Development of a Human–Machine Mix for Forecasting Severe Convective Events

CHRISTOPHER D. KARSTENS,[a,b,c] JAMES CORREIA JR.,[a,c] DAPHNE S. LADUE,[d] JONATHAN WOLFE,[e] TIFFANY C. MEYER,[a,b] DAVID R. HARRISON,[f,a,b] JOHN L. CINTINEO,[g] KRISTIN M. CALHOUN,[a,b] TRAVIS M. SMITH,[a,b] ALAN E. GERARD,[b] AND LANS P. ROTHFUSZ[b]

[a] *Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
[c] *NOAA/NWS/Storm Prediction Center, Norman, Oklahoma*
[d] *Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma*
[e] *NOAA/NWS/WFO Duluth, Duluth, Minnesota*
[f] *University of Oklahoma, Norman, Oklahoma*
[g] *Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin–Madison, Madison, Wisconsin*

(Manuscript received 20 December 2017, in final form 14 February 2018)

ABSTRACT

Providing advance warning for impending severe convective weather events (i.e., tornadoes, hail, wind) fundamentally requires an ability to predict and/or detect these hazards and subsequently communicate their potential threat in real time. The National Weather Service (NWS) provides advance warning for severe convective weather through the issuance of tornado and severe thunderstorm warnings, a system that has remained relatively unchanged for approximately the past 65 years. Forecasting a Continuum of Environmental Threats (FACETs) proposes a reinvention of this system, transitioning from a deterministic product-centric paradigm to one based on probabilistic hazard information (PHI) for hazardous weather events. Four years of iterative development and rapid prototyping in the National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT) with NWS forecasters and partners has yielded insights into this new paradigm by discovering efficient ways to generate, inform, and utilize a continuous flow of information through the development of a human–machine mix. Forecasters conditionally used automated object-based guidance within four levels of automation to issue deterministic products containing PHI. Forecasters accomplished this task in a timely manner while focusing on communication and conveying forecast confidence, elements considered necessary by emergency managers. Observed annual increases in the usage of first-guess probabilistic guidance by forecasters were related to improvements made to the prototyped software, guidance, and techniques. However, increasing usage of automation requires improvements in guidance, data integration, and data visualization to garner trust more effectively. Additional opportunities exist to address limitations in procedures for motion derivation and geospatial mapping of subjective probability.

1. Introduction

For several decades, the National Weather Service (NWS) has invested in the development and maintenance of various technologies that assist forecasters tasked with making real-time warning decisions for severe convective hazards. Perhaps the most prominent of these technologies is the Weather Surveillance Radar (WSR) network (e.g., Polger et al. 1994; Torres and Curtis 2007; Istok et al. 2009; Daniel et al. 2014). The current Weather Surveillance Radar 1988-Doppler (WSR-88D) network offers many capabilities, including Doppler and reflectivity products for detecting individual hazards and for identifying storm mode, rapid volumetric scanning strategies for improving detection, and dual-polarization products for providing complementary evidence. In addition to real-time applications such as warning decision-making, the aggregation of cases identified from the WSR-88D network, combined with contextual information (e.g., environment), can lead to the development of or improvements to existing conceptual models (e.g., Moller et al. 1994). However, the application of conceptual models relies on human interpretation for severe convective weather detection, either through the use of patterns learned (i.e., heuristics

Corresponding author: Christopher D. Karstens, chris.karstens@noaa.gov

DOI: 10.1175/WAF-D-17-0188.1

or intuition) and/or by following a stepwise analytical procedure that is often acquired from training.

The proliferation of computational technology in the 1990s enabled a real-time codification of analytical procedures to work in conjunction with the WSR-88D network (i.e., algorithms), with the goal of rapid detection of radar signatures to assist in warning decision-making. Development of single-site radar algorithms ensued, such as the Hail Detection Algorithm (HDA; Witt et al. 1998), the Mesocyclone Detection Algorithm (MDA; Stumpf et al. 1998), and the Tornado Detection Algorithm (TDA; Mitchell et al. 1998). These algorithm development efforts quickly evolved to include multisensor-enabled algorithms using four-dimensional radar data (Lakshmanan et al. 2007) with the focus shifting to automated detection of severe weather phenomena. This Multi-Radar Multi-Sensor (MRMS) system (Smith et al. 2016) has been combined with satellite and environmental information to drive artificial intelligence applications (e.g., Cintineo et al. 2014; McGovern et al. 2017; Gagne et al. 2017) and emerging convection-allowing model ensemble systems (e.g., Stensrud et al. 2009; Wheatley et al. 2015) that aim to predict the likelihood of severe convective weather occurrence, all with the goal of enhancing forecaster situational awareness and extending warning lead time.

With the continued emergence of analytical and predictive techniques, how can all of the aforementioned information be utilized by forecasters, particularly within the time constraints associated with warning decisions? The Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2014) project proposes a reinvention of the current NWS watch/warning system, transitioning from a deterministic product-centric paradigm to one that utilizes the previously discussed forms of probabilistic guidance to provide NWS forecasters and partners with probabilistic hazard information (PHI; Karstens et al. 2015). Karstens et al. (2015) found that automated PHI guidance has the potential to produce more reliable probabilistic forecasts with less false-alarm area, but with reduced verification metrics, compared to human-generated forecasts. In addition, forecasters issued forecasts in a reasonable and timely manner and could save more time by using the automated guidance as a first guess. Thus, a PHI approach to forecasting severe convective events likely requires a human–machine mix (Snellman 1977; Moller et al. 1994; NRC 2014) in which guidance is partially or completely used to generate and/or maintain a forecast, to both assist and augment forecasting of all potential hazard areas regardless of severity (i.e., exceeding and falling below warning thresholds).

In practice, however, algorithms and other automated predictive techniques lack forecaster trust (Hoffman et al. 2013), a perception derived from factors such as erroneous detections (Andra et al. 2002), competitive dynamics (Stuart et al. 2006), and misuse or overreliance (Klein 2000) in place of basic subjective analysis (e.g., radar interrogation and conceptual understanding; Wilson et al. 2017). Within deterministic NWS warning operations, such techniques traditionally serve as ''safety nets,'' cuing the forecaster to reengage in subjective analysis, when time permits (Andra et al. 2002). However, these perceptions and strategies can limit the potential that automated technologies offer (NRC 2014), particularly in rapidly changing convective situations with high severity and/or coverage that can reduce or eliminate opportunities for maintaining subjective analysis (Bosart 1989; Brooks et al. 1992). Thus, groundwork for adapting a human–machine mix approach to warning applications is needed.

The purpose of this paper is to build on the efforts of Karstens et al. (2015) through the exploration of a human–machine mix as applied to severe convective events. This paper summarizes the iterative progress and discoveries from four annual testing cycles, yielding several insights into what a baseline probability-based warning paradigm could look like. Section 2 provides information about the chronological sequence and design of annual National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT) testing cycles. Detailed qualitative explanations and quantitative results of the human–machine mix development are given in sections 3 and 4, respectively, followed by a summary and discussion including future work in section 5.

2. Testing cycles to facilitate a human–machine mix system design

a. *Practitioner's cycles methodology*

Expanding on the previous findings from Karstens et al. (2015), this paper describes what is now four annually occurring tests (sometimes elastically referred to as ''experiments''), or practitioner's cycles (Hoffman et al. 2010), in the HWT that brought in one (2015), then two (2016–17), key user groups. Practitioner's cycles are derived from empirical data on actual software development failures. Accounting for the continuous and highly interacting factors in complex macrocognitive work, these practitioner's cycles situate prototypes in the workplace whenever possible. Because this work pertains specifically to severe convective events, the next best strategy of creating a naturalistic environment

for an envisioned world problem in the laboratory (i.e., HWT) was used (Crandall et al. 2006; Kahneman and Klein 2009). Realism in the testing environment is critical for these practitioner cycles to be capable of directly addressing and mitigating many of the factors that create the ''valley of death'' (NAS 2000), wherein a concept is unable to transition from technology readiness level 3 (proof of concept) to readiness level 7 (sufficiently mature prototype to be tested in an operational environment; Deal and Hoffman 2010).

In 2014, it was immediately apparent that exploration of the advantages and disadvantages of generating rapidly updating, hazard-specific PHI on storm hazards required the presence of key user groups. The project thus appropriately increased in complexity, leveraging funds from several projects. However, the goals of these annual testing cycles were mutual: to test ways in which a variety of forms of automated guidance (Table 1) could be used by NWS forecasters and to identify the optimal modes of operation in conjunction with NWS partners to generate useful, usable, and understandable information. This work began in 2014 at practitioner cycle three: users interacting with a functioning prototype, revealing interface issues, usability issues, brittleness, integration issues, and additional desirements (Hoffman et al. 2010). The second year (2015) added one key user group, emergency managers (EMs), and the final two years (2016 and 2017) added broadcast meteorologists in order to assist in identification of optimal modes of operation. Complexity was maximized in 2016, when actions and decisions from every participant group were interjected into the other groups as they would be in real life.

This strategy yields high external validity, which then situates other aspects of this project in the initial, ''exploration and discovery'' phase of fundamental scientific processes necessary for a methodical entry into a long-term scientific endeavor (e.g., UCMP 2017). In other words, this design has permitted and enabled the generation of many hypotheses, only a few of which are suggested herein. Those hypotheses can be subsequently studied in isolation for high internal validity research efforts that focus on specific issues such as how to best convey probability to an emergency manager. However, a return to complex, high external validity designs during system design completion is prudent to assure this project successfully crosses the valley of death to technology readiness level 7.

b. Participant activities and researcher tools

Each year consisted of a three-week testing cycle held during the afternoon and evening hours in the peak of the severe convective weather season. Shifts on days 1–4 were spent working one displaced real-time event (meaning an archived event was timed to appear as though the event were occurring in real time) and one real-time event. Day 1 served mainly as a learning day, ending with a brief period of independent use of the tools. The selection of real-time events was sensitive to the potential for severe convective activity on a given day, whereas the displaced real-time case selection considered a variety of research needs, including convective modes, coverage, severity, and evolution for the mutual benefit of independent subprojects (Fig. 1; Table 1).

NWS forecaster participants issued forecasts for tornadoes (2014–17), combined severe thunderstorm hazards (wind and hail; 2015–17), and cloud-to-ground lightning (2014, 2016, and 2017; Table 1). Forecasters were provided with and able to use experimental probabilistic guidance to assist in ascertaining forecast confidence of hazard occurrence while evaluating the utility of the guidance; they also analyzed standard operational data feeds using the Advanced Weather Interactive Processing System, version 2 (AWIPS II). As previously mentioned, user participants representing the NWS partner community, including EMs (LaDue et al. 2016, 2017) and broadcast meteorologists (Obermeier et al. 2017; Nemunaitis-Berry et al. 2017), were incorporated into the testing cycles beginning in 2015 and 2016, respectively. During each event, all participants were encouraged to ''think aloud'' (Ericsson and Simon 1993) while researchers took notes and recorded their verbalizations (via interview protocols). Additional recordings (via screen capture videos, camcorders, passive software logging) captured the actions of NWS forecasters creating, issuing, and updating manual and partially automated probabilistic forecasts (Bosart 1989) for PHI objects that spatially denote the extent of a severe convective weather event (i.e., hazard) and temporally denote the projected movement of a hazard area using a web-based prototype (Karstens et al. 2015).

Forecasts of PHI generated by NWS forecasters were converted from object-based probabilistic forecasts to grid-based probabilistic forecasts (i.e., PHI swaths) and displayed in an experimental version of the NWS Enhanced Data Display (EDD; Wolfe 2014). EMs and broadcast meteorologists used the experimental EDD to make simulated real-time decisions (e.g., activate sirens, preposition personnel, initiate programming cutins). Additionally, an internal NWS instant messaging program for real-time communication with Integrated Warning Team (IWT; e.g., Cavanaugh et al. 2016) partners (NWSChat; NOAA/NWS 2016), was used as a means to communicate information from EMs and

TABLE 1. Overview of testing cycles occurring in the HWT.

