The 2021 Hazardous Weather Testbed Experimental Warning Program Radar Convective Applications Experiment: A Forecaster Evaluation of the Tornado Probability Algorithm and the New Mesocyclone Detection Algorithm

THEA N. SANDMÆL,[a,b] BRANDON R. SMITH,[a,b] JONATHAN G. MADDEN,[a,b] JUSTIN W. MONROE,[a,b] PATRICK T. HYLAND,[a,b] BENJAMIN A. SCHENKEL,[a,b,c] AND TIFFANY C. MEYER[a,b]

[a] *Cooperative Institute for High-Impact and Severe Weather Research and Operations, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
[c] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*

(Manuscript received 7 March 2023, in final form 2 May 2023, accepted 4 May 2023)

ABSTRACT: Developed as part of a larger effort by the National Weather Service (NWS) Radar Operations Center to modernize their suite of single-radar severe weather algorithms for the WSR-88D network, the Tornado Probability Algorithm (TORP) and the New Mesocyclone Detection Algorithm (NMDA) were evaluated by operational forecasters during the 2021 National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT) Experimental Warning Program Radar Convective Applications experiment. Both TORP and NMDA leverage new products and advances in radar technology to create rotation-based objects that interrogate single-radar data, providing important summary and trend information that aids forecasters in issuing time-critical and potentially life-saving weather products. Utilizing virtual resources like Google Workspace and cloud instances on Amazon Web Services, 18 forecasters from the NOAA/NWS and the U.S. Air Force participated remotely over three weeks during the spring of 2021, providing valuable feedback on the efficacy of the algorithms and their display in an operational warning environment, serving as a critical step in the research-to-operations process for the development of TORP and NMDA. This article will discuss the details of the virtual HWT experiment and the results of each algorithm's evaluation during the testbed.

SIGNIFICANCE STATEMENT: Before transitioning newly developed radar-based severe weather applications to forecasting operations, an experiment simulating the use of these tools by end users issuing severe weather warnings is helpful to identify both how they are best utilized and address any needed improvements to increase their operational readiness. Conducted in 2021, this study describes the forecaster evaluation of the single-radar Tornado Probability Algorithm (TORP) and the New Mesocyclone Detection Algorithm (NMDA) in one of the first completely virtual Hazardous Weather Testbed (HWT) experiments. Participants stated both TORP and NMDA offered marked improvement over the currently available algorithms by helping the operational forecaster build their confidence when issuing severe weather warnings and increasing their overall situational awareness of storms within their domain.

KEYWORDS: Severe storms; Algorithms; Radars/radar observations; Nowcasting; Operational forecasting

1. Introduction

The Tornado Detection Algorithm (TDA; Mitchell et al. 1998) and the Mesocyclone Detection Algorithm (MDA; Stumpf et al. 1998) were developed by the National Oceanic and Atmospheric Administration's (NOAA) National Severe Storms Laboratory (NSSL) in the 1990s to aid operational forecasters in severe weather warning decision making, and are part of the current suite of operational Next Generation Weather Radar (NEXRAD) Level III algorithms. The TDA and MDA both utilize WSR-88D single-radar data to identify potential tornadoes and mesocyclones in radial velocity data, respectively, by using and expanding upon the pattern vector technique first developed in Hennington and Burgess (1981), further outlined in Zrnić et al. (1982), and applied by subsequent studies that the TDA/MDA development was built upon (Zrnić et al. 1985; Wieler 1986; Desrochers and Donaldson 1992). Unlike some of the early tornado detection algorithms, the TDA does not require the presence of a mesocyclone to generate a detection and utilizes less stringent criteria to identify locations of vortices in radial velocity data. Similarly, the MDA is more thorough than earlier versions as it detects a much larger spectrum of storm-scale rotational features by putting more of the feature thresholding emphasis on the final volumetric features rather than during the initial creation of the radial velocity shear vectors. When operating, the TDA produces a single product while the MDA produces both the mesocyclone detection (MD) product, generated only at the end of a radar volume scan, and a digital mesocyclone detection (DMD) product that is generated with each radar tilt (Warning Decision Training Division 2022b).

While performing well in the case studies presented in their seminal publications, studies evaluating the algorithms using

Meyer's current affiliation: Unidata, University Corporation for Atmospheric Research, Boulder, Colorado.

Corresponding author: Thea N. Sandmæl, thea.sandmael@noaa.gov

DOI: 10.1175/WAF-D-23-0042.1

larger datasets have shown that the detection rate of tornadic vortices by the TDA is very low at around 5%–17% (Jones et al. 2004; Sandmæl et al. 2023) and numerous spurious detections, largely due to poor velocity de-aliasing and ground clutter, plague the MDA (Mitchell et al. 2000; McGrath et al. 2002). Since the introduction of these algorithms, there have been several major advancements to the WSR-88D fleet that aid forecasters in the interrogation of severe and tornado-producing storms including dual-polarization (e.g., Ryzhkov et al. 2005), upgrades to scanning strategies in both spatial and temporal resolution (Torres and Curtis 2007; Kingfield and French 2022), and new derivative radar products (Mahalik et al. 2019). To fully leverage these modern products and advances in radar technology, the National Weather Service (NWS) Radar Operations Center (ROC) tasked the NSSL with developing two new algorithms, the Tornado Probability Algorithm (TORP) and New Mesocyclone Detection Algorithm (NMDA), as proposed replacements for the operational TDA and MDA, respectively. Described in detail by Sandmæl et al. (2023), TORP is a machine learning-based algorithm that provides object-based probabilities of a tornado occurring that are derived from a random forest model (Breiman 2001). The NMDA does not utilize machine learning but instead diagnostically detects and tracks mesocyclones and other rotational features within volumetric radar data by applying a defined set of statistical parameters and thresholds to various single-radar fields. While both algorithms utilize a linear least squares derivative (LLSD) azimuthal shear (AzShear; Mahalik et al. 2019) gradient of de-aliased radial velocity as the main product to drive identification of rotational interest areas, TORP provides a probability of an event occurring (a tornado) while the NMDA provides a binary yes/no decision on an event (a mesocyclone).

Operational products that are under development greatly benefit from going through a thorough iterative research-to-operations (R2O) process (Serafin et al. 2002; Kain et al. 2003; Clark et al. 2012; Gallo et al. 2017). Feedback from the products' intended end users is essential to ensure that a product addresses their operational needs. There are several success stories of transitioning products from research to operations, such as ProbSevere, a machine learning-based probabilistic guidance for severe storm warning operations (Cintineo et al. 2013, 2018, 2020), and the Hazard Services software toolkit that allows NWS forecasters to streamline the process of issuing products such as flood warnings (Argyle et al. 2017). As is the case with many other operational products, ProbSevere and Hazard Services have both gone through extensive testing through experiments in NOAA's Hazardous Weather Testbed (HWT) before being transitioned to fully operational forecasting tools.

One long-standing HWT program that has been part of the R2O process for many operational algorithms and techniques is the Experimental Warning Program (EWP; Calhoun et al. 2021). Developed and formalized within the HWT in the early 2000s, the EWP resulted from algorithm development collaborations between NSSL and local NOAA NWS Weather Forecast Offices (WFOs; e.g., Scharfenberg et al. 2005). The purpose of the EWP is to work toward an improvement in forecasting severe and high-impact weather on the convective warning time scale (0–2 h) by transitioning novel research products to operations and soliciting invaluable feedback from the intended users before the products are finalized, through the use of targeted experiments, such as the Radar Convective Applications (RCA) experiment. Originating in 2019 and organized by principal investigators (PIs) from the University of Oklahoma (OU) Cooperative Institute for Severe and High-Impact Weather Research and Operations (CIWRO), formerly known as the Cooperative Institute for Mesoscale Meteorological Studies (CIMMS), and NSSL, the RCA experiment serves to evaluate the utility of innovative single- and multiradar based convective algorithms, including TORP and NMDA.

In 2020, the planned evaluation of TORP and NMDA in the RCA experiment was postponed due to COVID-19 and moved to 2021 where it operated as a fully virtual experiment. While not the first HWT experiment to be held completely virtually (Clark et al. 2021), the RCA experiment was one of the first to utilize remote versions of the Advanced Weather Interactive Processing System II (AWIPS-II), which was hosted on the Amazon Web Services (AWS) cloud platform that allowed participants to evaluate the algorithms using AWIPS-II just as they would within a typical in-person HWT experiment. Combining this with the use of online communication platforms such as Google Meet and Slack, PIs were able to closely recreate the in-person HWT experience, providing participants with the necessary resources to properly assess the algorithms via this new virtual HWT format.

Considered a success by both PIs and participants, the virtual 2021 HWT EWP RCA experiment provided a wealth of important feedback to drive the development of TORP and NMDA, even in the midst of a fledgling virtual HWT environment. This article serves as a complete overview of the experiment by providing a summary of the two algorithms evaluated, the experiment design, and the results of the experiment through detailing the reception and evaluation of each algorithm. The subjective assessment of the algorithms explores how they can be used operationally and how the iterative development and testing process involving communication between the research and operational communities can be utilized to benefit algorithm development.

2. Algorithm descriptions

TORP and NMDA are part of a larger effort by the ROC to modernize their suite of single-radar severe weather algorithms for the WSR-88D network. Both algorithms utilize maxima of velocity-derived AzShear in the identification of rotational interest areas within the single-radar data, which is different from the pattern vector techniques used by the TDA and MDA to identify rotation with a radial velocity field. LLSD AzShear helps to provide a cleaner, more easily quantifiable radar field on which to examine for rotation, which is achieved through removing outliers in the data by smoothing and using a local neighborhood of values for a more robust gradient calculation. These interest areas are then further interrogated by each algorithm using different methods in order

to ascertain the probability of a tornado occurring (TORP) or whether an area of rotation exists (NMDA).

The random forest machine learning model used by TORP to calculate tornado probabilities is based on radar data from a wide variety of past weather events. TORP utilizes all available base radar (Level-II) products from single-radar single-tilt scans (0.5° by default) from the WSR-88D network, including dual-pol variables (NOAA/NWS/ROC 1991), to create predictors for the random forest model. It also uses the LLSD gradients (azimuthal, radial, and total gradients) of each radar product (Mahalik et al. 2019). The model compares these predictor values with thresholds found from random samples of the radar data from past weather events. In overview, TORP 1) reads in single-tilt single-radar data and the random forest model summarized in a text file, 2) creates objects based on a 0.006-s^{-1} threshold of single-radar AzShear, 3) extracts data within 2.5 km of the center of each object, 4) assigns a tornado probability based on the random forest model output, and 5) tracks the objects in time. More details can be found in Sandmæl et al. (2023).

The NMDA operates over all available tilts and uses a radial velocity-derived LLSD AzShear field in conjunction with other radial velocity-derived products and base radar (Level-II) fields to drive detection generation, perform quality control, and derive detection-based attributes that provide information such as feature strength and size. To track detections over time, mean storm motion is calculated from a model-derived sounding table that is used as a method to initially link detections before an object-centric tracking method gradually takes over as a detection builds a longer track history. Additional details regarding the NMDA's detection generation and tracking process, along with descriptions of the algorithm's strengths and limitations, are outlined in the appendix.

