be a nice addition if the authors know.

The authors agree, this would be an interesting question to explore. It's difficult to tell with this single year of data, but it certainly seems that population density plays a role (especially in regard to use of the *catastrophic tag, the Wichita case study being the prime example). The data also seem to suggest tag decisions (with respect to population and otherwise) probably varied between WFOs also, and perhaps even between individual warning forecasters. During the 17 Nov outbreak, considerable tags were included in many IBTWs, even when a town was not necessarily in the direct path of the anticipated tornado. However, population density is generally higher in the eastern portions of the Central Region where this outbreak occurred. Reviewer B (Brotzge) also noted that exploring the potential impact of population density on tag decisions would be interesting. This is a needed study as more data is acquired.*

[Minor comments omitted...]

Second Review:

Recommendation: Accept.

Summary: After a thorough review of the paper and the comments of others, I feel satisfied that the authors have addressed my concerns, and the added content (based on the points of other reviewers) does not raise any flags for me. I'll be curious to see how Patrick feels, but will leave those points to him.

I have no further comments, and recommend the article for publication.