Year	Dates	Displaced real-time events	Participants	Object-based guidance	Object-based forecasts
2014	5–9 May 19–23 May 2–6 Jun	13 Jun 2012 [Dallas/Fort Worth, TX (FWD)] 5 Aug 2013 [Springfield, MO (SGF)]	6 forecasters	Hail (MRMS MESH)[a]	Tornado, hail, wind, lightning
2015	4–8 May	22 May 2010 [Aberdeen, SD (ABR)]	6 forecasters	Tornado (MRMS azimuthal shear)[b]	Tornado
	19–23 May 1–5 Jun	3 June 2014 [Omaha, NE (OAX)]	10 EMs	Severe (ProbSevere)[c]	Severe
2016	9–13 May	6 May 2015 [Norman, OK (OUN)]	9 forecasters	Tornado NEWS-e (NSSL WoF)[d] V_{rot} (SPC)[e]	Tornado
	23–27 May	31 Mar 2016 [Huntsville, AL (HUN)]	11 EMs	Severe (ProbSevere)[c] Storm duration[f] Damaging straight-line winds[g]	Severe
	6–10 Jun	24 Jun 2015 [Peachtree City, GA (FFC)]	3 broadcasters	Lightning (MRMS)[h]	Lightning
2017	8–12 May	25 May 2016 [Topeka, KS (TOP)]	9 forecasters (three returners: one from 2014 and two from 2016)	Severe (ProbSevere)[c] Storm duration[f] Storm classification[i] Damaging straight-line winds[g]	Tornado Severe Lightning
	22–26 May	9 May 2016 (OUN)	6 EMs		
	5–9 Jun	24 May 2016 [Dodge City, KS (DDC)] 1 Sep 2016 [Melbourne, FL (MLB)], lightning only	3 broadcasters	ProbHail and ProbWind[j] Real-time best track[k] Probability trend prediction[l] Tornado NEWS-e (NSSL WoF)[m] V_{rot} (SPC)[e] ProbTor[j] Lightning (MRMS)[n]	

[a] Karstens et al. (2015).
[b] Karstens et al. (2016).
[c] Cintineo et al. (2014).
[d] Correia et al. (2016).
[e] Thompson et al. (2017).
[f] McGovern et al. (2017).
[g] Lagerquist et al. (2017).
[h] Meyer et al. (2016).
[i] McGovern et al. (2018).
[j] Cintineo et al. (2018).
[k] Harrison et al. (2017).
[l] Harrison et al. (2018).
[m] Correia et al. (2018).
[n] Calhoun et al. (2018).

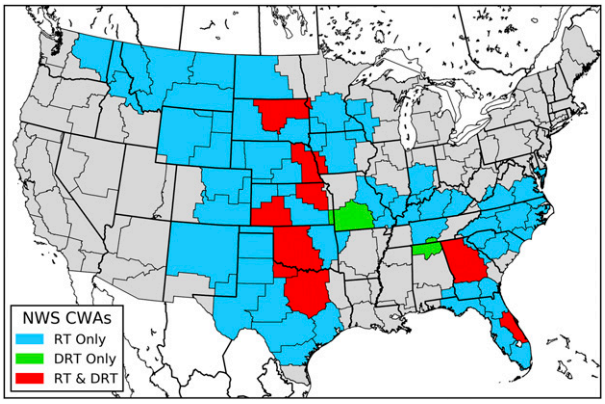

FIG. 1. Map of CWAs in which forecasters operated during the HWT PHI testing cycles (2014–17), denoted by real-time (RT; blue), displaced real-time (DRT; green), and both RT and DRT (red) events.

broadcast meteorologists to/from NWS forecasters via a warning coordinator. This simulated warning system grounded our work in observations rather than speculation on how such a system might work.

Immediately following each testing period, participants individually filled out a NASA Task Load Index (NASA-TLX; Hart and Staveland 1988) survey and a confidence continuum survey (adapted from Heinselman et al. 2015), followed by a verbal debrief or a recent case walkthrough (day 4 displaced real-time case in 2017 only; Militello and Hutton 1998; Hoffman 2005). Thereafter, the participants convened to form an IWT. Conducted as focus groups (Kamberelis and Dimitriadis 2005), these were moderated by a researcher and focused on a variety of themes, including recollection of working the event(s), elements of the forecast information (e.g., tools, probabilities, visualization, communication), methodological shortcomings, and ideas for generating systematic performance improvements. At the end of each week (day 5), participants filled out an end-of-week survey and convened for a final focus group activity to provide summary feedback and desirements (the "end" of the practitioner's cycle) for the software, similar input to experimental guidance, and expert direction on this potential future warning paradigm. Using this framework allowed the participants to have a role in contributing to and shaping a potential future warning system, helping to finish the system while under construction.

This study documents results from a subset of these instruments. Specifically, all statistics are derived from passive software logging of forecaster actions and one end-of-week survey question for EMs (discussed in section 4). The software statistics were reviewed annually in conjunction with the focus group discussion to inform development strategies for the next practitioner cycle (hereafter referred to as a testing cycle) the following year. In addition, evaluation of probabilistic forecasts is restricted to tornadoes and severe thunderstorms herein in an effort to draw from factors that were consistent in the three most recent testing cycles (using 2014 as a reference) and to draw some comparison with the severe convective warning paradigm currently employed by the NWS.

Inherent to these HWT testing cycles is the conundrum of drawing conclusions from a small sample size of participants. However, the aggregation of four annual testing cycles yielded 30 NWS forecaster and 27 EM participants, with hundreds of forecasts generated in each testing cycle. This study used recruitment tactics to deliberately introduce variation inherent of a larger sample size. Specifically, varied expertise was sought across the NWS regions and thus experience with hazards, policies, and demographics (Harrison and Karstens 2017). This approach brings the world into the laboratory, with all the real-world concerns, constraints, and opportunities, and forced all of us (researchers and participants) to deal with that complexity while, together, evaluating this new potential warning system. Note, three of the forecaster participants in 2017 (one per week) also participated in a previous testing cycle (one from 2014, two from 2016) to evaluate our development efforts and provide veteran insight. Emergency managers from a wide range of jurisdiction sizes and emergency support functions similarly helped broaden the representativeness of these results. That said, the findings of this study may not be generalized to all forecasters and emergency managers, but they are representative of a fairly broad constituency.

3. Qualitative evolution of a human–machine mix

a. Object-based guidance

To facilitate a human–machine mix for warning applications within the prototype system developed by Karstens et al. (2015), a consolidation of automated guidance into an object-based framework must occur. Key components of this process include object identification (i.e., vectorizing) of atmospheric processes (i.e., severe convective storms) from continuous fields with self-describing attributes (e.g., Lakshmanan et al. 2009), object tracking through time (e.g., Lakshmanan and Smith 2010; Lakshmanan et al. 2015), update frequency sufficient to resolve hazard evolution (e.g., LaDue et al. 2010; Heinselman et al. 2012; Wilson et al. 2017), and hazard prediction (e.g., Cintineo et al. 2014). Three forms of automated object-based guidance were incorporated and tested with forecasters, with efforts beginning in 2015 (Table 1).

Guidance for the creation of combined severe thunderstorm forecasts was based on the NOAA ProbSevere model (Cintineo et al. 2014) via objects identified from MRMS composite reflectivity with an update frequency of approximately 2 min (Fig. 2). The ProbSevere model combines data from Geostationary Operational Environmental Satellites (GOES; Menzel and Purdom 1994), the WSR-88D network, and the Rapid Refresh model (RAP; Benjamin et al. 2006) to probabilistically forecast the severe weather likelihood from a developing storm. The ProbSevere model evolved to incorporate lightning predictors and offers probabilistic predictions for individual severe thunderstorm hazards (hail, wind, and tornado; Table 1; Cintineo et al. 2018). ProbSevere guidance was also augmented with predictions of duration (McGovern et al. 2017), storm classification (McGovern et al. 2018), damaging straight-line winds (Lagerquist et al. 2017), and probability trend prediction (Harrison et al. 2018) using machine learning techniques. Additionally, a real-time best-track algorithm (Harrison et al. 2017, manuscript submitted to *Wea. Forecasting*, hereafter HKM) was implemented in 2017 to reduce the number of unjustified tracking breakage instances.

Guidance for the creation of tornado forecasts in 2015 (only) included objects identified from MRMS azimuthal shear, as discussed in Karstens et al. (2016). Beginning in 2016, guidance was made available from the National Severe Storms Laboratory (NSSL) Experimental Warn-on-Forecast System for Ensembles (NEWS-e; Wheatley et al. 2015) for identifying mid- and low-level storm rotation (Correia et al. 2016, 2018; McDonald and Correia 2016). Finally, guidance for lightning hazards in 2016 and 2017 included the probability of cloud-to-ground lightning prediction (Meyer et al. 2016; Calhoun et al. 2018) via objects identified from the MRMS reflectivity at −10°C with an update frequency of approximately 2 min (not discussed).

b. Levels of automation

Forecaster usage of the automated guidance followed a similar procedure as described in Karstens et al. (2015, section 2c, steps 1b–4b), but with some modifications. First, the display of the automated guidance changed from point markers to objectively identified diagnostic polygons (i.e., objects; Fig. 2a). Second, all first-guess forecast attributes could be optionally overridden by a forecaster, an effort that evolved throughout 2015–17 through iterative testing and evaluation (discussed in sections 3c and 4; Fig. 2b). The set of overrideable first-guess forecast attributes included the object geometry (shape and position), motion vector (speed and direction), motion uncertainty (speed and direction), probability trend, and discussion text.

After undergoing testing and evaluation with forecasters in 2015, four preferential modes of operation with the automated guidance were identified:

level 1, manual forecast;
level 2, forecaster geometry;
level 3, automated geometry; and
level 4, automated forecast.

For convenience, these modes are classified as four levels of automation, following a simplified linear classification method developed by Sheridan and Verplank (1978), ranging from completely manual (level 1; see Karstens et al. (2015) section 2c, steps 1a–6a) to completely automated forecast generation (level 4). Levels 2 and 3 are representative of a human–machine mix distinguished by forecaster override (level 2) or automated control (level 3) of the object geometry from automation.

Levels 2 and 3 are hereafter referred to as the human–machine levels of automation. In 2015, the delineation between these two levels was determined by whether or not a forecaster adjusted the object geometry on the map. After observing a lack of intuition with the initial functionality to make this delineation, a checkbox was added in 2016 enabling binary, nonsequential control, thus allowing the forecaster to toggle control of the object geometry to (checked) or from (unchecked) automation (Fig. 2b; "Auto" checkbox on Object Shape/Pos line). However, controls for adjusting the object shape [see Karstens et al. (2015) , their Fig. 4] were still present with the automated geometry enabled (level 3). In 2017 these controls were disassociated from level 3 of the automation process (Fig. 2b) after observing forecaster usage/goals in 2016, therefore limiting these controls to levels 1 and 2 of automation.

c. Conditional usage of automation

Although initial testing with human–machine levels of automation was encouraging, a few critical limitations were identified. First, the automatically identified hazard areas were at times too small or too numerous. For example, too many instances of azimuthal shear objects created too many first-guess tornado objects in 2015 (Table 1). Forecasters end up deleting these objects and starting over, while recommending that tornado objects remain manually generated (level 1). This recommendation is supported by Bruick and Karstens (2017), who show that most, approximately 93%–97%, simultaneous peak tornado warning occurrences [i.e., maximum number of tornado warnings in effect within an NWS county warning area (CWA) at a given time] were associated with no more than four or five warnings (generated manually), respectively. Although it can be