3. 2019 HWT EWP RCA experiment—NMDA evaluation

The NMDA's initial prototype was first evaluated in the 2019 HWT EWP RCA experiment, two years prior to the 2021 virtual experiment described in this study. Occurring over 6 weeks in the spring of 2019, the NMDA was evaluated by 35 participants, including operational forecasters from both the NWS and Department of Defense (DoD), who provided feedback on its performance in a pseudo-operational warning environment. Participants evaluated the ability of the NMDA to detect and track mesocyclones relative to the MDA, which included both its MD and DMD products (Warning Decision Training Division 2022b). Since the initial NMDA development task only included replacing the core of the MDA, the NMDA utilized the preexisting MDA AWIPS-II visualization package for this HWT evaluation. Due to TORP being in the early stages of development in 2019, it was not included in this experiment.

The experiment participants found that the performance of the NMDA displayed promising results in its ability to aid the operational forecaster in detecting and tracking rotation within thunderstorms, outperforming the MDA across the same metrics. Relating directly to the MDA, the overwhelming majority of participants stated, both informally in discussion and formally through specified feedback collection methods (i.e., surveys), that they did not operationally use the MDA products. Their reasoning stemmed from three main arguments: 1) low trust in the MDA due to the known high false-alarm rates, 2) official training that encourages the use of radar base data when making warning decisions (Warning Decision Training Division 2022a), and 3) a cumbersome visualization package that features a large unmovable table listing algorithm detections. These MDA-related findings combined with the promising performance of the NMDA provided sound justification to forego a second evaluation between both algorithms in the 2021 HWT experiment. Additionally, it elevated the need to update the existing AWIPS-II visualization package.

New AWIPS-II algorithm visualizations

Following these results, the main focus of the work following the 2019 HWT experiment was to develop a new AWIPS-II visualization package to provide a modern design that allowed the full potential of TORP and NMDA to be utilized by the operational end users. The 2019 participants suggested that the ideal NMDA visualization (and eventually TORP) would display detections as icons with a mouse-over feature that would produce a dropdown list of detection attributes, similar to that of the preexisting ProbSevere visualization plugin within AWIPS-II (Fig. 1 from Cintineo et al. 2018). This simple read-out would allow participants to view additional details in a compact and user-friendly graphic, replacing the static tables used by the TDA and MDA products to display a tabulated list of the current detections and their attributes (Fig. 1). The amount of screen space that is occupied by an AWIPS-II product is very important to forecasters and the use of large graphics, such as those with the TDA and MDA, can quickly overwhelm an AWIPS-II display interface.

By the start of the 2021 HWT evaluation in April 2021, a prototype of the new AWIPS-II visualization plugin was ready for participants to evaluate (Fig. 2). The strengths of this new visualization lies in its simple and easy-to-use format that is highly customizable by the end user to tailor it to their specific operational needs. Detection icons are shaded by probability (TORP) and AzShear strength (NMDA) to provide a quick method for prioritizing the importance of each detection. For both algorithms, the mouse-over readout displays detection attributes such as strength, size, tracking information, and details specific to the feature each algorithm is engineered to detect. For example, TORP provides a list of the machine learning model predictors, such as the nearby AzShear maximum associated with the rotational strength of a circulation or the local minimum in correlation coefficient, which could be indicative of a tornado debris signature (Ryzhkov et al. 2005), while the NMDA displays information important to mesocyclones such as depth characteristics. Certain attributes, such as maximum AzShear, also contain contextual labels (e.g., low, medium, high, extreme) that are based on climatology-derived distributions of severe storm reports. Additionally, since both algorithms track their detections over time, trend indicators

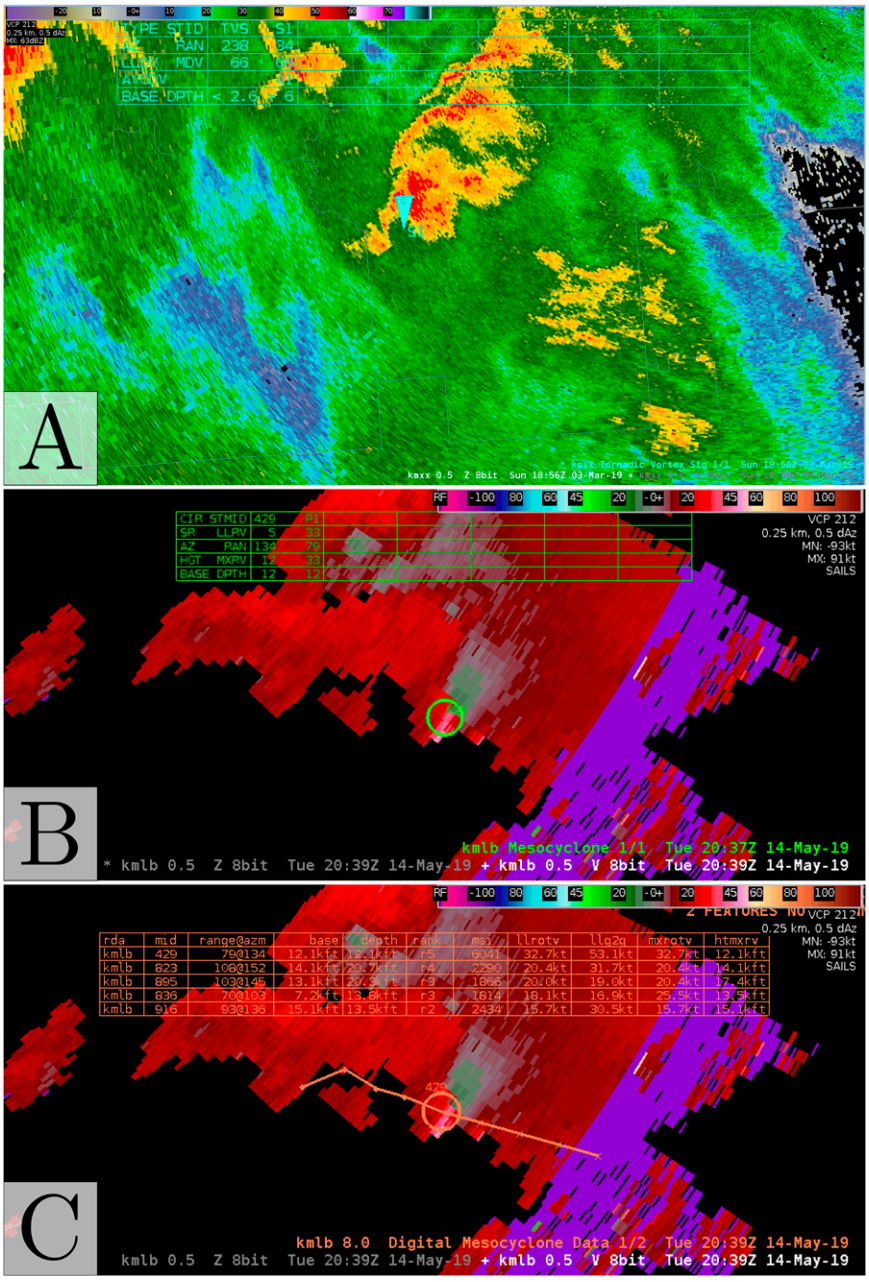

FIG. 1. Example of (a) a TDA detection, (b) an MD detection, and (c) a DMD detection as they are currently displayed in AWIPS-II for two different storms on 3 Mar 2019 (TDA) and 14 May 2019 (MD and DMD).

are added next to trackable attributes to provide context on whether its value is changing. All attributes displayed by the mouse-over readout are optional to display, allowing the end user to decide what detection information is pertinent to them.

4. 2021 Virtual HWT experiment design

The 2021 HWT RCA experiment was conducted over three weeks (19–23 April, 3–7 May, 17–21 May) and designed/facilitated by six CIWRO meteorologist staff members affiliated with the NSSL, who were all certified for social/behavioral/educational human subject research through the Collaborative Institutional Training Initiative, also known as the CITI Program (Collaborative Institutional Training Initiative 2023). 16 forecasters from the NWS and two from the U.S. Air Force participated in the experiment, with six forecasters participating each week. The forecaster participants were chosen from a pool of applicants that were asked to list the length of their forecasting

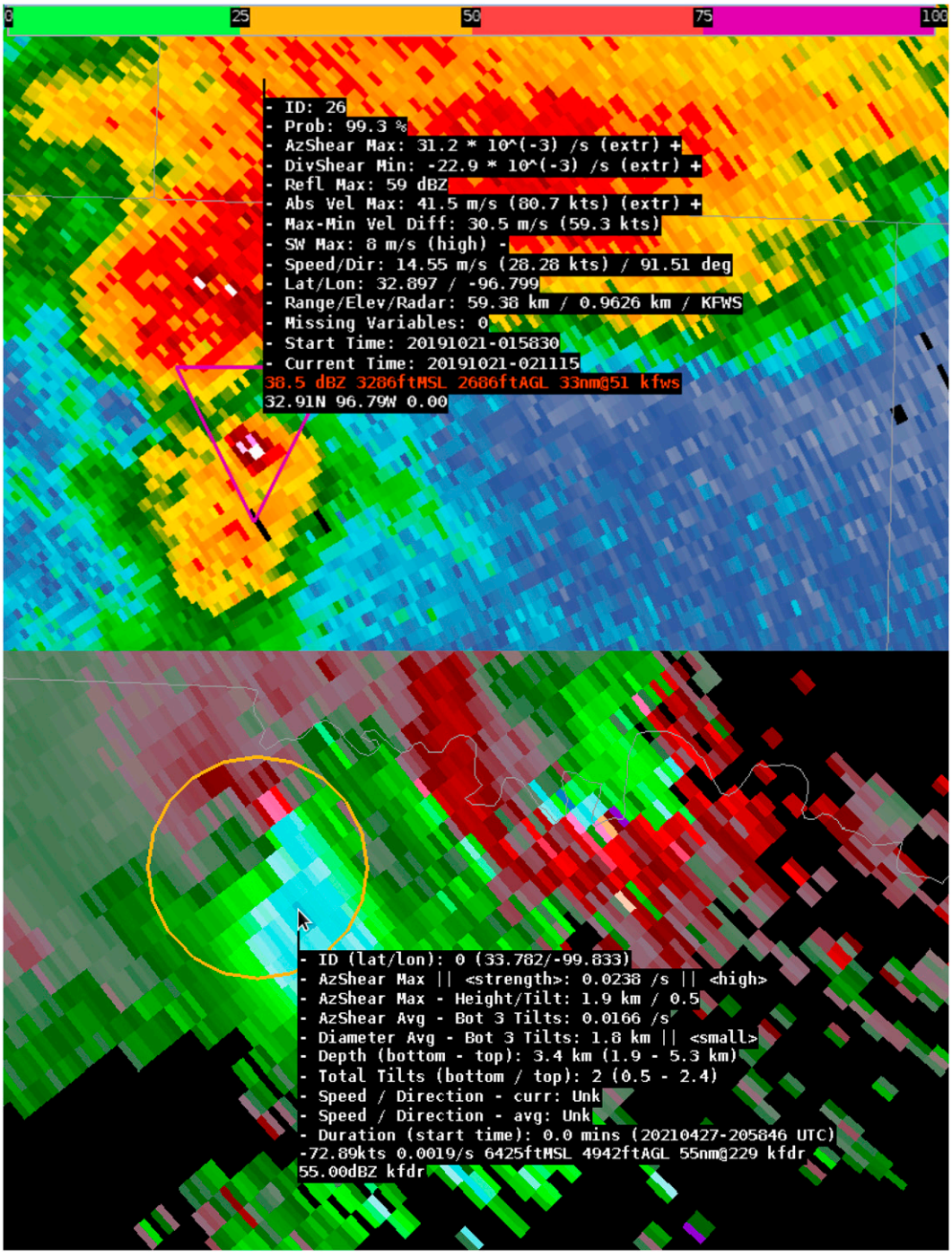

FIG. 2. TORP and NMDA as displayed in AWIPS-II, where TORP is displayed as triangles and NMDA is displayed as circles.