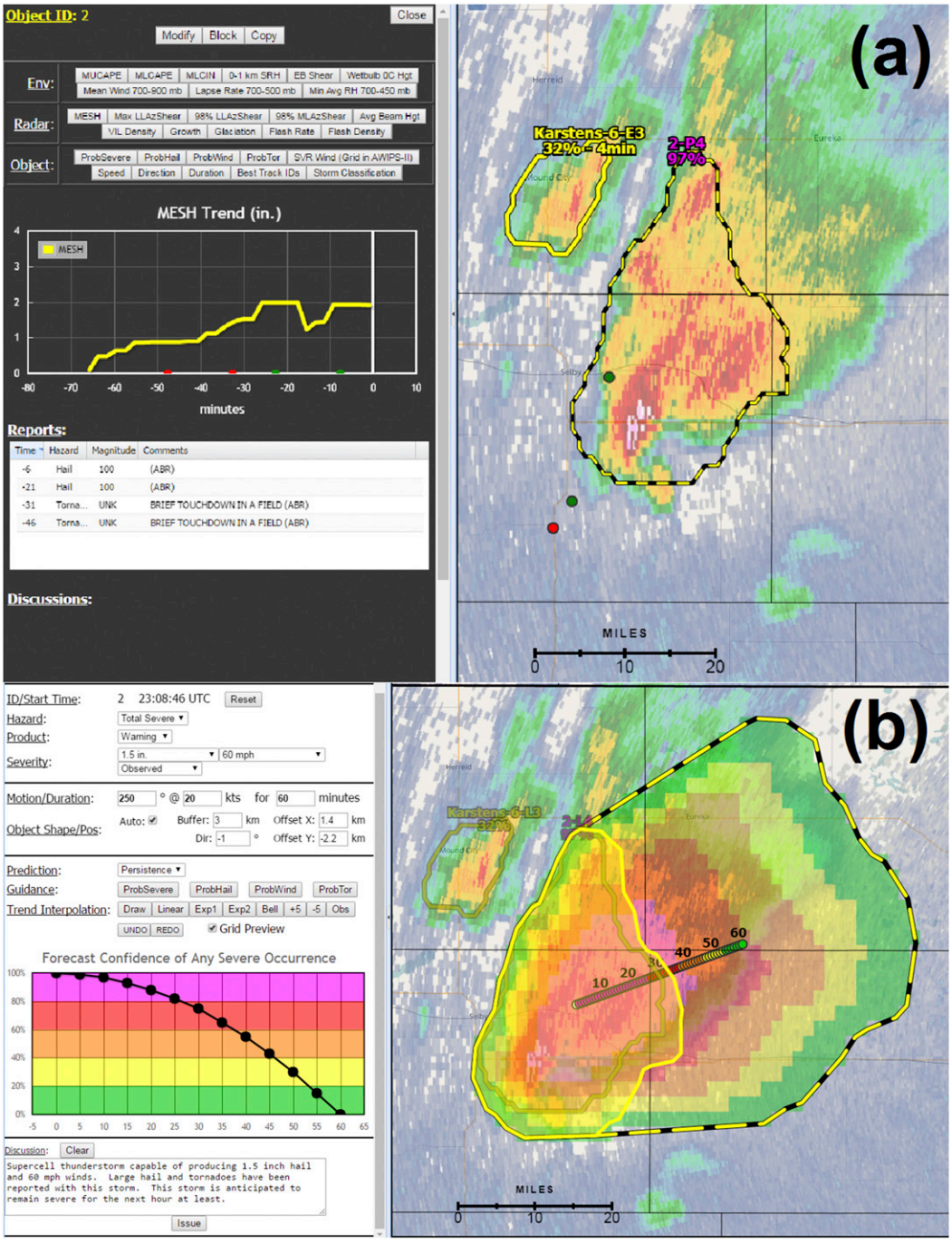

FIG. 2. Visualization of (a) analyzing and (b) editing first-guess ProbSevere objects (yellow/hatched polygons on map) in the prototype PHI tool. Automated objects [hatched object in (a)] are labeled with their identification number, prediction (P = persistence, E = explicit), level of automation (1–4), and probability prediction value. Clicking on an object allows for analysis of historical trends in the predictors, object attributes, and previously issued discussions [see (a)]. Additionally, LSRs are spatially/temporally matched with objects and provided in a table [left side of (a)] and visualized as a breadcrumb trail of color circles on the map. Clicking the Modify button displays controls and tools for editing forecast attributes and object properties [see (b)]. Clicking the Issue button gives forecasters partial or complete control of the object, with enhanced situational awareness attributes added to the object label, including the forecaster name and time (min) since last update [see small object in (a)].

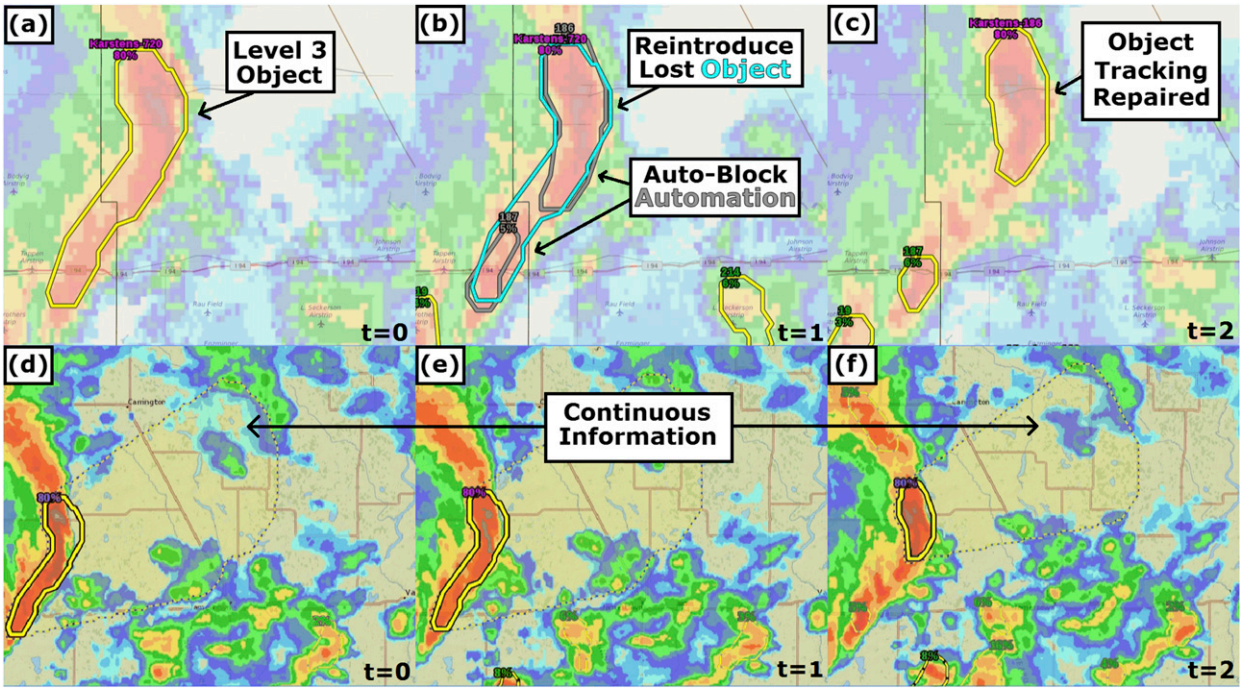

FIG. 3. Illustration of a tactic developed to triage discontinuous automated object tracking, as viewed by (a)–(c) NWS forecaster participants and by (d)–(f) NWS partner participants, at three time steps. In (a)–(c) the underlying color fills are composite reflectivity imagery and low-level radar reflectivity mosaic imagery in (d)–(f).

implied that events requiring tornado warnings can be maintained manually (i.e., four or five objects per forecaster; Karstens et al. 2015), NWS warnings and PHI are not directly comparable, particularly in consideration of the FACETs concept of forecasting additional subsevere hazard areas (Rothfusz et al. 2014).

Second, automated object identification and tracking is not a steady or entirely predictable process. Object tracking methods rely on two-dimensional continuous data for atmospheric processes that are three-dimensional (excluding time), leading to inconsistent object identification numbers while tracking objects. Thus, the identification and tracking of objects is sometimes justifiably (e.g., merging/splitting/growing storms) and unjustifiably (e.g., algorithm limitations) discontinuous. From an automation perspective, these tracking breakage instances can result in a loss of continuous information to NWS partners, requiring subsequent intervention. From an EM perspective, the loss of continuous information is detrimental, as storm history is important for maintaining situational awareness and informing decisions (discussed in section 4).

When presented with these situational impasses, forecasters in 2015 were observed to preferentially assume control of the object (from level 3 to levels 1 or 2) as a way of eliminating the error, resurfacing workload issues like those observed in 2014 (Karstens et al. 2015). To address these limitations, week one of the 2016 testing cycle began with forecaster tools identical to those from 2015 to clarify challenges associated with automated object identification and tracking, particularly when the tracking breaks. This breakage occurs when the original object cannot be identified on the successive data layer and will manifest as one of three potential situations:

1) The original object disappears.
2) The original object is replaced with a new object or set of objects.
3) The original object is merged with another previously identified object or set of objects.

To address these three tracking issues, a tactic was developed to reintroduce any forecaster-modified object that undergoes a tracking failure into the spatial display while automatically masking any overlapping object not being maintained by the forecaster. At this juncture, the forecaster is presented with the opportunity to decide how to proceed, depending on which of the three tracking situations have been incurred. In situation 1, the forecaster can take no action or expire the object. In situations 2 and 3, the forecaster can repair the broken object tracking by transferring attributes from

one object (original) to another [new object(s) that automatically replaced the original]. An example of the situation 2 triaging sequence is provided in Fig. 3.

Adopting this new tactic quickly revealed new results and challenges. In convective events, particularly those with minimal spatial coverage, where tracking issues happen infrequently, the tactic appears to work well. Forecaster actions are constrained in a manner that is aligned with their goals (repair the tracking situation quickly, but with time to make an informed decision if needed) and, therefore, are able to intuitively overcome the three situational impasses quickly and decisively without interrupting the flow of information to NWS partners. However, some convective events appear to trigger these tracking situations frequently and randomly, leading to an additional workload to maintain a coherent geospatial representation of the hazard areas. For example, adjacent cellular storms that grow upscale may undergo a sequence of oscillating mergers and splits before eventually merging. Although adjustments to the object identification and tracking algorithm may alleviate a portion of these issues, it is clear that tracking discontinuities are a feature of convective hazard evolution. A method for extracting postevent storm tracks (Lakshmanan et al. 2015) was adapted for real-time use (HKM) in the 2017 testing cycle, anecdotally revealing a significant reduction in the number of unjustified tracking breakage instances.

With continued testing and evaluation of the human–machine levels of automation, it became quickly apparent through observations of forecasters in 2016 that these levels are not optimal for all convective modes and evolutions. In other words, this process is not unidimensional (Bradshaw et al. 2013). Forecasters needed tools to facilitate their goals, allowing them to transfer between all levels of automation as seamlessly as possible and establish effective forecaster–computer interdependence (Andra et al. 2002; Hoffman et al. 2017). Usage with automated guidance appears to scale predominantly with hazard severity, with subsevere hazards better left to automation (level 4) and significant severe hazards under complete manual control (level 1). Temporal evolution of hazard severity appears to determine the instances when a transfer (i.e., *overriding* or *releasing* control) is needed from one level of automation to another. This transfer process has been categorized into six conditional usages of automation as follows:

1) complete *override* of automation;
2) *override* automated geometry, at least one other attribute automated;
3) automated geometry, *override* at least one other attribute;
4) *release* attributes to automation, maintain geometry;
5) *release* geometry to automation, maintain at least one other attribute; and
6) *release* full control to automation.

Usage with the levels of automation that resulted from this transfer process is described in section 4.

d. Deterministic product generation and usage

Efforts for generating forecast information for user interpretation in the 2015 testing cycle included no deterministic products; only PHI could be used by EMs to make simulated decisions. However, EMs in that testing cycle explained that deterministic warning decisions made by NWS forecasters are codified in standard operating plans, and EMs attempted to deduce determinism from PHI on their own for individual events. Consequently, an obvious need emerged in 2015 to reintroduce deterministic products in subsequent testing cycles. However, Karstens et al. (2015) summarize technological limitations with current storm-based warnings, issued as static, county-clipped polygons with no ability to add area, resulting in a series of several polygons issued for highly dynamic hazards. A reengineering of the system presents an opportunity to address these limitations.

In 2016 forecasters were given the ability to determine which objects to assign a warning (analogous to the current warning system) or a subsevere deterministic product labeled an advisory [analogous to special weather statements (SPS)]. All ProbSevere objects from automation were assigned advisories by default to maintain the expert human judgment expected of severe weather warnings while facilitating product generation across a full spectrum of automated PHI, with forecaster ability to upgrade to a warning or keep as an advisory (Brooks et al. 1992). Additionally, forecasters were tasked with assigning a probability threshold for defining the warning polygon boundary, similar to that proposed by Rothfusz et al. (2014), and a nonnumeric label of forecast confidence at issuance (low, medium low, medium, medium high, high).

Figure 4a shows rather diffuse distributions of assigned thresholds for tornado and severe thunderstorm hazards, with the daily progression of the median on days 2–4 (day 1 was mostly learning) toward lower threshold values (Fig. 4b). This result could suggest the application of thresholds among forecasters and/or among convective situations was inconsistent. Usage of nonzero thresholds for defining deterministic product polygon boundaries necessarily causes a reduction in area, adding geospatial precision already inherent to the object-based method for hazard identification that may

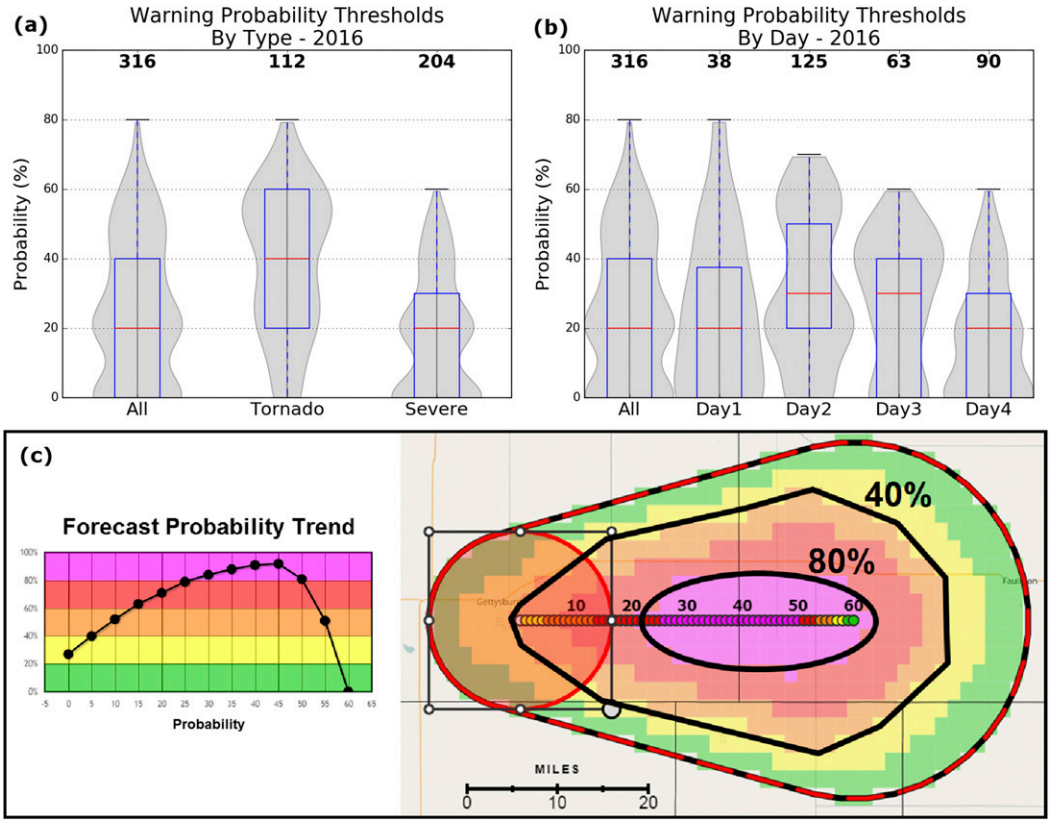

FIG. 4. Distributions [violin plots; http://matplotlib.org/api/pyplot_api.html#matplotlib.pyplot.violinplot) with box-and-whisker diagrams, diamonds (if shown) are outliers beyond the whisker lengths of Q1 − (1.5 × IQR) and Q3 − (1.5 × IQR), where Q1 indicates the first quantile, Q3 indicates the third quantile, and IQR stands for interquartile range] of probability thresholds assigned by forecasters in 2016 for deriving warning polygons from PHI swaths by (a) hazard type and (b) day of testbed testing. Counts are labeled above each distribution. (c) Example of using two probability thresholds (40% and 80%; black polygons) for deriving deterministic warning polygons geospatially using a probabilistic swath and temporally using a forecast probability trend.

be inconsistent with forecaster judgment. This added precision was observed to adversely affect hazard detection, particularly if the product was not updated frequently. This problem is exemplified in Fig. 4c for two probability thresholds. Assignment of a high threshold of 80% results in a polygon area that is sensitive to

1) the forecast probability trend exceeding 80% (pink area on graph),
2) when in the forecast this exceedance occurs (25–50 min), and
3) the method of interpolating the forecast probability trend geospatially (two-dimensional Gaussian).

The result of this sequence of decisions is a small downstream warning polygon (denoted by the black polygon encompassing the 80% area on the map) going into effect several minutes from issuance, an artifact that is inconsistent with traditional warning decision-making (at issuance). Choosing a lower threshold of 40% introduces similar complex technical considerations, but with less area reduction of the resulting polygon not too dissimilar from the lowest nonzero probability threshold (~0%) of the probabilistic swath. In addition, IWT discussions frequently mentioned differing sensitivities, depending on the archetype of severe weather (marginal vs outbreak cases), to the magnitude of probability required to issue a warning. For example, lower probability thresholds to warn might be warranted on outbreak days.

Consideration of these results led to adjustments in the methodology for deriving deterministic products prior to the 2017 testing cycle. The task of assigning a probability threshold by forecasters was removed, and by default all deterministic products were derived using the outer boundary (~0% contour) of the probabilistic swath. This refined definition of determinism reframes the inner probabilistic swath as a relative geospatial representation of forecast confidence (subjective

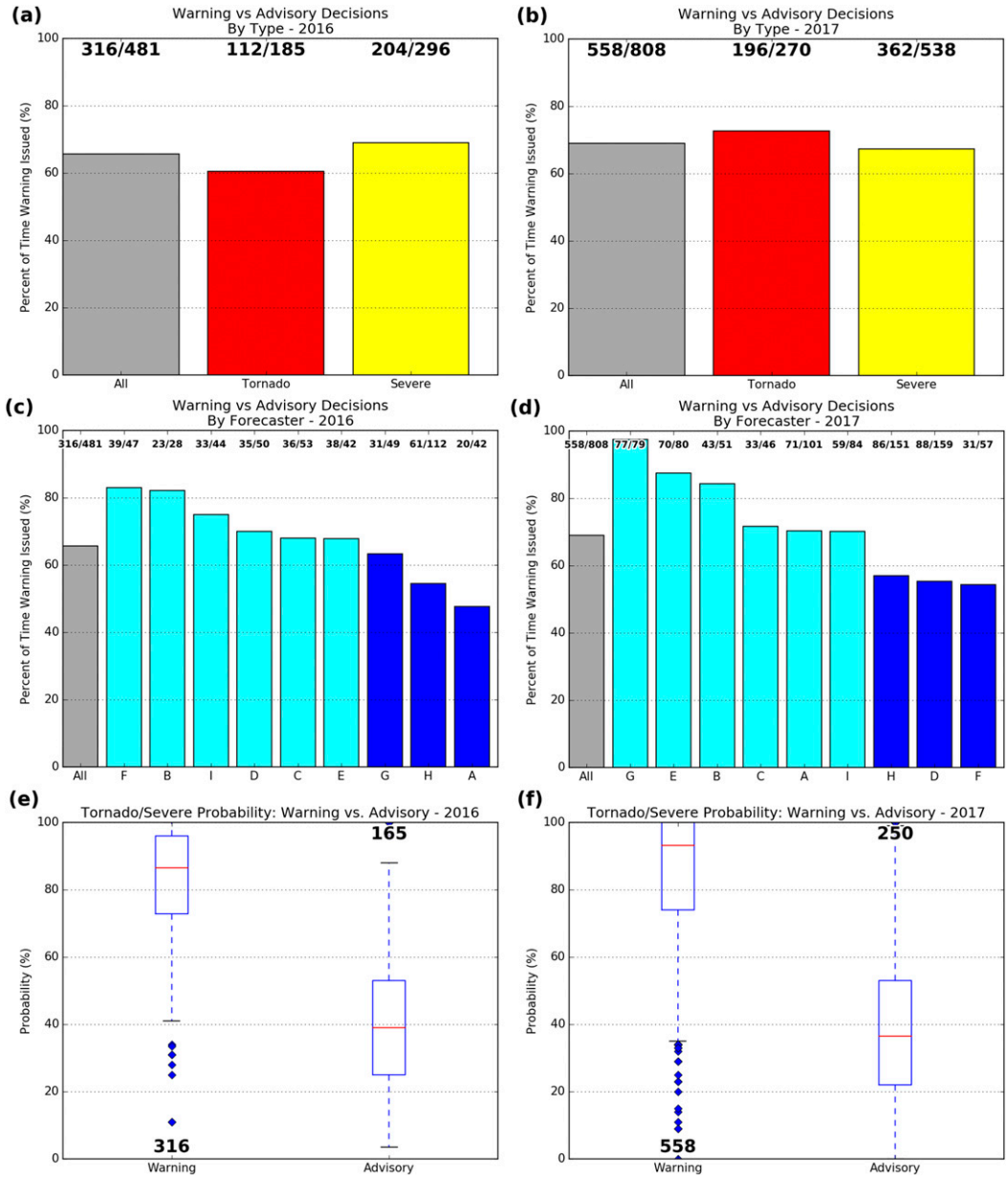

FIG. 5. Percentage of warning issuances vs advisory issuances by (top) hazard type and (middle) forecaster [sorted by percentage and colored by above (cyan) or below (blue) all forecasters average] from (a),(c) 2016 and (b),(d) 2017. Box plots of diagnostic probabilities assigned to warnings and advisories from (e) 2016 and (f) 2017. Numbers above each bin in (a)–(d) give the number of warning issuances compared to the total number of issuances. Counts are provided above/below the box plots in (e) and (f).

probability; Kahneman and Tversky 1972; Brooks et al. 1992; discussed further in section 4) consistent with a user inference of uncertainty within deterministic forecasts (e.g., Morss et al. 2008; Ash et al. 2014; Lindell et al. 2016; Schumann et al. 2017). HKM performed conditional verification with ProbSevere swath polygons [those containing or nearest to a local storm report (LSR)] to test this revised methodology compared to NWS warnings issued during the same period. In this complementary study, detections of LSRs with increasing lead time for both ProbSevere swaths and NWS warnings show similar hit rates as a function of forecast lead time, signaling that similar performance may be obtained if forecasters assign warnings to objects associated with storms ordinarily assigned a storm-based warning.

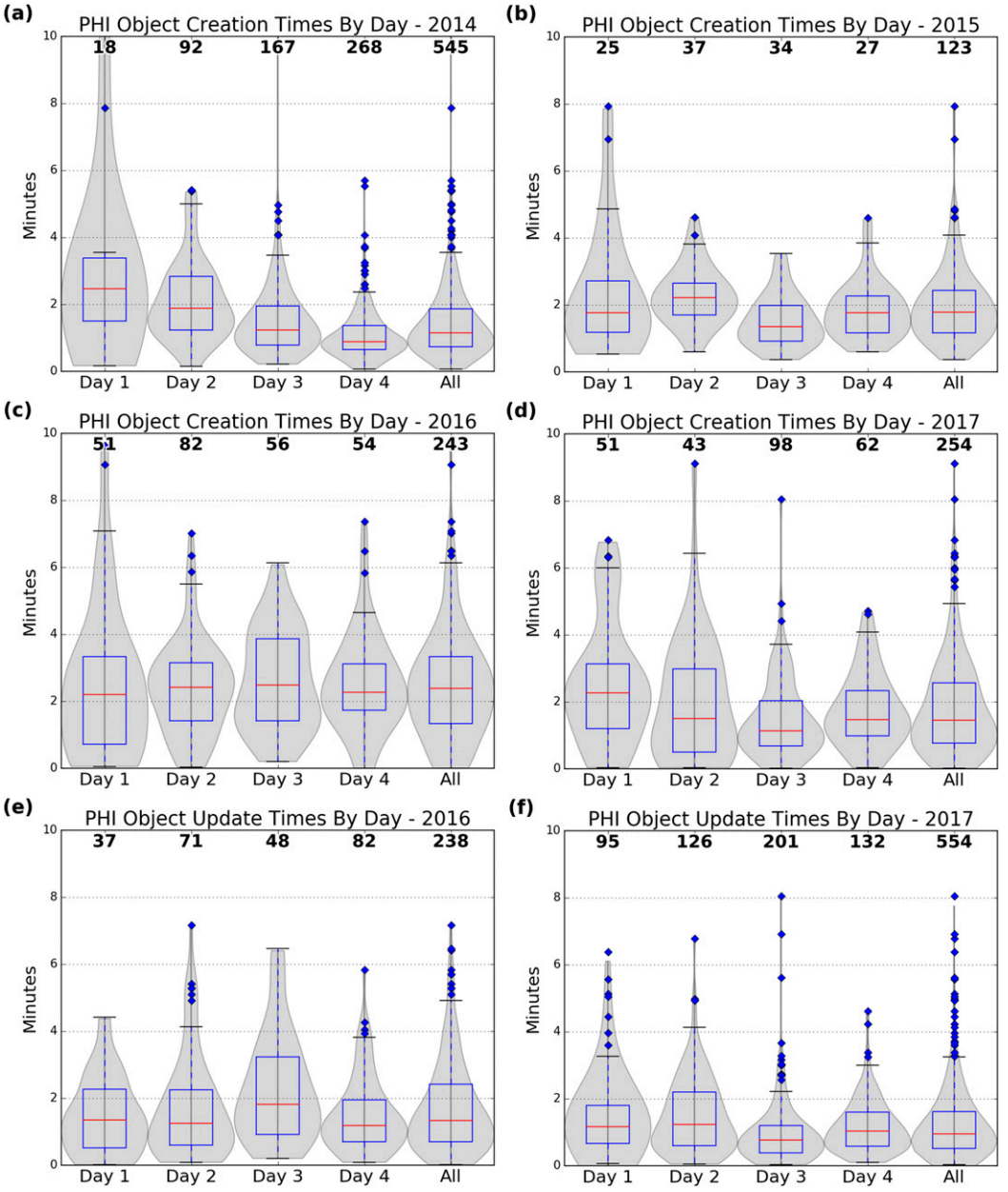

FIG. 6. Time duration distributions for (top),(middle) creating and (bottom) updating PHI objects from the (a) 2014, (b) 2015, (c),(e) 2016, and (d),(f) 2017 testing cycles by day of testing. Counts are labeled above each distribution.