experience with the NWS and how long they have been issuing warnings, as well as provide a personal statement describing why they were interested in participating. The applicants were chosen primarily based on their personal statements, but were also picked to represent a group diverse in experience and geographic location. How many times the applicants had applied and how many times they had been chosen to be a part of prior HWT experiment were also taken into account, giving forecasters who had low chosen-to-applied ratios the opportunity to participate. This led to a group of ten general forecasters, five lead forecasters, and one science and operations officer from the NWS, with 2–29 years of warning experience with a median of

FIG. 3. The virtual setup for the 2021 Hazardous Weather Testbed Experimental Warning Program Radar Convective Applications.

9.5 years. Additionally, the ROC selected two meteorological trainers from the U.S. Air Force, each with decades of forecasting experience to represent the DoD, one of the funding agencies for the ROC. The experiment consisted of the participants evaluating the operational performance and usefulness of TORP and NMDA through a blend of real-time and displaced real-time weather events (hereinafter referred to as archived cases) displayed in AWIPS-II hosted virtually on AWS's cloud platform.

Prior to the start of the experiment week, participants were provided with materials that gave a short overview of the algorithms and how to use them. Each experiment week operated Monday through Friday, and began with group introductions and an overview of the experiment on Monday morning (Fig. 3). To further familiarize the participants with the products and the technology they would be evaluating during the experiment, two training cases were conducted; one archived case in the morning and a real-time case in the afternoon. These cases were also used to identify and remedy any technological problems that the participants might experience in regard to operating in the virtual HWT environment. The core of the algorithm evaluations took place Tuesday through Thursday where participants used a mix of archived—that is, displaced real-time simulations of past weather events—and real-time cases during the evaluation process (see Table 1 for the full list of cases evaluated). In the event that there was not any active weather in the contiguous United States, the afternoon real-time case would be replaced by an additional archived case. Postevaluation group discussions were conducted at the conclusion of each case, followed by online surveys soliciting additional forecaster feedback. Friday concluded with an end-of-week group debrief and final survey covering all facets of the experiment week.

The experiment was designed to focus on gathering feedback and ideas for product improvement and subjective algorithm performance evaluation, rather than controlled experiments to measure the objective benefits of the algorithms in a warning environment (i.e., comparing results between a control group without access to the algorithms to a group that did). Therefore, the experiment design was flexible in terms of evaluation and feedback methods, while still retaining an overarching structure. During each evaluation case, participants were split into pairs to work as a team, with each pair assigned to a Google Meet video chat room, while working with a facilitator who would encourage discussion about what the participants were observing and provide technical assistance as needed. To promote fresh perspectives during the evaluations, participants were assigned a new partner each day. While working an event, forecasters were largely encouraged to evaluate the algorithms using four different methods: 1) investigating a single algorithm, 2) investigating both algorithms, 3) utilizing both algorithms in a simulated real-world warning operations environment, or 4) using both algorithms in a simulated mesoscale analyst role.

Participant feedback was collected via multiple methods to foster different types of feedback that allowed project PIs to accurately judge the performance of the algorithms, identify problems, and prioritize future product improvements and additions. Online surveys provided an even baseline throughout

TABLE 1. List of cases. "Archived training" and "real-time training" refer to the cases that were used to familiarize participants with the technical aspects of the experiment and were not formally used to evaluate the algorithms.

Date	Time (UTC)	County warning area(s)	No. of forecasters evaluated	Type of case	Convective mode
21 Oct 2019	0100–0300	Fort Worth–Dallas, TX (FWD)	18	Archived training	Supercell
15 May 2018	1810–2025	Albany, NY (ALY); Binghamton, NY (BGM); New York/Upton, NY (OKX)	17	Archived	Supercell/linear
28 May 2018	0100–0315	Indianapolis, IN (IND); Northern Indiana (IWX); Wilmington, OH (ILN)	18	Archived	Supercell
21 Oct 2019	0230–0445	Norman, OK (OUN); Tulsa, OK (TSA)	18	Archived	Linear
31 May 2018	2045–2300	Pocatello, ID (PIH)	6	Archived backup	Supercell
19 Jun 2018	2000–2200	Boulder, CO (BOU); Cheyenne, WY (CYS)	12	Archived backup	Supercell
19 Jul 2018	1930–2200	Des Moines, IA (DMX)	18	Archived backup	Ordinary/supercell
19 Apr 2021	1930–2100	Melbourne, FL (MLB); Miami, FL (MFL)	6	Real-time training	Multicell (weak)
3 May 2021	1900–2015	Columbia, SC (CAE); Greenville-Spartanburg, SC (GSP); Peachtree City/Atlanta, GA (FFC)	6	Real-time training	Supercell/multicell
4 May 2021	1950–2215	Birmingham, AL (BMX); Jackson, MS (JAN); New Orleans/Baton Rouge, LA (LIX)	6	Real-time	Linear
6 May 2021	2000–2230	IND; Lincoln, IL (ILX); Memphis, TN (MEG); Paducah, KY (PAH)	6	Real-time	Linear/supercell (weak)
17 May 2021	1900–2030	Albuquerque, NM (ABQ); Amarillo, TX (AMA); Austin/San Antonio, TX (EWX); FWD; Lubbock, TX (LUB); San Angelo, TX (SJT)	6	Real-time training	Supercell
18 May 2021	1950–2315	AMA; EWX; Corpus Christi, TX (CRP); FWD; LIX; Shreveport, LA (SHV)	6	Real-time	Linear/multicell

the experiment upon which to judge each algorithm's usefulness and performance across the various cases. Verbal communication was an extremely important part of the feedback process, both through the one-on-one discussions that occurred between participants and facilitators during an evaluation and the larger group discussions that occurred after each evaluation. The open verbal discussion allowed the PIs to ask pointed questions that dug deeper into particular aspects of each algorithm and to obtain feedback that might not be captured via the surveys. Finally, blog posts were used during algorithm investigation to provide more detailed analysis by participants of important algorithm behavior, including screenshots, or suggestions on how the products could be improved (publicly available at https://inside.nssl.noaa.gov/ewp). While broad examples of blog usage were provided, no specific guidance was given other than to document positives and negatives of the algorithms, allowing participants to judge what level of detail or frequency was appropriate for their particular use of the blog during an evaluation. These feedback methods were analyzed through manual review and summarized by grouping each piece of feedback into common themes and selecting relevant quotes.

The virtual nature of the HWT experiment combined with novel algorithm and visualization development presented unique challenges and opportunities. Both updates to the experiment design, as well as updates to the algorithm products themselves, were sometimes performed immediately after evaluation periods based on feedback received from the participants. A positive to a dynamic HWT evaluation like this is that it enabled the experiment to adapt to participant needs and to provide a better experience for those attending during later weeks. Moreover, it increased the diversity of the feedback by enabling a shift in focus toward different areas of interest in the event an early forecaster consensus on a particular topic was evident. However, these week-to-week changes can sometimes make it difficult to glean consistent results across participants, particularly if the changes involve the algorithms themselves and the performance metrics that are being calculated (i.e., warning verification). In the case of this experiment, no updates were implemented to the core algorithm logic, rather, only to the downstream visualization package. While this did have a positive impact on algorithm usability, the benefits of being able to improve the visualization in real time before a future transition to operations far outweighed any potential drawbacks (see results section below for further discussion). Beyond this update to the visualization package, adjustments to the experiment format included providing additional background information for archived cases, such as local terrain maps and surface observations, as well as more detailed instructions and expectations of how the virtual product evaluations were going to occur. Participants in the second and third weeks were also provided local storm reports for archived events as if they were occurring in real time, announced both verbally and sent via Slack messages by a staff member, as opposed to only viewing tornado tracks of the event after the evaluation period (Fig. 4). The forecasters had most, if not all, of the products that they would have had available operationally in AWIPS-II accessible to them during the case evaluations.

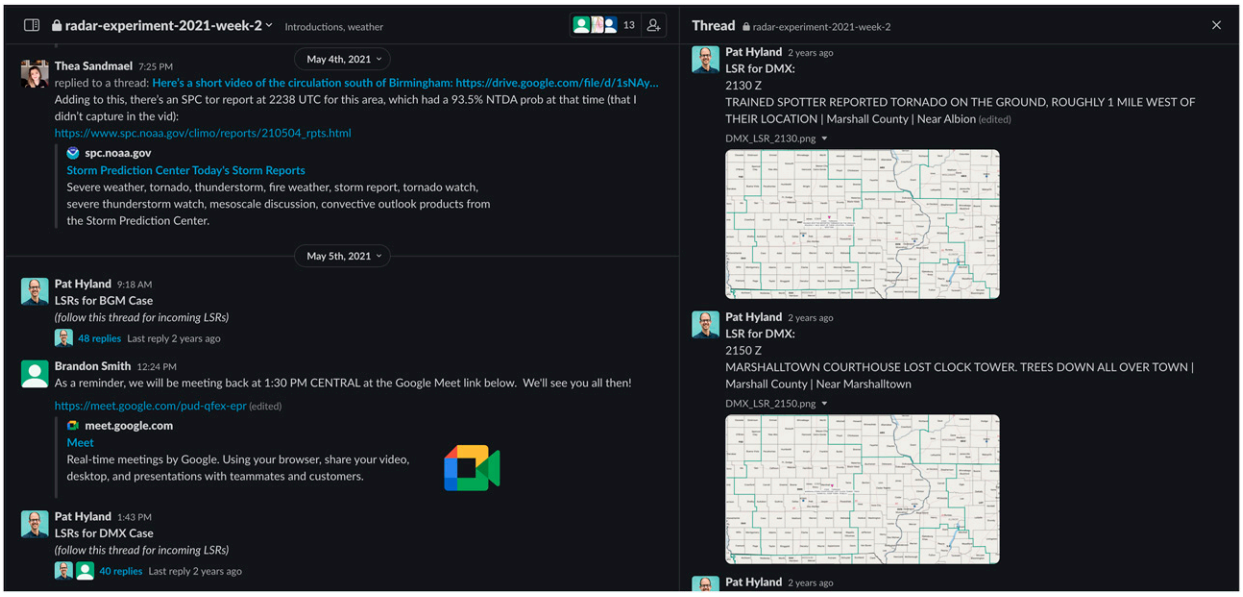

FIG. 4. Screenshot of the experiment's Slack workspace.

5. 2021 Virtual HWT experiment results and discussion

The overarching goal of this HWT experiment was to use participant exploration of the algorithm products and forecaster feedback to guide further algorithm development, advancing them to a readiness level for operational transition. Participants were provided key aspects to focus on during their evaluations, including the utility of the algorithms during the warning decision-making process and the presentation of algorithm information within the software. Forecasters were asked how they would use the algorithms while working as a radar operator, during the warning process, or other pertinent activities. More specifically, project PIs were interested in how the information provided by the algorithms influenced their decision to warn (or not warn) a storm. Additionally, participants evaluated the new AWIPS-II visualizations for each algorithm and were asked to provide feedback on topics such as the usefulness of the display, detection icons, attribute fields, and several different types of customizable features.

The participants provided a wealth of feedback on their impressions of the algorithms during the experiment. Findings can be largely summarized into the main topics of algorithm performance, impacts to warning decisions, operational readiness, comparisons with the legacy products (in the case of TORP), and the visualization package. This section serves as an overview of the key findings associated with each algorithm.

a. *TORP*

1) QUALITATIVE COMPARISON WITH THE TDA

The main goal of TORP is to replace the TDA, so it was important to gauge the forecasters' opinions on how the two algorithms compare in performance. In addition to the visualization upgrades, TORP provides additional storm-based information and a probability-based likelihood of a detected tornado versus TDA's binary detections, where a detection always infers the presence of a tornado. TORP also utilizes dual-polarization products associated with tornado debris signatures (Ryzhkov et al. 2005), which is something that was unavailable during the development of the TDA.

While TORP was not assessed during the 2019 HWT experiment, many of the results pertaining to the use of the currently operational MDA products also applied to the operational use of the TDA. Citing an abundance of false alarms and recommendations against using it in operations by experienced forecasters or training instructors, five (~28%) of the participating forecasters said that they had never used the TDA. Of the forecasters that had used it before, 69% stated that they never or rarely use it (Fig. 5), claiming it is "*terrible*," "*useless*," and "*unusable*" due to "*ridiculous false alarm rates*," recommending against operational use of the TDA.

When asked at the end of the week whether the forecasters would use the TDA, TORP, or neither, 100% of the forecasters replied that they would use TORP. Similar to the results from the 2019 NMDA HWT evaluation, some of the forecasters mentioned that the TDA visualization was a deterrent to using the TDA. One forecaster found it "*almost hard to compare*" the two algorithms since "*TORP is much, much better*," while another described TORP as "*lightyears better*" than the TDA. These findings also reflect the results in Sandmæl et al. (2023) that showed TORP performs better than the TDA in every objective performance metric. The forecasters showed excitement about the prospect of TORP replacing the TDA, including one who said they were "*really hoping to see TORP in operations*" after evaluating it and comparing it with the TDA.

2) FORECASTER-EVALUATED ALGORITHM PERFORMANCE

Overall, forecaster reception of TORP was very positive, with 100% of the forecasters rating its ability to detect

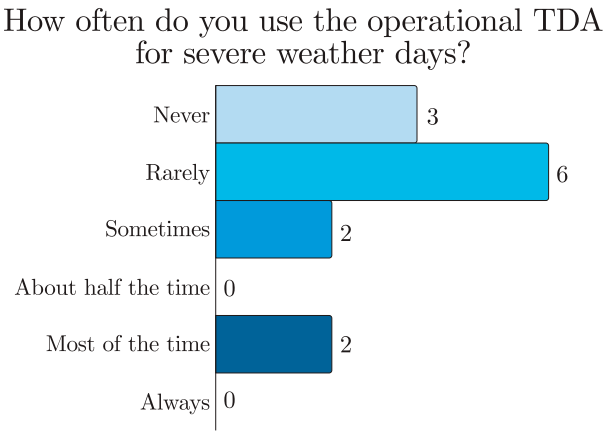

FIG. 5. Weekly survey feedback describing frequency of TDA use ($N = 13$). The number of responses for each option is listed to the right of each bar.

tornadoes as "good" or "very good" (Fig. 6). The participants found the product useful overall (Fig. 7), including one who emphasized that "*with the abundance of information forecasters have, even 'Slightly' or 'Moderately' useful can be big*," referring to TORP's ability to condense and summarize large quantities of single-radar information into one product. The only forecaster that was unsure if they would use TORP participated in the first week of the evaluation when the algorithm was less refined. This forecaster said they would use TORP if it had a more organized readout to better utilize the information and additional filters for non-meteorological detections, both of which were at least partially addressed by the introduction of the customizable menu options during the end of the second week of the experiment.

3) OPERATIONAL UTILITY

After evaluating TORP in a simulated warning environment, its noted advantages included its ability to provide good guidance for rapid decision making and boosting forecaster confidence in high-stress environments. 94% of the forecasters agreed that TORP provided an overall increase in confidence in warning situations. As indicated by the results in Fig. 7, many forecasters also verbally expressed that they would use TORP as a storm-interrogation tool for both warning decisions as well as for overall situational awareness. By looking at the color-indicated probability from the icons (colorbar shown in Fig. 2), forecasters could quickly prioritize storms and identify which circulations should be interrogated first. Additionally, TORP alerted forecasters to storms that were intensifying, which was especially helpful during busy events. TORP sometimes even prompted forecasters to investigate circulations earlier than they would have otherwise.

When asked after each case whether or not TORP helped increase tornado warning lead time, the forecasters answered "yes" 71% of the time, citing that TORP increased their confidence in issuing a warning. Forecasters indicated that having their reasoning backed by the algorithm would remove their need to wait for one more radar scan to make the final decision to issue a tornado warning.

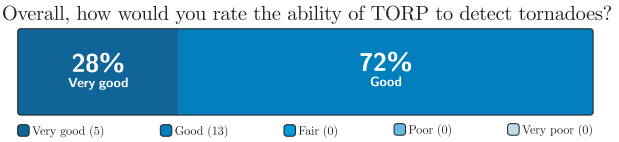

FIG. 6. Weekly survey feedback rating TORP's ability to detect tornadoes ($N = 18$). The answer options are listed in the legend under the bar, with the number of responses for each option listed in parentheses and the fraction of the responses displayed on the bar itself.

Several forecasters mentioned that TORP was also especially helpful for both situational awareness and warning issuance for marginal cases. For example, during an evaluation of a displaced real-time case featuring a quasi-linear convective system (QLCS), one forecaster missed a weak tornado early in the simulation while in tornado-warning operations, only realizing this after being alerted of a tornado report. The TORP probabilities that were associated with this missed tornadic circulation were high relative to other circulations in the case and the forecaster used this experience to recognize a potential developing tornado. When the forecaster noticed TORP probabilities in the 40%–45% range for a new circulation, the same range as the previously missed tornado, it increased their confidence enough to issue a tornado warning that provided a 10–12-min lead time on what eventually became a new QLCS tornado. They explained that this lead time would have been reduced without the assistance of TORP, as they would have waited to see more QLCS tornado confidence builders before making the tornado warning decision. QLCS tornadoes represent a difficult forecasting challenge (Brotzge et al. 2013) and many forecasters agreed that identifying areas of concern in QLCSs is one of TORP's main strengths. This type of mental probability "re-calibration" was also observed with several forecasters when they were evaluating an archived case that included multiple storms with weak radar returns that produced tornadoes. One forecaster explained that they "*warned on a storm that [they] would otherwise never have thought to warn on given the radar fields.*"

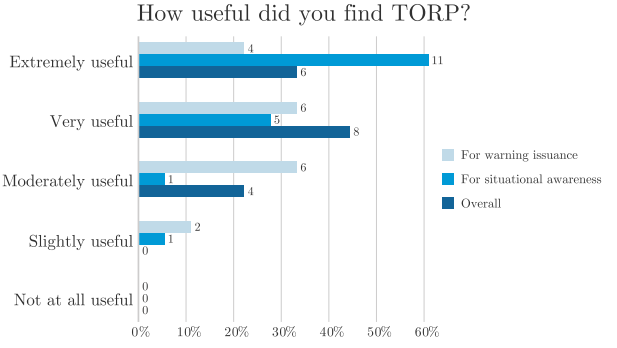

FIG. 7. Weekly survey feedback concerning TORP's usefulness ($N = 18$). The number of responses for each option is listed to the right of each bar.

This warning ended up verifying with a tornado report soon after the forecaster issued a tornado warning, claiming that "*without TORP highlighting that storm, [they] would have missed it completely.*"

Though some forecasters were comfortable using TORP as a primary decision maker for tornado warnings in certain situations, many stressed that TORP is not going to make warning decisions for them. It is rather something that they will take into consideration when they are on the fence about issuing a tornado warning, considering whether to add impact tags, or when to discontinue a warning.

4) Addressing limitations and suggested additional capabilities

To prepare TORP for operational use, it was important to identify areas of improvement before the transition to operations. The main limitation that the forecasters identified for TORP was the occurrence of false detections, mainly caused by ground clutter, poor de-aliasing, and sidelobe-contaminated velocity data. By design, TORP will evaluate every area with high rotational shear thresholded using a 20-dBZ reflectivity filter dilated to expand the area covered by higher reflectivity values and despeckled to reduce noise, which will retain the highest possible number of real tornado detections to be evaluated by TORP's random forest model. Noted in the blog posts written by the forecasters, this relatively relaxed filtering leads to numerous low-probability detections, typically ranging from 0% to 20%, in association with obvious false detections based on their meteorological experience. The majority of these blog posts were written during the first week of evaluations, before the implementation of the ability to filter TORP detections based on an adjustable probability threshold, indicating that this feature alleviated some of the issues with low-probability noise detections.

Even though the false detections generally were assigned very low probabilities, the forecasters would still refer to them as false alarms. However, through the majority of the case evaluations, the forecasters thought TORP handled the false alarm detections well, with 64% rating the algorithm's ability to handle false alarms as "good" or "very good" with 31% rating it as "fair". The forecasters were divided on whether the developer should prioritize implementing a more aggressive detection filtering method. However, they agreed that users would likely trust the algorithm more if non-meteorological and obviously non-tornadic detections were filtered out. During the first two weeks of the experiment, participants were asked if they would like a way to hide detections based on probability, a feature that most of them felt was necessary. To address this feedback, a probability slider was introduced during the second week and tested during the third week of the experiment (Fig. 8). The slider allows users to threshold detections based on their preferred probability values. While some participants from the third week still suggested that the algorithm should have more internal filtering, most felt that the slider enabled them to sufficiently filter out detections caused by bad data (Fig. 9). The majority of participants also agreed that relying solely on the probability slider would be preferable to inadvertently filtering out legitimate tornado detections in areas of weak reflectivity by applying excessive filtering.