Using these adjustments in 2017 yielded long lead times for deterministic products, which consequently gave EMs who use such products as triggers more time to consider and reconcile warnings containing PHI (as opposed to warnings alone) for augmenting decision-making within the context of expert-forecaster discussion. Among forecasters, the switch to relative probability may have contributed to a reduction in forecast generation times (discussed in section 4). Approximately two out of three deterministic products issued were warnings during both 2016 and 2017 (Figs. 5a,b), indicating a preference for information dissemination driven by forecast confidence in exceeding severe threshold(s), although variation among individual forecasters can be noted (50%–95%; Figs. 5c,d). Likewise, the distributions of diagnostic probabilities associated with warning (relatively high values) and advisory (relatively low values) products

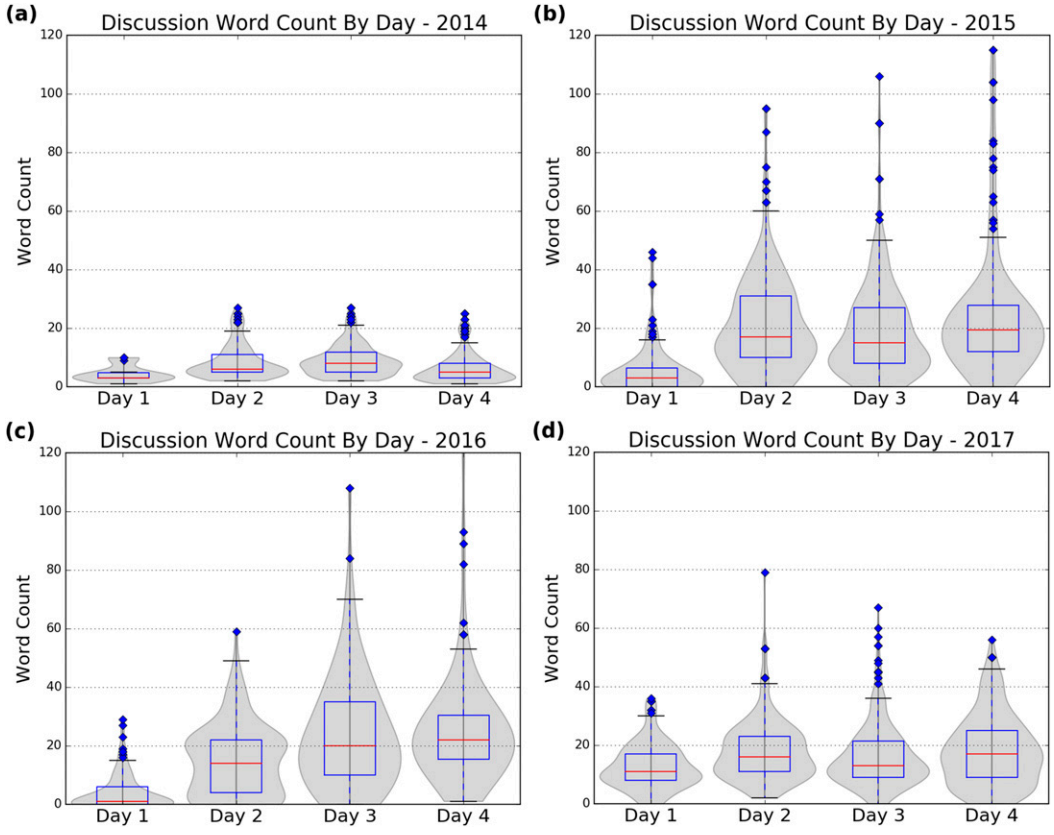

FIG. 7. Forecast discussion word count distributions from the (a) 2014, (b) 2015, (c) 2016, and (d) 2017 testing cycles by day of testing.

supports a severity preference, using probability to convey the likelihood of hazard occurrence (Figs. 5e,f). Analysis of these results in 2016 informed refinement of the probability trend definition (based on hazard occurrence) prior to the 2017 testing cycle, and the congruence in results between 2016 and 2017 supports this change. In the absence of any explicit training, forecaster preference to stratify probabilities by product type reduces, but does not completely avoid, alternative contradictory interpretations (e.g., high-confidence advisory, low-confidence warning); however, the geospatial representation of high-confidence probability trends may unintentionally introduce such interpretations if geospatial relativity is not considered (i.e., low probabilities along the edges of the swath; e.g., Fig. 4c).

4. Quantitative evolution of a human–machine mix

From the 2014 testing cycle it was learned that forecasters could create, issue, and update PHI forecasts within a reasonable amount of time, ranging from 30 s to 2 min (Karstens et al. 2015; Fig. 6a). These forecast creation times decreased in magnitude throughout the week of testing, indicative of forecasters working to rapidly develop intuition with the new tools. However, the creation time distributions from the three testing cycles since 2014 show no daily decrease (Figs. 6b,c,d). The addition of NWS partners in these three testing cycles led to a shift from primarily focusing on learning the tool (Fig. 7a) to adding more information in the forecast discussion throughout the week of testing (Figs. 7b–d). This occurred despite increasing the number of tasks to complete PHI issuance (Fig. 2b). The downward shift of the word count distributions in 2017 (Fig. 7d) compared to 2015 and 2016 (Figs. 7b,c) is likely attributable to a change in the EDD software display. While updating a forecast, the forecasters in 2015 and 2016 were observed to append and time stamp their forecast discussions in the discussion text box (Fig. 2b), creating a running discussion history, based on interactions in the IWTs. To simplify this task, each forecast discussion was automatically made visible and sequentially ordered in the EDD (Fig. 8), compared to earlier visualizations that listed only the most recent forecast discussion. Forecasters also found their own discussions useful when returning to a storm (Fig. 2a), particularly while managing a heavy workload.

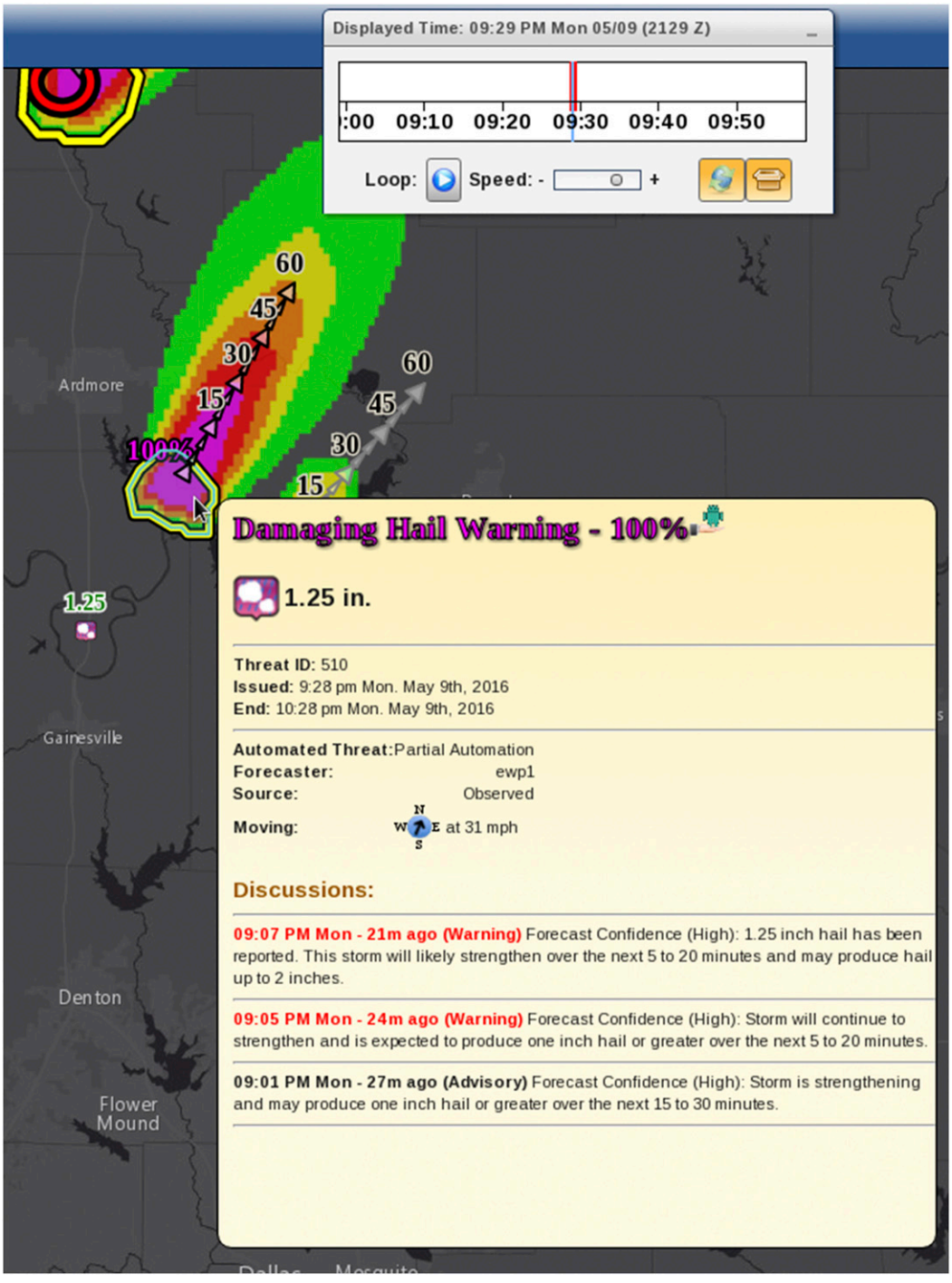

FIG. 8. Screen capture of PHI displayed in the experimental NWS EDD used by EM and broadcast meteorologist participants in 2017. Hovering over a PHI object displays the downstream PHI swath (colors) and a popup box with forecast information, including a sequential listing of forecast discussions associated with the object. The object depicted is a ProbSevere object that underwent a deterministic progression from an advisory to a warning for severe hail, under level 3 of automation.