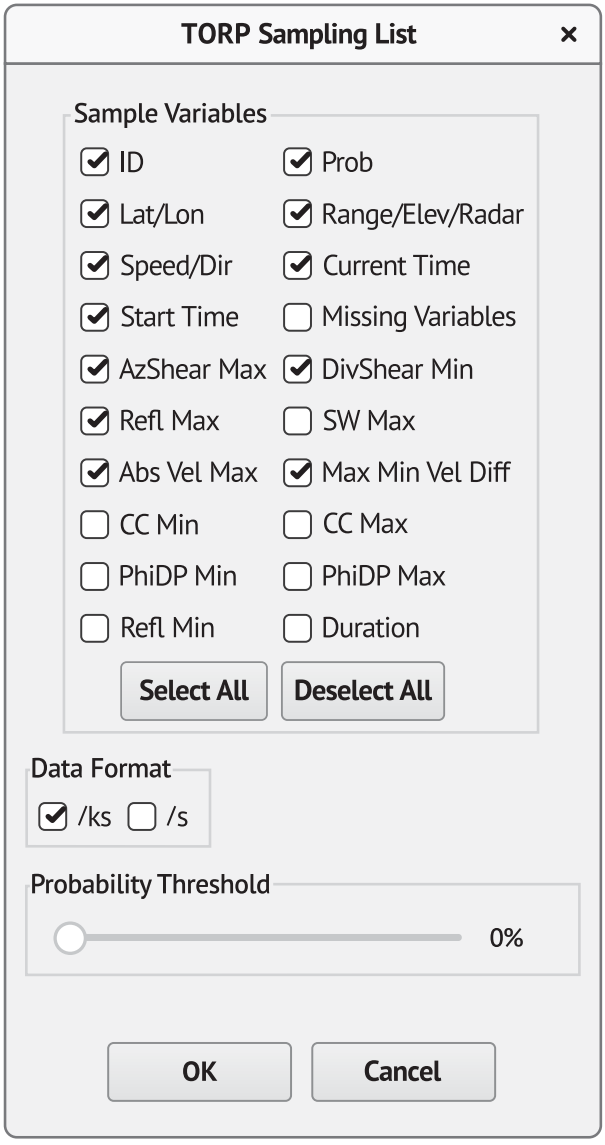

FIG. 8. The TORP AWIPS-II menu available at the end of the experiment for customizing the output displayed to the user.

In addition to the probability slider, one idea that was suggested to mitigate false-alarm detections was to include an option to turn on a method that aggressively filtered detections. To accommodate the forecaster preference to preserve weak tornado detections, a new output flag was added to TORP following the experiment, which allows visualization software to provide users with the option to toggle between regular and aggressive detection filtering. This aggressive filtering can be turned off if a weather event appears to be conducive to rotating storms with weak reflectivity signatures or those exhibiting displacement between velocity and reflectivity signatures. Another measure to limit false detections that has been

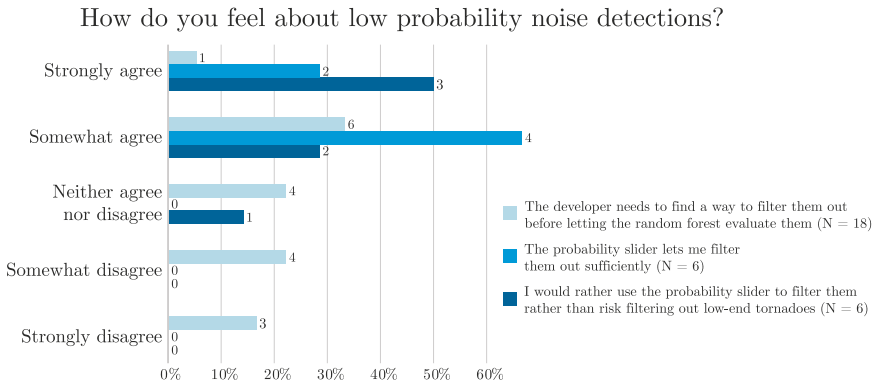

FIG. 9. Weekly survey feedback concerning TORP's low-probability false alarms. The number of responses for each option is listed to the right of each bar.

implemented since the HWT experiment was to provide the option to flag and remove low-probability detections within 30 km of a radar, allowing the user to view the flagged detections in the event that a storm is passing over or near a radar site.

The forecasters also noted that issues in velocity data, such as velocity de-aliasing failures or sidelobe contamination caused by large hail, are more likely to produce higher tornado probabilities than the obvious noise detections based on clutter. The effects of this type of contaminated data could also cause more erratic detection locations when tracking a circulation over time. While experienced forecasters can recognize signs of hail contamination, it can be challenging to identify for anyone with less training in examining radar data. Forecasters suggested that these commonly encountered noise detections should be emphasized in TORP's training documents. Many forecasters said they were more accepting of false alarms once they understood exactly how and why TORP responded to certain types of bad data, stressing the need for thorough training that explores the algorithm's behavior and caveats.

The participants were also able to identify a few rare cases where TORP did not behave as anticipated. As such, TORP's internal object identification logic was revised to properly handle certain artifacts in the radar data that could lead to multiple detections for one tornadic storm or cause missing detections, which was a very important internal algorithm improvement directly resulting from feedback gathered in forecaster blog posts. There were also several improvements to TORP that participants said would expand the current capabilities of the algorithm. Forecasters emphasized a desire for TORP to provide predictive information, such as increases in rotation that were forecast. There are ongoing efforts to expand TORP from a detection algorithm to a detection and prediction algorithm. Currently, a dataset including the 0–60-min pretornadic period of all tornadic storms in the dual-polarization era (2013 to present) is being generated. This dataset will be used to develop pretornadic probabilities separate from the current tornado detection probabilities, potentially allowing the users to better forecast increasing tornado potential. Participants also suggested expanding TORP's detection generation to other radar tilts, as TORP was only available with the 0.5° tilt during the experiment. Following the HWT, this particular algorithm expansion has been completed and TORP can now process any radar tilt. Limited testing of this new capability on tilts below 1.9° showed very little, if any, change in objective performance metrics from the 0.5° tilt. Finally, an upgrade to include probabilities for whether a detection has the potential to produce a significant (EF2+) tornado was discussed with the forecasters, which was generally a well-received idea. A preliminary machine learning model has been trained to provide these probabilities, which can be incorporated into the list of TORP outputs and used to support impact-based tornado warning tag decisions.

b. NMDA

1) FORECASTER-EVALUATED ALGORITHM PERFORMANCE

To evaluate the performance of the NMDA during the 2021 HWT, key questions in both the individual and weekly surveys asked how well the NMDA performed regarding its ability to both detect and track mesocyclones. Similar survey questions during the 2019 HWT evaluation allowed perceived impacts of algorithm improvements on the NMDA's performance to be measured. When participants were asked to rate the NMDA's ability to detect individual mesocyclones, 83% of 2021 participants provided a "good" or better rating, a 16% increase over the 2019 evaluation (67%; Fig. 10a). Similarly, there was a 14% increase between the 2019 (58%) and 2021 (72%) evaluations for those participants that provided a "good" or better rating when asked about the NMDA's ability to track an individual mesocyclone (Fig. 10b). While the evaluation ratings showed modest increases between experiment iterations, it shows that the additional improvements to the NMDA trended in the right direction. Additionally, no participants found the detecting and tracking performance to be "poor" or "very poor," revealing that even in difficult environments, the NMDA still performed to acceptable standards.

Along with their performance ratings, participants were also asked to provide a brief explanation regarding their

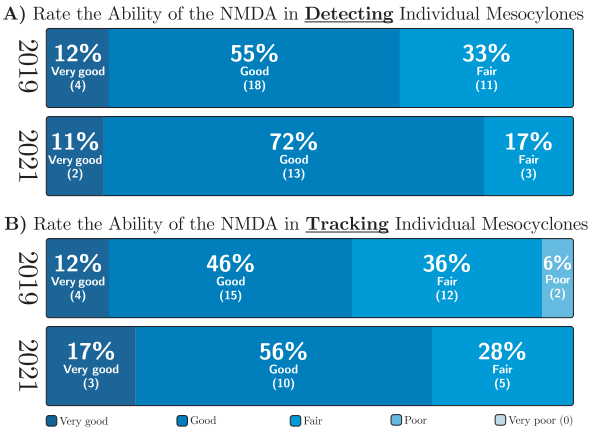

FIG. 10. Weekly survey results of the participants' perception of the NMDA's ability to (a) detect and (b) track mesocyclones from the 2019 to 2021 HWT evaluations ($N = 33$ and 18, respectively). Note that there were no "very poor" ratings for either detecting or tracking in the 2019 and 2021 HWT evaluations. In addition, the only ratings of "poor" (two total) were for tracking in the 2019 HWT evaluation.

choice for the evaluation rating. Even though 83% and 72% of participants provided a "good" or better rating for the NMDA's detecting and tracking abilities, respectively, only 11% (detecting) and 17% (tracking) of those were associated with the top rating of "very good." Examining the participants' justifications for ratings of "good" or lower, many mentioned false detections as the contributing factor to their specific rating. When asked to rate the NMDA's ability to handle false alarm detections in the end-of-week survey, 61% of participants stated that the NMDA performed "fair" overall. However, on a case-by-case level, this percentage would fluctuate dependent on the storm type due to the aforementioned detection generation difficulties outlined in the appendix. As a quick method to reduce false alarm detections, a week 1 participant, along with a few others, suggested that "*a lot of these [false detections] can be eliminated by hiding the lower end detections.*" Due to this, a user-adjustable detection-thresholding tool based on the detection's AzShear value (strength) was introduced during week 2 and tested during week 3 of the experiment that helped participants to reduce the number of false detections displayed. While the sample size of the end-of-week surveys is too low to notice any marked change in the participants perceptions toward the NMDA's handling of false detections, examining the surveys given at the end of every evaluation case shows slight improvements between week 1 and weeks 2–3 that are likely attributed to the thresholding tool. Utilizing the surveys from archived cases to provide a fair comparison between each week, there was a 10% reduction in ratings of "fair/poor" combined with a 10% increase of "good/very good" between weeks 1 and 2, something that was also mirrored between weeks 1 and 3 where a 22% decrease/increase of "fair/poor" and "good/very good" occurred, respectively. While this thresholding tool proved useful, reducing false detections internally in the algorithm prior to output is the most ideal solution to this problem.

2) Impact to warning decisions and other operational utilities

While the NMDA performs well in its intended purpose of detecting and tracking rotational features, it is extremely important to understand its usefulness in the operational warning environment. The goal of the NMDA is to assist end users with storm interrogation and provide additional confidence measures during the warning-decision process. Participants were asked how they would utilize the algorithm during their warning-issuance workflow and how this integration would impact their warning-decision process. Participants stated that their use of the NMDA converged on two methods during warning operations: 1) as a situational awareness tool and 2) a confidence builder when already considering a warning.

The NMDA helped as a situational awareness tool by alerting forecasters to strengthening storms and as a prioritization tool for deciding which storms required attention first, especially in complex situations or for those participants who were playing the role of a mesoscale analyst forecaster. Storm prioritization is an important skill for a warning forecaster and one that is aided by the NMDA, shown by one participant who stated "*I really liked overlaying NMDA detections from several radars, as this helped my situational awareness, and helped me prioritize storms in the SRAD [Screen, Rank, Analyze, Decide] process.*" When participants were asked if the NMDA made them consider warning a storm that they otherwise would not have noticed, eight out of 18 participants (44%) answered "yes," with a few of the participants stating this for cases occurring in low-end severe situations that can be challenging to the warning forecaster. An additional four participants responded with "maybe," stating that while the NMDA detections did not move them to consider warning a storm, the detections did draw their attention to a storm that they were not already watching. The NMDA is well suited in aiding in situational awareness during chaotic or mixed-mode storm environments, which according to several participants, would have an immediate operational impact to those serving as a mesoscale analyst forecaster.

When a participant was already considering issuing a warning, the NMDA acted, to varying degrees, as a warning-confidence builder. The amount of increased confidence differed depending on the storm environment. For more straightforward warning cases, such as those containing isolated supercells, the NMDA contributed less to their warning confidence since radar base data already confirmed their warning decision. The NMDA generally helped increase confidence for scenarios such as QLCS circulations or marginally severe storms, where the NMDA might highlight a feature that was not readily apparent to a participant in the radar base data. It should be noted that while the NMDA can build warning confidence, most participants stated they would not necessarily use it to directly issue a warning, instead relying heavily on their own warning experience and analysis of the base data to make those final decisions. As one participant noted, "*it is extremely difficult*

if not impossible to train a computer algorithm to see what the human eye is seeing with storm structure, etc.," highlighting the importance of the human element in severe weather forecasting. This suggests that, for certain aspects of storm analysis, experienced forecasters must rely on their own visual interpretation of the raw data.

At the end of each evaluation week, participants were asked how their confidence in utilizing the NMDA in the warning decision process changed relative to the start of the week. Using a 7-point confidence scale that ranged from "decreased considerably" to "increased considerably," 15 of the 18 participants stated that their confidence increased, several of which had "considerably" or "moderately" increased confidence. They attributed this confidence to the multitude of storm modes and environments that they evaluated the NMDA with over the course of the experiment. This allowed them to visually see the NMDA's successes and limitations, which enabled them to mentally calibrate to what the algorithm was displaying. Others acknowledged that their current understanding of the subpar performance of the operational MDA products helped to quickly instill confidence in the NMDA when they saw how well it performed in comparison.

3) IDENTIFIED LIMITATIONS AND METHODS TO ADDRESS THEM

False detections were the most impactful limitation of the NMDA that directly affected forecaster use of the algorithm. When participants provided lower scores for their perceived performance of the NMDA's detecting and tracking abilities, many of them listed these false detections as a contributing factor to their decision. In fact, when participants were asked if they could request one improvement for the NMDA, the leading performance-related improvement was to reduce false detections. These unwanted detections within the NMDA are caused by various means, including turbulent velocity fields and contaminated velocity data. While quality-control methods were in place for this particular prototype, the HWT evaluation illustrated the need for additional work to reduce false detections. These efforts focused on improving the techniques used in the generation of rotational interest objects within each individual tilt, as well as the vertical linking of these objects to build the final detections (see the appendix for an overview of the detection generation process). The largest reduction of false detections involved developing a technique that combines and re-evaluates preliminary rotational interest objects that occur along the radar radial, helping to reduce the false detections that occur along linear features, such as QLCSs. Another false detection reduction endeavor involved the complete reconstruction of the mechanism used to build the final detections. A switch to a more dynamic and computationally efficient approach in vertically linking rotational interest objects between neighboring tilts helped reduce incorrect linkages that could led to the creation of false detections. This was accomplished by examining whether a specific interest object on a higher tilt had multiple possible links from the neighboring lower tilt, and if so, comparing the object attributes from the lower tilt to determine the best match—a check that was not done in the previous vertical linking method. Additionally, the updated detection construction methodology enabled the removal of a previous restriction aimed to reduce false alarms that only allowed the base of a detection to begin within the first three tilts of a volume scan, hence permitting more midlevel mesocyclones to be properly detected by the NMDA, which was an additional performance-related improvement that was requested by several of the participants, with one stating that "*the single-radar objective analysis through the depth of the storm is the most valuable part of the algorithm.*"

In addition to reducing false detections, participants also noted that some sporadic detection tracking inconsistencies occasionally introduced errors into the detection's trend attributes. Participants considered this trend information to be an important component of the NMDA as it provides a sense on how rotational features are changing over time. These inconsistencies were found to be related to the tracking-derived forward-motion estimates of detections that would cause incorrect linking with downstream detections. This was corrected by relaxing or further constraining thresholds that were set to limit the distance and direction between linked detections, as well as checking for unrealistic attributes trends, such as storm motion. Overall, combining the improved detection tracking methodology with that of the refined detection building process worked to reduce the identified limitations of the NMDA associated with false detections and object tracking.

c. *Concurrent use of both algorithms by participants*

While the evaluations of both the NMDA and TORP were independent, many participants examined both simultaneously during case evaluations. This inadvertently led to discoveries on how participants used both algorithms together to help in their decision-making process. Some participants found concurrent use to be particularly useful, especially when detections overlapped, indicating that the feature of interest had both tornado potential (TORP) and vertical continuity (NMDA). This provided them additional confidence that the storm feature they were analyzing on radar during warning operations had merit and was unlikely to be caused by spurious radial velocity data.

While there were no apparent performance-related drawbacks to using both algorithms together, one participant did acknowledge that the slightly different color tables used to display a detection icon's strength (NMDA) or probability (TORP) caused some confusion if detections were overlaid. Their experience made them associate the color of the detection icon with the severity of the threat, and in their particular example, the color of the NMDA detection icon indicated a lower threat than that of the TORP detection icon, even though both detections were sampling the same feature. They acknowledged that they knew both algorithms were displaying different data but it caused them to pause and critically reason why, which could potentially delay them in their warning issuance process. Ultimately, this is an aspect that will have to be addressed in future training materials to make

users who plan to use the algorithms together aware of this possibility.

d. Algorithm visualizations within AWIPS-II

The new AWIPS-II visualization plugin, a major point of the evaluation during the 2021 HWT, received positive feedback from participants. Due to the simple and easy-to-use format that efficiently utilized AWIPS-II screen space, participants stated that it was a large improvement over the long-standing visualizations used by the TDA and MDA. The mouse-over detection attribute list was welcomed by participants because it provided detection information on-demand; however, proper organizational listing of the attributes was deemed a necessity to facilitate quick use of the algorithm in a warning environment. For those attributes that contained trend information and climatology-derived categorical severity labels, participants deemed them useful, but would prefer raw climatological percentile values in addition to the categorical levels, and requested long-term trend information, preferably in the form of a trend graph.

Many participants suggested changes to enhance the usability of the visualization tool for forecasters in an operational environment. During the first week, participants suggested two features: 1) the ability to customize the list of detection attributes displayed via the mouse-over feature, as well as 2) an adjustable thresholding slider to filter the detections based on tornado probability (TORP; Fig. 8) or AzShear strength (NMDA). Both suggestions came from a desire to only view detections and attributes that were important to the forecaster's specific needs. One forecaster stated that "*[they] likely would not use the NMDA or would only use it sparingly*" due to "*the organization of the readout and the false alarm detections*" when asked whether they would use the NMDA in operations as it was presented during the HWT, showing that they felt that these particular improvements were necessary for them to use the algorithm.

The HWT technical staff were able to implement these improvements for participants to fully review starting on Thursday during the second of the experiment. The customizable attribute list was welcomed by participants, but they thought that too many attributes were displayed by default, forcing the end user to adjust the list prior to algorithm use. Participants stated they would prefer to start with a simplified list of the commonly used attributes that could be adjusted as needed. During the third week of the experiment, the default list of output variables was reduced, while the attributes available for display with TORP detections were expanded. This expansion included all of the predictors that showed statistical differences between tornadic and non-tornadic populations. The thresholding slider was also met with wide acclaim as it gave participants the ability to threshold which detections were displayed, prompting survey feedback such as "*the slider was a very useful addition*" and that they "*love[d] the slider bar.*" Some participants commented on the slider's effectiveness in filtering out false alarm detections in the NMDA, stating that while there were "*still a few false alarms around areas of rotation*," they found that increasing the threshold was "*very beneficial at reducing these false signatures.*"

The continuous evolution of the visualization package during the course of the experiment allowed PIs to obtain near real-time feedback on how these changes performed in the pseudo-operational environment of the HWT. By the end of the experiment, PIs had a targeted list of improvements for the visualization plugin. Minor technical improvements included time matching of the products with other data sources and differentiating between periods when data are not available and those when no detections are present due to a lack of rotating storms. Other larger improvements centered around increased end-user customization that would provide additional utility to the visualization. One customization included optional trailing tracks attached to the detection icons that are shaded by the historical values of the object's probability (TORP) or strength (NMDA; Fig. 11). The length of the tracks can be adjusted based on the number of detections or time duration, allowing users to see both trends and the past track of each detection. Other customization features included a sortable detection attribute list, letting the users fully customize how to display the text output of the products, and the ability to share algorithm settings between multiple AWIPS-II panes.

While AWIPS-II provides a robust and expandable platform to visualize TORP and NMDA, operational end users such as those within the DoD do not currently have access to AWIPS-II while performing their duties. To ensure that these important members of the operational forecaster community have access to these new and powerful tools, those responsible for the operational transition of these algorithms will need to develop alternative methods to display TORP and NMDA on non-AWIPS-II third-party radar data visualization platforms, such as the Gibson Ridge (Gibson Ridge Software LLC 2023) suite of visualization software that was suggested by the DoD participants, who attended both the 2019 and 2021 HWT.