Daily median update times are smaller in magnitude compared to the creation time distributions (Figs. 6e,f). This reduction in update timing is likely attributable to a reduction in changes needed to the forecast objects and/or attributes (populated from previous issuance) compared to initial creation. This time reduction is also

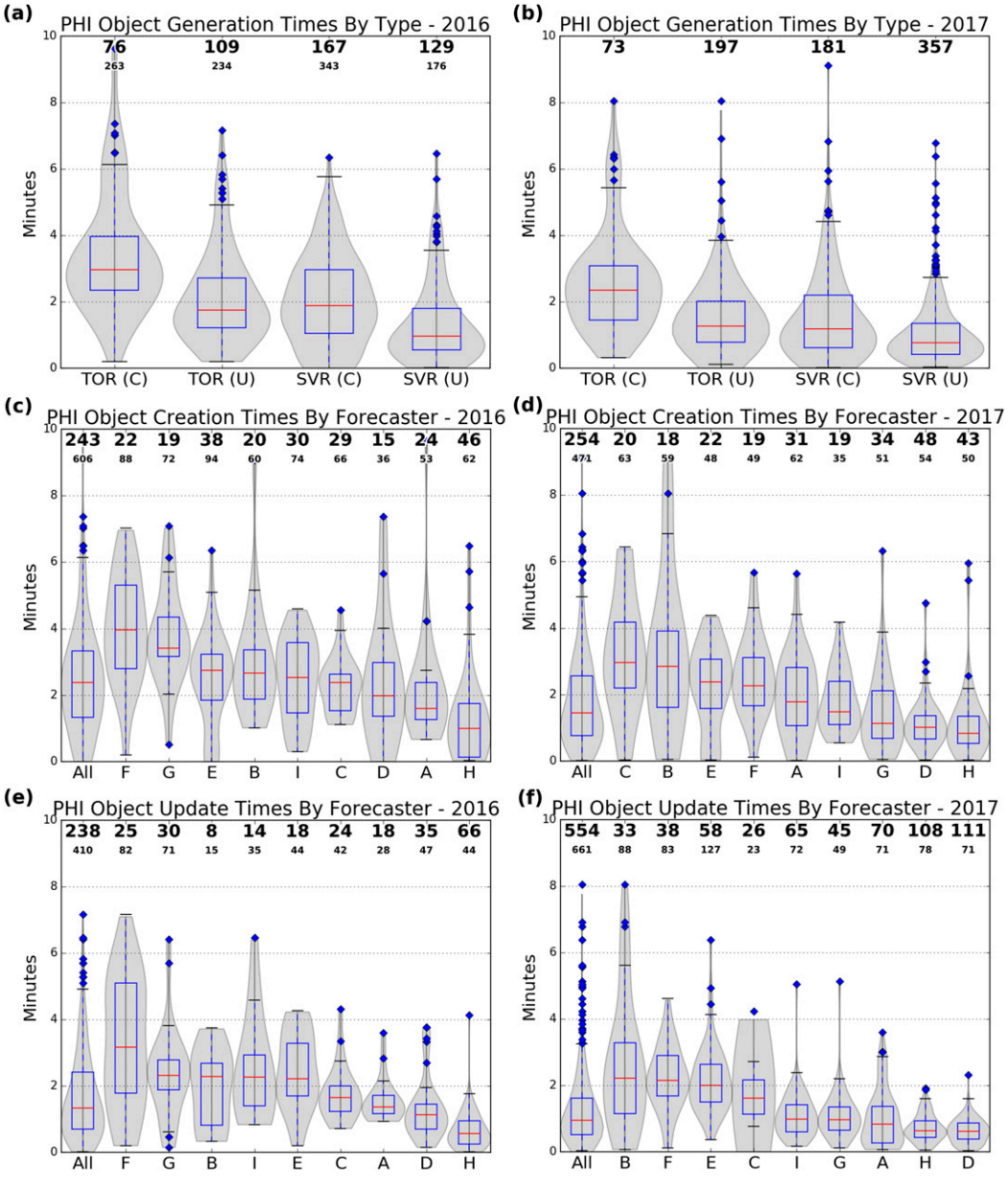

FIG. 9. Time duration distributions for (a)–(d) creating [denoted as C in (a) and (b)] and (a),(b),(e),(f) updating [denoted as U in (a) and (b)] PHI objects from (left) 2016 and (right) 2017 by hazard type in (a) and (b) and by forecaster in (c)–(f). Counts are labeled at the top of each distribution (large font). In (c)–(f), we include labels of total forecast generation minutes (number of forecasts × time to generate each forecast; small font) and individual forecaster distributions are sorted by median values.

apparent when analyzed by hazard type (Figs. 9a,b), with faster generation times for severe hazards than tornadoes. This hazard type disparity is likely attributable to the purely level 1 manual usage associated with tornado objects, compared to the levels 1–3 human–machine options associated with severe objects in 2016 and 2017. Differences in the creation and update timing distributions are also evident when analyzed by forecasters (Figs. 9c–f). Interestingly, there is also an apparent relationship between faster creation/updates resulting in a greater number of forecasts produced. However, temporal integration of the creation/update time distributions (small labels in Figs. 9c–f) suggests that forecasters spent their time in a variety of ways, with some focusing more of their available time on creating forecast information (large values) and others focusing more of their available time on maintaining subjective analysis by analyzing radar observations and guidance (small values).

This variability in strategy among forecasters was observed and sometimes vocalized. A commonly observed strategy among fast generators was to focus primarily on the geospatial aspect of the forecast in the initial creation, issue the forecast, and then immediately update the discussion and other pertinent communication elements. Development and usage of such strategies implies that the software was agile enough to meet forecaster desirements (Deal and Hoffman 2010). The flexibility to decide workflow prioritization allowed a few forecasters to innovate a way to become faster generators of PHI content (2016's forecaster H, 2017's forecasters D and H), thus meeting partner needs more quickly.

Notably, the distributions from 2016 shown in Figs. 6 and 9 are systematically higher in magnitude, by approximately 1 min, compared to other testing cycles. One factor that may contribute to this annual variability is the presentation and usage of first-guess forecast guidance. Figure 10a shows low conditional usage of automation in 2015, with approximately 60% of all forecasts generated manually, and a split between levels 2 and 3 of automation in the remaining 40%. Automation was used more with severe hazards, and quite little with tornado hazards. Figure 10b shows almost an inverse in usage in 2016, with over 80% of all forecasts generated in level 3 of automation and a split among levels 1 and 2 within the remaining 20%. As mentioned previously, both of these years included tools that were insufficient to allow forecasters to transfer among the various levels of automation. With such tools in place for 2017, Fig. 10c shows a slightly more balanced conditional usage of automation, though still favoring level 3 of automation. It is also noteworthy to compare the annual usage among individual forecasters and as a collective. Although some variability in the usage of the three levels of automation is evident among forecasters in a given year, variability among forecasters as a collective is more evident on an annual basis, as observed and as is apparent in Figs. 10a–c. This annual variability is attributable to changes, modifications, and the availability of tools to complete forecast generation tasks, as the fundamentals of the underlying ProbSevere guidance changed little from year to year (Table 1).

Comparing changes in the conditional usage distributions between 2016 and 2017 (Figs. 11a and 11b), the level 1 and 2 distributions appear quite similar, and the level 3 generation times decreased in 2017. From this comparison, it could be surmised that the aforementioned systematic increase in forecast generation times in 2016 was attributable to difficulties in generating forecasts using level 3 of automation. As discussed in section 3b, initially forecasters were observed to make adjustments to automated geometries, and upon

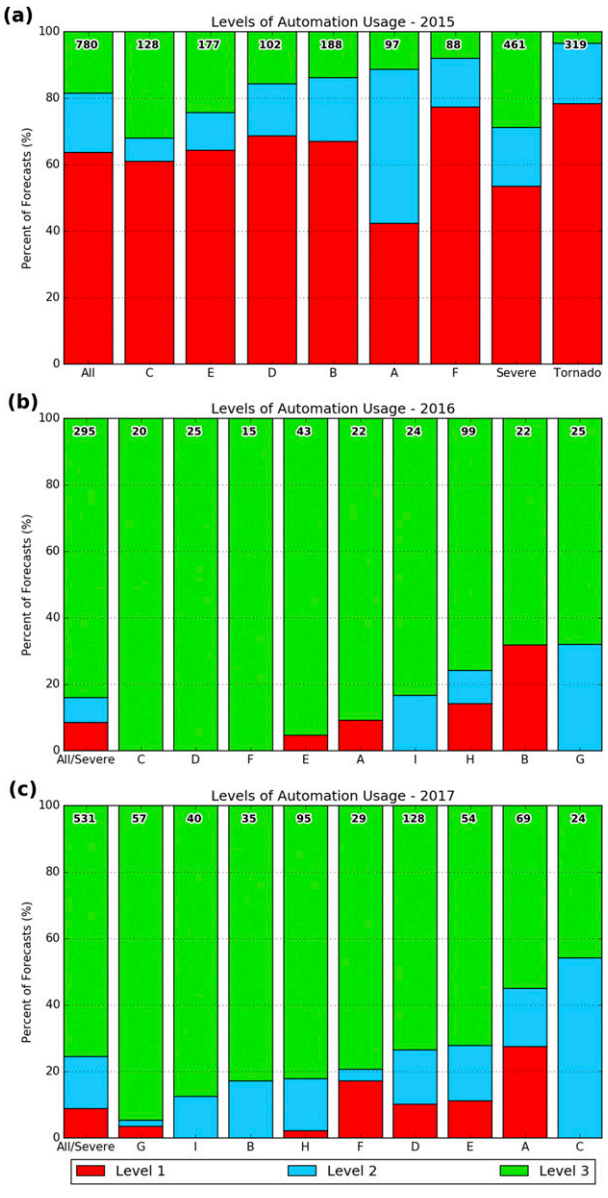

FIG. 10. Normalized conditional usage of levels 1–3 of automation for all automatable forecasts, automatable by forecaster, and automatable by hazard type from (a) 2015, (b) 2016, and (c) 2017. The total number of forecasts is labeled at the top of each bin. Individual forecaster bins are sorted by level 3 of automation usage.

issuance such changes were not saved. This predicament was partially addressed with training in 2016 and completely addressed with the disassociation of object editing controls from level 3 of automation in 2017 (requiring little training). Thus, part of the increase in forecast generation times in 2016 may be explained by this system design limitation, supported by a lack of level 2 conditional usage of automation in 2016 (Fig. 10b). Additionally, the reduction in generation times with level 3

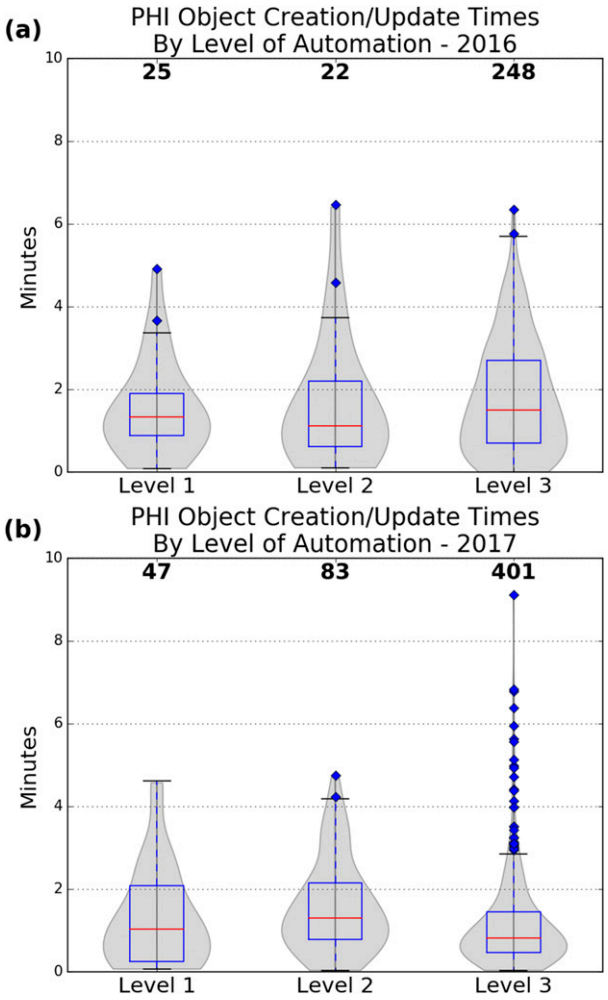

FIG. 11. Time duration distributions for creating and updating PHI objects by level of automation from (a) 2016 and (b) 2017.

of automation in 2017 could signify the development of tools that better facilitate forecaster goals. However, it is important to consider the sample sizes of the distributions (and as shown in Figs. 10b,d), which indicate that most forecasts were generated using level 3 of automation in both years.

Within the human–machine levels of automation, the most frequent overrides to the individual first-guess forecast guidance elements were made to the discussion and probability trend elements (>80%; Figs. 12a,b), indicative of a preference to focus on communication and expressing forecast confidence. Coincidentally, the discussion element was overridden most frequently, followed by the probability prediction element in 2017, which was a reversal from 2016. Figures 12c and 12d show usage of the first-guess probabilistic predictions among all forecasters increased rather significantly between 2016 and 2017, from approximately 5% to 30% usage, which is likely attributable to the addition of explicit probabilistic trend predictions, as opposed to using a default linear decay (Doswell 2004) of the ProbSevere probability prediction in the years prior. The relatively fewer overrides made to the object (10%–20%), speed (40%–50%), and direction (40%–50%) elements may be an implicit indication of the object identification and tracking quality relative to the hazard coverage and anticipated movement. Tools for systematically adjusting the automated object shape (via buffering) and position (via repositioning) were added in 2017 (Fig. 2b) to lower combatable object maintenance issues associated with level 2 usage of automation. Yet, there is a relative increase in overriding the object in 2017 from approximately 10% to 20%, indicating that tools for conditionally using automation were likely preferred over systematic object adjustments (Fig. 10c).