6. Summary and conclusions

In summary, the virtual 2021 HWT EWP RCA experiment was considered a success by both PIs, HWT technical staff, and participants. Through verbal discussions and surveys that were conducted throughout the week, the participants conveyed their satisfaction with the virtual experiment experience. The quality and details of the feedback obtained during the experiment has played a crucial role in the continued development of the algorithms. TORP and NMDA received positive feedback from the participants overall, with many hoping to see them available to the operational field in the near future. Several forecasters stated that the new algorithms exceeded their expectations based on their previous experience with the TDA and MDA, emphasizing that the upgrades are sorely needed. In the operational environment, both algorithms scored high on usefulness as a situational awareness tool and as a confidence builder in severe warning operations. During tornado warning operations, TORP was able to provide additional details that aided quick decision making

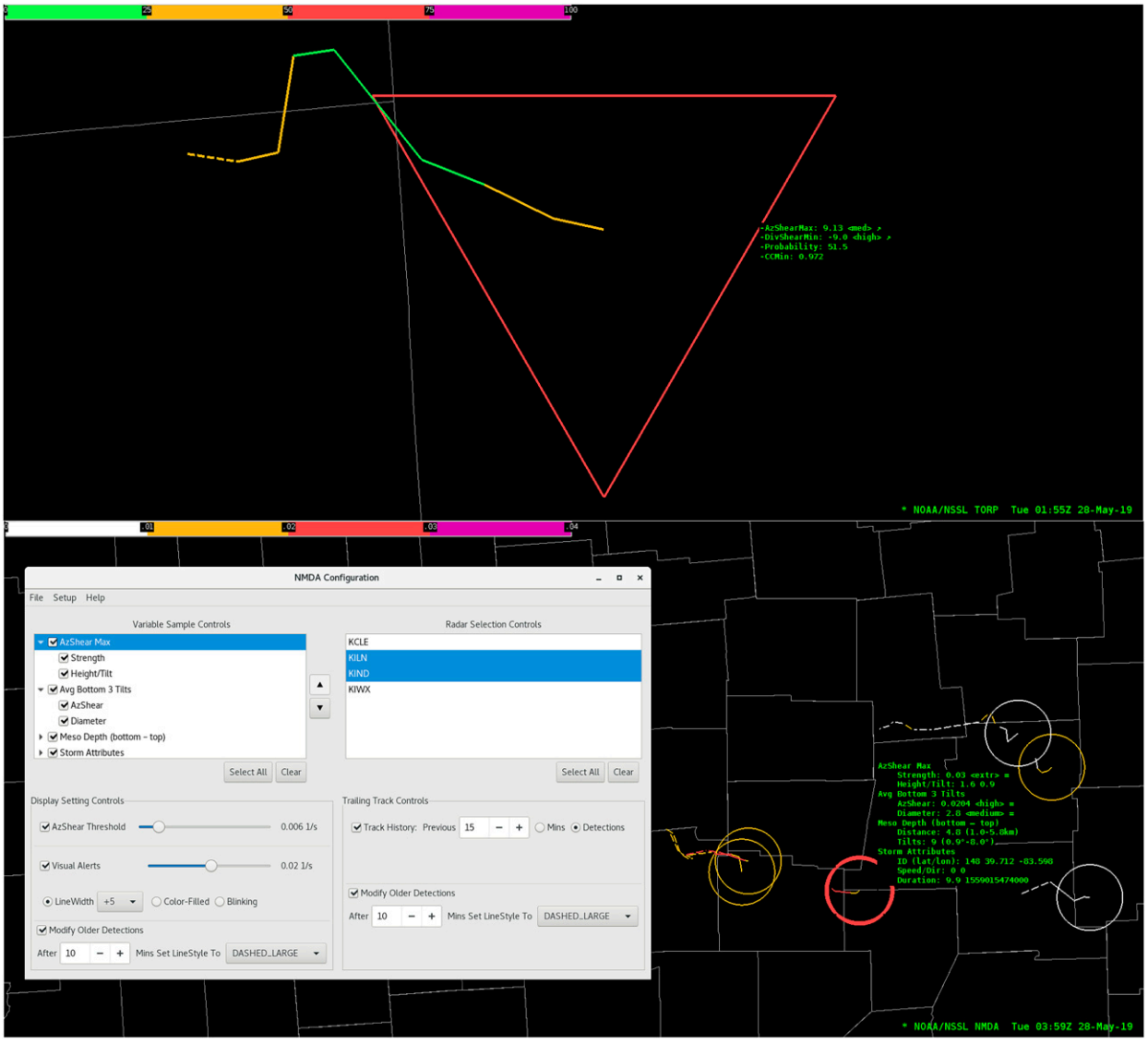

FIG. 11. Example of the new visualization features added to the prototype TORP/NMDA AWIPS-II plugin following experiment feedback. (top) A TORP detection with a track shaded by the previous probabilities associated with the detection, with the dashed line style signifying tracks that are older than a set time threshold, which in this example is 10 min. (bottom) NMDA detections with a new menu prototype, which lets the user adjust a multitude of different parameters, such as showing detections from multiple radars, enabling custom sorting of the readout variables, and setting visual alerts for detections exceeding a certain AzShear threshold.

within the warning-issuance process, with several forecasters stating that TORP provided nearly everything they would want in a tornado detection tool. Both algorithms can provide a concise summary of various single-radar products, enabling the end-user to relate the information to the convective warning process by pinpointing areas of rotation, something that may not be detected by operational tools that are grid-based or focus on larger reflectivity-based storm objects. The completely new AWIPS-II visualization package for the algorithms also received high marks from the participants who were especially pleased with the ability to customize and filter the detections.

With some additional adjustments, most participants felt that the algorithms were very close to achieving a final operational readiness level. These suggested adjustments were mainly centered on the continued reduction of false detections, most of which have been addressed since the conclusion of the experiment. The feedback received by project PIs aided in the advancement of the R2O process for both TORP and NMDA, enabling both algorithms the necessary information to progress to the point of being ready for operational transition. Discussions between project PIs and the ROC regarding the operational transition are currently underway. Many of the suggested features for both the algorithms and

the AWIPS-II visualization were incorporated, further improving their utility for the end user. In addition to the AWIPS-II product that was tested in this experiment, TORP has also been integrated into another HWT EWP experiment that involves the Forecasting a Continuum of Environmental Threats (FACETs) program (Rothfusz et al. 2018), where the algorithm provides tornado probability guidance for the forecaster creation of Probabilistic Hazards Information (PHI) in an experimental version of Hazard Services (Argyle et al. 2017; Calhoun et al. 2021).

Relating to the structure of the experiment, the project PIs and HWT technical staff learned a great deal about conducting a virtual HWT experiment utilizing AWIPS-II remotely via a cloud platform, ensuring future virtual HWT experiments can successfully operate. While AWIPS-II operated well in the AWS cloud, the various types of hardware and connection locations (e.g., home or office) used by participants did cause some problems, such as firewalls preventing entry to the cloud instance or the participants' hardware not allowing the cloud instance to display properly. While these were largely able to be addressed quickly and efficiently, in an effort to save time and keep the experiment running smoothly, it is suggested to work with participants prior to their HWT week to test the Internet connection and hardware capabilities that they will be using to remotely attend. Beyond the technical aspects associated with a virtual HWT, project PIs need to consider the breadth of various time zones that their participants could reside in and generate their daily HWT operations schedule accordingly. This will likely shorten the daily operating hours of the experiment, which can take away valuable real-time operating periods in the late afternoon and evening (typically the peak time for severe weather) and shorten the amount of time for training and learning at the beginning of the experiment. With this in mind, it is especially important to have any necessary training and information prepared and available to the participants well in advance of the start of their experiment week. Despite some of these challenges that were encountered as a consequence of being fully virtual, the participants were impressed by the efficiency of the virtual experiment. End-of-week surveys indicated that 100% of the participants would "definitely" or "probably" participate in a future HWT experiment if it required forecasters to participate virtually, and would recommend the experience to a fellow coworker. The success of this particular virtual HWT framework validates that virtual participation can be effective, potentially expanding the availability of the HWT to a broader set of participants.

Overall, this experiment continued the long-standing history of communication between researchers and the operational community through the use of testbeds, where the intended end users can explore new tools and provide feedback during the R2O process. This type of interaction between developers and users can help identify operational needs that might not be discovered otherwise, likely leading to the creation of more robust and useful products. Virtual evaluations of tools like these could also be imitated in a less formal setting, such as remote demonstrations with NWS Weather Forecast Offices, to showcase tools to local forecasters during real-time severe weather, providing new avenues of communication to further increase accessibility to prototype tools and techniques that are destined for operations. The authors of this publication hope the findings of this HWT experiment will serve to guide future single-radar algorithm development in considering key aspects to focus on that will impact their usefulness in the operational warning environment.

Acknowledgments. We thank all of the forecasters participating in the virtual 2021 HWT EWP RCA experiment. The authors appreciate the feedback these participants provided to help us improve our products. We thank Adam J. Clark, who provided an internal review of this paper, as well as Katie Magee and two other anonymous reviewers for their feedback. We also acknowledge Donald W. Burgess and Kodi Berry for their support. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA21OAR4320204, U.S. Department of Commerce. Support was also provided by the NOAA/NWS Radar Operations Center through the technology transfer memorandum of understanding agreement.

Data availability statement. The nonidentifiable data collected in this experiment are available upon request. The blog post feedback is publicly available (https://inside.nssl.noaa.gov/ewp/). These posts are written by forecaster participants using pseudonyms and were posted by PI accounts.

APPENDIX

NMDA Algorithm Description

In overview, the NMDA operates using the following method:

1) *Find preliminary interest areas for each individual tilt*—As each radar tilt is received, the algorithm overlays the input products (Fig. A1) for that tilt and determines the areas of rotational interest, which are refined to keep only the locations of peak AzShear values. Each peak undergoes a quality-control process that uses reflectivity, spectrum width, and a velocity-derived total-shear LLSD gradient (TotalShear) to help reduce false detections. This process is performed on all tilts received by the algorithm.
2) *Build detections by combining interest areas from multiple tilts*—Once a triggering event occurs (a specific radar tilt is ingested or the end of the volume is reached), the areas of peak rotation are compared between neighboring tilts. If their attributes meet certain criteria, they are joined to vertically build detections. Detection building is completed once all available tilts and their associated rotational peaks have been exhausted.
3) *Tracking detections temporally and spatially*—After two or more occurrences of detection building, the NMDA will attempt to link these detections over space and time, allowing trend information to be generated (i.e., change