Considering the preceding analysis, the observed increase in forecast generation times in 2016 relative to other testing cycles is hypothesized to be attributable to three factors. These factors include the deficient tools delineating levels 2 and 3 of automation (addressed in 2017 as discussed above), assigning a probability threshold defining a deterministic product boundary (eliminated in 2017, as discussed in section 3c), and assigning a nonnumeric label of forecast confidence at issuance. Forecasters quickly realized, through interaction with EMs, that their labels of forecast confidence (diagnostic) needed to be consistent with their forecast probability trend (Fig. 4c), but had difficulty accomplishing this task due to a temporal incongruence of these tools. In 2017, these two tools were combined, with a "forecast confidence" label (presented in numeric form) attached to the probability trend tool (Fig. 2b), defining it as a subjective probability of hazard occurrence (using traditional thresholds defining severe cases) within the object over an assigned duration. Usage of subjective probability is perhaps quite natural for forecasters, who regularly combine numerous sources of past, present, and future information while immersed in the weather analysis and forecasting process (Bosart 2003; Doswell 2004). Labels of forecast confidence provided to NWS partners were derived from this trend, including both numeric and nonnumeric labels, in addition to conveyance via a forecast discussion (Fig. 8).

5. Importance of forecast elements to EM decision-making

In the end-of-week survey used in the 2017 testing cycle, EMs were asked to rank the importance [on a scale from 1 (not important) to 10 (extremely important)] of forecast elements for informing potential

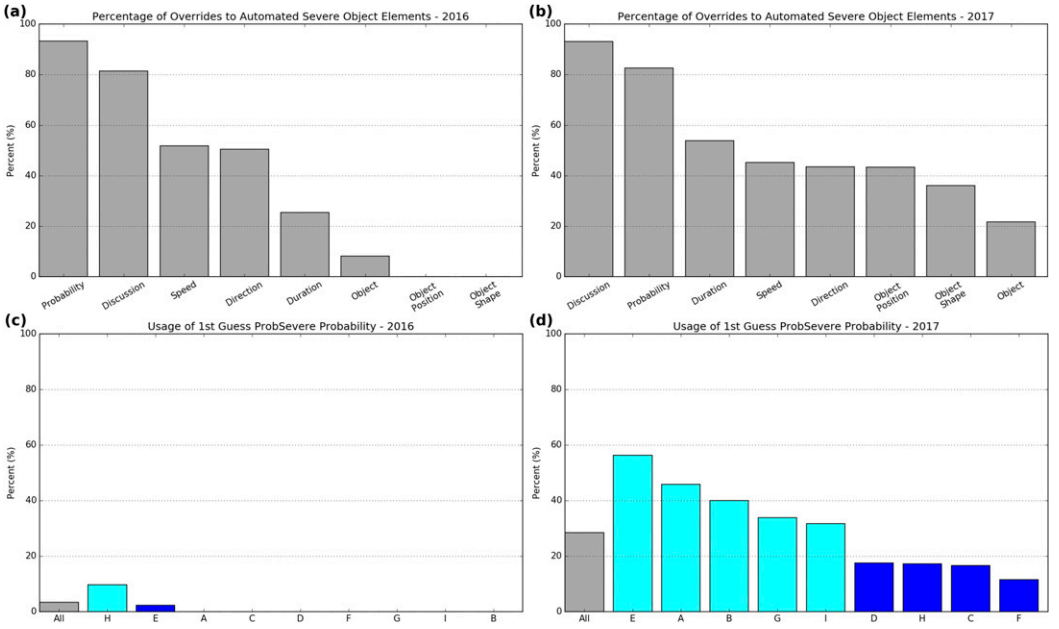

FIG. 12. Percentage of overrides made by forecasters to individual, automated, first-guess forecast elements from ProbSevere objects in (a) 2016 and (b) 2017, sorted in descending order. Also shown is the percent usage of the first-guess diagnostic probability from ProbSevere objects by forecasters in (c) 2016 and (d) 2017, visualized as in Fig. 5.

decision-making. Figures 13a and 13b show that for both severe thunderstorm and tornado hazards, EMs unanimously ranked the discussion element with the highest level of importance (10 out of 10), both informing and substantiating the effort forecasters made to communicate pertinent forecast information (Fig. 12b). This finding signifies a fundamental and irreplaceable role forecasters have in a human–machine mix for severe convective events. Geospatially filtering the sequence of discussions, or log of information (Fig. 8), as opposed to a unidimensional log (e.g., NWSChat), was found to be of significant benefit to facilitating rapid and (often) proactive decision-making, even well before a warning was issued.

The forecast element ranked second highest by EMs was time of arrival, and for some EMs it superseded the warning for initiating actions. Weather-savvy EM participants had well-developed plans of action and an estimation of the amount of time it takes to execute these plans. Time of arrival information, in addition to traditional warning and probabilistic information, meets important needs of this subset of NWS partners. However, providing accurate and reliable timing information *requires* frequent updates to the hazard location, area, and movement. Timing is determined by a space-to-time conversion of the object crossing a user-specified location using the object motion vector. Thus, forecasters have little direct control over timing calculations other than by making frequent updates to the object and motion vector elements. This lends additional support for the human–machine levels of automation. In particular, level 3 of automation rapidly maintains object geometry at a frequency sufficient to resolve hazard evolution. Observations of forecasters working high-impact tornado and severe thunderstorm events revealed that forecasters prioritize situations requiring a near-continuous flow of information, attempting to maintain geospatially precise objects. Forecasters who leverage level 3 of automation appeared to maximize productivity, repurposing time otherwise spent in maintenance activities in a variety of ways such as creating additional forecast information or maintaining subjective analysis by further assessing radar observations and guidance.

Interestingly, probability was ranked with lesser importance for informing decision-making among EMs relative to the discussion and timing elements of the forecast, as well as other elements. In addition to obscurity, the definition of probability can present obstacles to understanding, including uncertainties in the magnitude and number of events being forecast, and the spatial and temporal scales over which the probability is valid (Perry et al. 2016). Joslyn et al. (2009) demonstrate that participants can make better decisions with probabilistic information when the information is framed properly. It appears that framing probability as forecast confidence of hazard occurrence, using traditional severe thresholds, effectively aligned with forecaster and user expectations for representing the concept of probability as applied to severe convective hazards. This is evident by the relative ranking of forecast confidence (third) compared

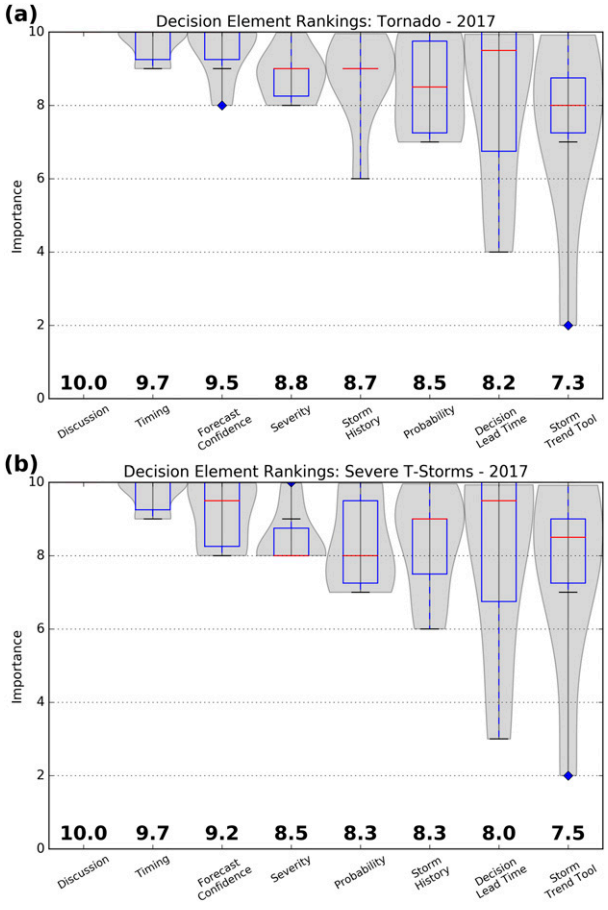

FIG. 13. Distributions of decision element importance ranked independently by six EMs in 2017 for informing simulated decisions for (a) tornado and (b) severe thunderstorm hazards. Averages are labeled below each distribution.

to probability (fifth and sixth), and by the aforementioned hypothesis concluding section 4 to explain why forecast generation times decreased from 2016 to 2017.

6. Summary and discussion

This paper presents the development of a human–machine mix for severe convective events utilizing PHI, building on 2014 HWT testing and evaluation discussed in Karstens et al. (2015), and summarizing subsequent HWT development and testing efforts with NWS forecasters and EMs occurring annually through 2017. The primary themes of the human–machine mix are anchored in the evolutionary usage of automated object-based guidance for augmenting forecaster workflows, generating and communicating a flow of information sufficient to resolve hazard evolution, and expressing forecaster confidence in the likelihood of hazard occurrence. By integrating various algorithms into a web-based prototype that served as a new warning system, forecasters could efficiently take advantage of guidance for the construction of PHI and warnings to meet partner needs.

It was learned that a human–machine mix for severe convective events comprises four levels of automation, with level 1 representing manual forecast generation, level 4 representing automated forecast generation, and the human–machine levels distinguished by forecaster and automated control of object geometry in levels 2 and 3, respectively. Forecasters require an ability to apply these levels selectively to individual convective situations, based on hazard severity, and an ability to easily transfer from one level of automation to another, based on temporal evolution of hazard severity, to forecast and adapt to hazard evolution until hazard demise. However, automated identification and tracking of severe convective processes is justifiably discontinuous, as a result of natural processes such as merging and splitting convection. By implementing a real-time best-track algorithm, a reduction in the number of unjustified tracking breakage instances occurred, but objects still incurring a tracking breakage instance were reintroduced into the spatial display while automatically blocking any newly identified object(s) introduced by automation. This tactic gave forecasters decision-making power and time for repairing tracking breakage instances, consistent with recommendations from NRC (2014). Distributions of the forecast creation and update times by day of testing, hazard type, and forecaster exhibit reasonable timing with stability achieved in 2017.

The presence of EMs provided forecasters an audience for their communication, as well as feedback for what particular information about storms was helpful for decision-making. EMs explained that deterministic warning decisions made by NWS forecasters are codified in standard operating plans. Deterministic product generation from PHI was initially reintroduced in 2016 as a by-product of the PHI swaths according to forecaster-assigned probabilistic thresholds. It was hypothesized that this task contributed to a systematic increase in forecast generation times, along with a limitation in tools for conditionally using automated guidance, and adversely affected hazard detection. Adjustments were made in 2017 such that default deterministic product generation encompassed the entire PHI swath, supported by a three-month conditional verification study (HKM). This change reframed the geospatial representation of PHI as relative subjective probabilities contained within deterministic products, giving EMs longer lead times to consider PHI for augmenting decision-making and with refined geospatial precision compared to current storm-based warnings. However, the process of creating deterministic products from PHI plumes remains unclear. Adjustments to this methodology are still needed, along with an ensuing systematic evaluation of forecaster-generated deterministic products.

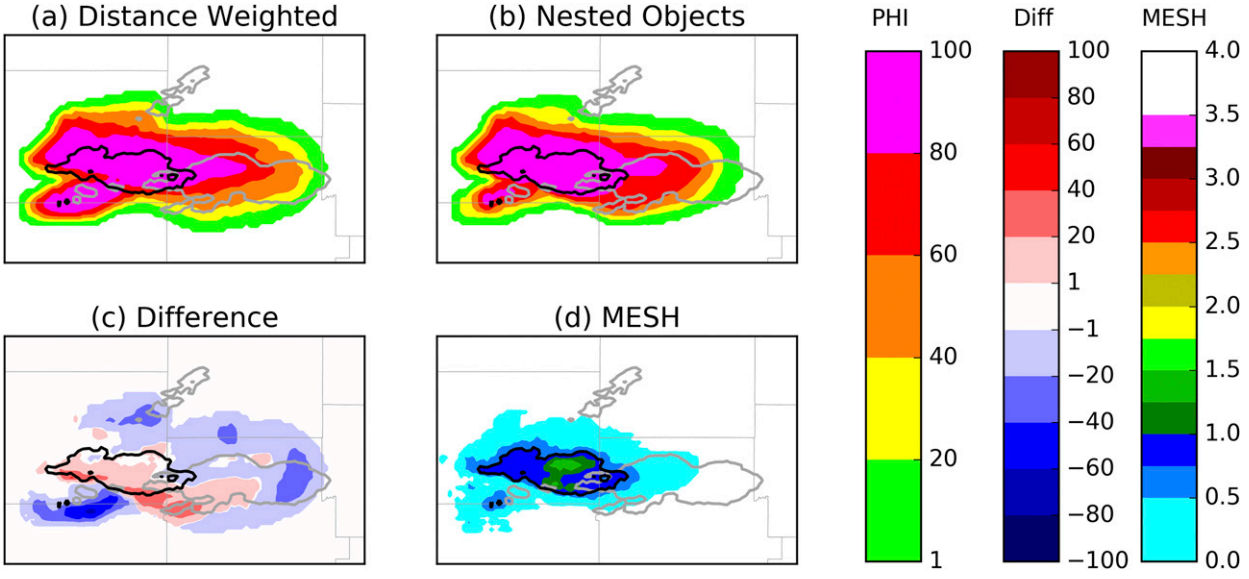

FIG. 14. Accumulated PHI forecasts derived from (a) distance weighting and (b) nested objects methods for two-dimensionally mapping object-based probabilistic forecasts, and (c) difference grid [(b) − (a)], issued 2320 UTC 25 May–0020 UTC 26 May 2016 from automated ProbSevere objects. (d) MRMS MESH observations valid 2320 UTC 25 May–0020 UTC 26 May 2016. All panels include accumulated 0.75-in. MESH contours valid 2320 UTC 25 May–0020 UTC 26 May 2016 (black contours) and 0020–0120 UTC 26 May 2016 (gray contours).