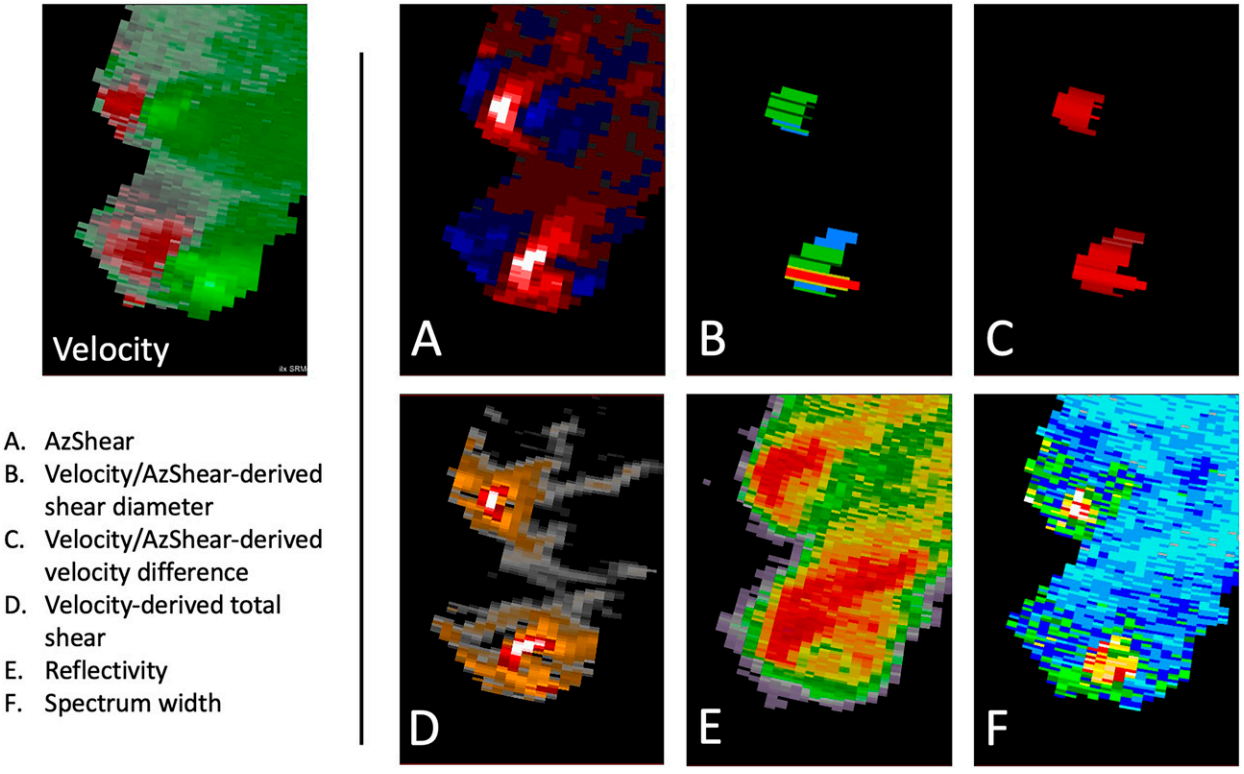

FIG. A1. Input products for the NMDA (the velocity field is provided for context).

in strength, traveling speed). A model-derived sounding table, if available, is used to calculate an approximate mean storm motion that initially is used to link downstream detections. Once two or more detections have been linked, tracking gradually transitions to a completely object-based approach. If multiple detections are found downstream, detection attributes are compared to link the detections that most closely match.

As with any radar-based algorithm, there are strengths and weaknesses associated with the NMDA's detecting ability, largely owing to storm type and quality of the input radar data. The NMDA excels with isolated storm environments, such as those associated with supercells and ordinary cell convection, and in certain linear situations (e.g., QLCS) in which the linear feature is more perpendicular to the beam. The NMDA encounters degraded performance, largely in the form of false detections, in linear storm environments where the linear feature is parallel to the radar beam and occasionally in postfrontal stratiform precipitation regions where turbulent wind fields are present. Recent advancements to the algorithm, which are discussed in this publication, have served to reduce false detections in these situations to improve overall algorithm performance. In general, NMDA performance will likely be affected if any of the input products are affected by contaminated data, such as with hail contamination or terrain effects.

REFERENCES

Argyle, E. M., J. J. Gourley, Z. L. Flamig, T. Hansen, and K. Manross, 2017: Toward a user-centered design of a weather forecasting decision-support tool. *Bull. Amer. Meteor. Soc.*, **98**, 373–382, https://doi.org/10.1175/BAMS-D-16-0031.1.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Brotzge, J. A., 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, https://doi.org/10.1175/WAF-D-12-00119.1.

Calhoun, K. M., K. L. Berry, D. M. Kingfield, T. Meyer, M. J. Krocak, T. M. Smith, G. Stumpf, and A. Gerard, 2021: The experimental warning program of NOAA's Hazardous Weather Testbed. *Bull. Amer. Meteor. Soc.*, **102**, E2229–E2246, https://doi.org/10.1175/BAMS-D-21-0017.1.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, and A. K. Heidinger, 2013: Evolution of severe and nonsevere convection inferred from GOES-derived cloud properties. *J. Appl. Meteor. Climatol.*, **52**, 2009–2023, https://doi.org/10.1175/JAMC-D-12-0330.1.

——, and Coauthors, 2018: The NOAA/CIMSS ProbSevere model: Incorporation of total lightning and validation. *Wea. Forecasting*, **33**, 331–345, https://doi.org/10.1175/WAF-D-17-0099.1.

——, M. J. Pavolonis, J. M. Sieglaff, L. Cronce, and J. Brunner, 2020: NOAA ProbSevere v2.0—ProbHail, ProbWind, and ProbTor. *Wea. Forecasting*, **35**, 1523–1543, https://doi.org/10.1175/WAF-D-19-0242.1.

Clark, A. J., and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, https://doi.org/10.1175/BAMS-D-11-00040.1.

——, and Coauthors, 2021: A real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bull. Amer. Meteor. Soc.*, **102**, E814–E816, https://doi.org/10.1175/BAMS-D-20-0268.1.

Collaborative Institutional Training Initiative, 2023: CITI program. CITI, accessed 28 April 2023, https://about.citiprogram.org/.

Desrochers, P. R., and R. J. Donaldson Jr., 1992: Automatic tornado prediction with an improved mesocyclone-detection algorithm. *Wea. Forecasting*, **7**, 373–388, https://doi.org/10.1175/1520-0434(1992)007<0373:ATPWAI>2.0.CO;2.

Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1.

Gibson Ridge Software LLC, 2023: GRLevelX. GRLevelX, accessed 24 April 2023, http://www.grlevelx.com/.

Hennington, L. D., and D. W. Burgess, 1981: Automatic recognition of mesocyclones from single-Doppler radar data. *20th Conf. on Radar Meteorology*, Boston, MA, Amer. Meteor. Soc., 704–706.

Jones, T. A., K. M. McGrath, and J. T. Snow, 2004: Association between NSSL mesocyclone detection algorithm-detected vortices and tornadoes. *Wea. Forecasting*, **19**, 872–890, https://doi.org/10.1175/1520-0434(2004)019<0872:ABNMDA>2.0.CO;2.

Kain, J. S., P. R. Janish, S. J. Weiss, M. E. Baldwin, R. S. Schneider, and H. E. Brooks, 2003: Collaboration between forecasters and research scientists at the NSSL and SPC: The spring program. *Bull. Amer. Meteor. Soc.*, **84**, 1797–1806, https://doi.org/10.1175/BAMS-84-12-1797.

Kingfield, D. M., and M. M. French, 2022: The influence of WSR-88D intra-volume scanning strategies on thunderstorm observations and warnings in the dual-polarization radar era: 2011–20. *Wea. Forecasting*, **37**, 283–301, https://doi.org/10.1175/WAF-D-21-0127.1.

Mahalik, M. C., B. R. Smith, K. L. Elmore, D. M. Kingfield, K. L. Ortega, and T. M. Smith, 2019: Estimates of gradients in radar moments using a linear least squares derivative technique. *Wea. Forecasting*, **34**, 415–434, https://doi.org/10.1175/WAF-D-18-0095.1.

McGrath, K. M., T. A. Jones, and J. T. Snow, 2002: Increasing the usefulness of a mesocyclone climatology. *21st Conf. on Severe Local Storms*, San Antonio, TX, Amer. Meteor. Soc., 5.4, https://ams.confex.com/ams/pdfpapers/44897.pdf.

Mitchell, E. D. W., S. V. Vasiloff, G. J. Stumpf, A. Witt, M. D. Eilts, J. T. Johnson, and K. W. Thomas, 1998: The National Severe Storms Laboratory tornado detection algorithm. *Wea. Forecasting*, **13**, 352–366, https://doi.org/10.1175/1520-0434(1998)013<0352:TNSSLT>2.0.CO;2.

Mitchell, E. D., K. L. Elmore, K. Angle, C. Hannon, and N. J. Eckstein, 2000: A radar signature climatology using WSR-88D level II data. *20th Conf. on Severe Local Storms*, Orlando, FL, Amer. Meteor. Soc., P3.7, https://ams.confex.com/ams/Sept2000/techprogram/paper_16416.htm.

NOAA/NWS/ROC, 1991: NOAA Next Generation Radar (NEXRAD) Level II base data. NOAA/National Centers for Environmental Information, accessed May 2019 to April 2020, https://doi.org/10.7289/V5W9574V.

Rothfusz, L. P., 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.

Ryzhkov, A. V., T. J. Schuur, D. W. Burgess, and D. S. Zrnić, 2005: Polarimetric tornado detection. *J. Appl. Meteor.*, **44**, 557–570, https://doi.org/10.1175/JAM2235.1.

Sandmæl, T. N., and Coauthors, 2023: The tornado probability algorithm: A probabilistic machine learning tornadic circulation detection algorithm. *Wea. Forecasting*, **38**, 445–466, https://doi.org/10.1175/WAF-D-22-0123.1.

Scharfenberg, K. A., and Coauthors, 2005: The Joint Polarization Experiment: Polarimetric radar in forecasting and warning decision-making. *Wea. Forecasting*, **20**, 775–788, https://doi.org/10.1175/WAF881.1.

Serafin, R. J., A. E. MacDonald, and R. L. Gall, 2002: Transition of weather research to operations: Opportunities and challenges. *Bull. Amer. Meteor. Soc.*, **83**, 377–392, https://doi.org/10.1175/1520-0477(2002)083<0377:TOWRTO>2.3.CO;2.

Stumpf, G. J., A. Witt, E. D. Mitchell, P. L. Spencer, J. T. Johnson, M. D. Eilts, K. W. Thomas, and D. W. Burgess, 1998: The National Severe Storms Laboratory mesocyclone detection algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 304–326, https://doi.org/10.1175/1520-0434(1998)013<0304:TNSSLM>2.0.CO;2.

Torres, S. M., and C. D. Curtis, 2007: Initial implementation of super-resolution data on the NEXRAD network. *21st Int. Conf. on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology*, San Antonio, TX, Amer. Meteor. Soc., 5B.10, https://ams.confex.com/ams/pdfpapers/116240.pdf.

Warning Decision Training Division, 2022a: Radar & Applications Course (RAC)—Base and derived products—Introduction to base and derived products. Warning Decision Training Division, accessed 10 January 2023, https://training.weather.gov/wdtd/courses/rac/products/intro/story.html.

——, 2022b: Radar & Applications Course (RAC)—Base and derived products—Mesocyclone (MD) and digital mesocyclone (DMD). Warning Decision Training Division, accessed 7 March 2023, https://training.weather.gov/wdtd/courses/rac/products/md-dmd/presentation_html5.html.

Wieler, J. G., 1986: Real-time automated detection of mesocyclones and tornadic vortex signatures. *J. Atmos. Oceanic Technol.*, **3**, 98–113, https://doi.org/10.1175/1520-0426(1986)003<0098:RTADOM>2.0.CO;2.

Zrnić, D. S., L. D. Hennington, and J. Skelton, 1982: Automatic recognition of mesocyclones from single Doppler radar data. NOAA Tech. Rep. AFGL-TR-82-0291, 52 pp., https://apps.dtic.mil/sti/pdfs/ADA125854.pdf.

——, D. W. Burgess, and L. D. Hennington, 1985: Automatic detection of mesocyclonic shear with Doppler radar. *J. Atmos. Oceanic Technol.*, **2**, 425–438, https://doi.org/10.1175/1520-0426(1985)002<0425:ADOMSW>2.0.CO;2.