Although efforts to improve the presentation and reliability of probabilistic predictions within the automated object-based guidance appeared to result in greater acceptance and usage by forecasters, it is clear that additional efforts are needed to improve forecaster trust with these predictions. Our imperfect understanding of atmospheric processes leads to imperfect deterministic predictions, making probability a convenient tool to represent forecast uncertainty (Doswell 2004). However, objective probabilities are typically derived from a collection of events (i.e., a training set) that are often ambiguous and/or obscure to forecasters. Rather, guidance is typically introduced through training, and intuition is developed by repeated real-time and postevent assessment of predictions for singular events. Capabilities are needed for actively visualizing the underlying reasoning and training set information inherent to automated techniques to reveal to the forecaster how the technique arrives at its answer (Hoffman et al. 2017). It is hypothesized that such efforts will improve usage and trust of the techniques and better facilitate instances when forecasters can strategically add forecast value. This strategic intervention can be aided by continued efforts to conduct verification while identifying strengths and weaknesses of the guidance through climatological assessments of performance.

In addition, continued development and systematic evaluations of object motion derivation and geospatial representation of forecast confidence are needed. Linear extrapolation was used for motion vector calculations despite underlying tool support for depicting dynamic hazard evolution. Opportunities exist to provide explicit guidance on nonlinear hazard evolution, such as that from a warn-on-forecast (WOF) system (Stensrud et al. 2009). To date, a distance-weighted two-dimensional Gaussian method has been used for mapping forecasters' probabilistic trends of forecast confidence geospatially. It is hypothesized that employing the concept of nested objects, representing objectively identified hazard-specific areas [e.g., MRMS maximum estimated size of hail (MESH)] with a tophat distribution, within objects representing storm-scale reflectivity structures and convective modes (e.g., ProbSevere) with a two-dimensional Gaussian distribution, could yield significant improvements to the relative representation and interpretation of subjective probability, particularly for hazards commonly displaced from the geometric center of storm-scale reflectivity structures. A comparison of these methods to MESH observations is provided in Fig. 14, applying the watershed segmentation algorithm to identify MESH objects at a 0.75-in. threshold. The difference grid in Fig. 14c highlights geospatial refinements from the nested object method, albeit for one case, with an increase in probabilities along the southern flank of the storm and a decrease in probabilities in peripheral areas. Implementation of such refined methods would enable readdressing the reliability of geospatial probabilistic forecasts for hazard-specific severe convective events.

Finally, at this juncture of insights generated with the use of the PHI-based human–machine mix warning system described herein, both forecaster and user decision-making could benefit from ensuing hypothesis testing. We offer the following list of topics to potentially motivate future research endeavors:

- decision-making with probabilistic information while simulating rapid changes in probability forecasts under time pressures,
- conflation of probability with alternative interpretations (e.g., intensity),
- comprehension of relative probability compared to other geospatial representations, and
- importance of forecast elements and their temporal sequencing relative to user decision-making timelines.

Acknowledgments. The authors thank the following people for assistance in the HWT Prototype PHI Tool testing cycles: Gabe Garfield, Darrel Kingfield, Amy McGovern, Kodi Nemunaitis-Berry, Holly Obermeier, Chen Ling, Joseph James, Casandra Shivers, Shadya Sanders, Justin Sieglaff, Mike Pavolonis, James Hocker, Susan Jasko, Gina Eosco, Kim Klockow, Harold Brooks, Robert Hoffman, Israel Jirak, and Greg Stumpf. The authors also thank the many NWS forecasters, emergency managers, and broadcast meteorologists for their diligent efforts and insightful thoughts from working real-time and displaced real-time severe weather events. The manuscript was substantially improved thanks to the constructive comments of four anonymous reviewers. Partial support for this research was provided by NOAA (Grants NA15OAR4590187 and NAISNWS4680019) and the Office of Weather and Air Quality through the U.S. Weather Research Program. Additionally, this paper was prepared by CDK with funding provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

REFERENCES

Andra, D. L., Jr., E. M. Quoetone, and W. F. Bunting, 2002: Warning decision making: The relative roles of conceptual models, technology, strategy, and forecaster expertise on 3 May 1999. *Wea. Forecasting*, **17**, 559–566, https://doi.org/10.1175/1520-0434(2002)017<0559:WDMTRR>2.0.CO;2.

Ash, K. D., R. L. Schumann III, and G. C. Bowser, 2014: Tornado warning trade-offs: Evaluating choices for visually communicating risk. *Wea. Climate Soc.*, **6**, 104–118, https://doi.org/10.1175/WCAS-D-13-00021.1.

Benjamin, S. G., and Coauthors, 2006: From the 13-km RUC to the Rapid Refresh. *12th Conf. on Aviation, Range, and Aerospace Meteorology*, Atlanta, GA, Amer. Meteor. Soc., 9.1, https://ams.confex.com/ams/Annual2006/techprogram/paper_104851.htm.

Bosart, L. F., 1989: Automation: Has its time really come? *Wea. Forecasting*, **4**, 271, https://doi.org/10.1175/1520-0434(1989)004<0271:AHITRC>2.0.CO;2.

——, 2003: Whither the weather analysis and forecasting process? *Wea. Forecasting*, **18**, 520–529, https://doi.org/10.1175/1520-0434(2003)18<520:WTWAAF>2.0.CO;2.

Bradshaw, J. M., R. R. Hoffman, D. D. Woods, and M. Johnson, 2013: The seven deadly myths of ''autonomous systems.'' *IEEE Intell. Syst.*, **28**, 54–61, https://doi.org/10.1109/MIS.2013.70.

Brooks, H. E., C. A. Doswell III, and R. A. Maddox, 1992: On the use of mesoscale and cloud-scale models in operational forecasting. *Wea. Forecasting*, **7**, 120–132, https://doi.org/10.1175/1520-0434(1992)007<0120:OTUOMA>2.0.CO;2.

Bruick, Z. S., and C. D. Karstens, 2017: An investigation of local and national NWS warning outbreaks for severe convective events. *J. Oper. Meteor.*, **5**, 14–25, https://doi.org/10.15191/nwajom.2017.0502.

Calhoun, K. M., and Coauthors, 2018: Cloud-to-ground lightning probabilities and warnings within an integrated warning team. *Special Symp. on Impact-Based Decision Support Services*, Austin, TX, Amer. Meteor. Soc., 4.4, https://ams.confex.com/ams/98Annual/webprogram/Paper329888.html.

Cavanaugh, D., M. Huffman, J. Dunn, and M. Fox, 2016: Connecting the dots: A communications model of the North Texas Integrated Warning Team during the 15 May 2013 tornado outbreak. *Wea. Climate Soc.*, **8**, 233–245, https://doi.org/10.1175/WCAS-D-15-0047.1.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, and D. T. Lindsey, 2014: An empirical model for assessing the severe weather potential of developing convection. *Wea. Forecasting*, **29**, 639–653, https://doi.org/10.1175/WAF-D-13-00113.1.

——, 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.

Correia, J., Jr., D. LaDue, K. H. Knopfmeier, C. D. Karstens, and D. M. Wheatley, 2016: Beyond probability: Providing information to warnings forecasters using the NSSL experimental Warn-on-Forecast System for Ensembles (NEWS-e). *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 5B.2, https://ams.confex.com/ams/28SLS/webprogram/Paper300778.html.

——, ——, C. D. Karstens, K. H. Knopfmeier, and D. M. Wheatley, 2018: Agile postprocessing: Towards user centered ensemble information extraction and visualization. *Eighth Conf. on Transition of Research to Operations*, Austin, TX, Amer. Meteor. Soc., 14B.3, https://ams.confex.com/ams/98Annual/webprogram/Paper332976.html.

Crandall, B., G. Klein, and R. R. Hoffman, 2006: *Working Minds: A Practitioner's Guide to Cognitive Task Analysis*. The MIT Press, 332 pp.

Daniel, A. E., J. N. Chrisman, S. D. Smith, and M. W. Miller, 2014: New WSR-88D operational techniques: Responding to recent weather events. *30th Conf. on Environmental Information Processing Technologies*, Atlanta, GA, Amer. Meteor. Soc., 5.2, https://ams.confex.com/ams/94Annual/webprogram/Paper241216.html.

Deal, S. V., and R. R. Hoffman, 2010: The practitioner's cycles, part 3: Implementation opportunities. *IEEE Intell. Syst.*, **25**, 77–81, https://doi.org/10.1109/MIS.2010.129.

Doswell, C. A., III, 2004: Weather forecasting by humans: Heuristics and decision making. *Wea. Forecasting*, **19**, 1115–1126, https://doi.org/10.1175/WAF-821.1.

Ericsson, K. A., and H. A. Simon, 1993: *Protocol Analysis: Verbal Reports as Data*. Rev. ed. The MIT Press, 496 pp.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

Harrison, D. R., and C. D. Karstens, 2017: A climatology of operational storm-based warnings: A geospatial analysis. *Wea. Forecasting*, **32**, 47–60, https://doi.org/10.1175/WAF-D-15-0146.1.

——, ——, and A. McGovern, 2018: Using machine learning techniques to predict near-term severe weather trends. *13th Symp. on Societal Applications: Policy, Research, and Practice*, Austin, TX, Amer. Meteor. Soc., 11.6, https://ams.confex.com/ams/98Annual/webprogram/Paper326626.html.

Hart, S. G., and L. E. Staveland, 1988: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Adv. Psychol.*, **52**, 139–183, https://doi.org/10.1016/S0166-4115(08)62386-9.

Heinselman, P. L., D. S. LaDue, and H. Lazrus, 2012: Exploring impacts of rapid-scan radar data on NWS warning decisions. *Wea. Forecasting*, **27**, 1031–1044, https://doi.org/10.1175/WAF-D-11-00145.1.

——, ——, D. M. Kingfield, and R. Hoffman, 2015: Tornado warning decisions using phased-array radar data. *Wea. Forecasting*, **30**, 57–78, https://doi.org/10.1175/WAF-D-14-00042.1.

Hoffman, R. R., 2005: Protocols for cognitive task analysis. Advanced Decision Architectures Collaborative Technology Alliance, 108 pp., www.dtic.mil/get-tr-doc/pdf?AD=ADA475456.

——, S. V. Deal, S. Potter, and E. M. Roth, 2010: The practitioner's cycles, part 2: Solving envisioned world problems. *IEEE Intell. Syst.*, **25**, 6–11, https://doi.org/10.1109/MIS.2010.89.

——, M. Johnson, J. M. Bradshaw, and A. Underbrink, 2013: Trust in automation. *IEEE Intell. Syst.*, **28**, 84–88, https://doi.org/10.1109/MIS.2013.24.

——, D. S. LaDue, H. M. Mogil, P. J. Roebber, and J. G. Trafton, 2017: *Minding the Weather: How Expert Forecasters Think*. The MIT Press, 470 pp.

Istok, M. J., and Coauthors, 2009: WSR-88D dual polarization initial operational capabilities. *25th Conf. on Int. Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Phoenix, AZ, Amer. Meteor. Soc., 15.5, https://ams.confex.com/ams/pdfpapers/148927.pdf.

Joslyn, S. L., L. Nadav-Greenberg, M. U. Taing, and R. M. Nichols, 2009: The effects of wording on the understanding and use of uncertainty information in a threshold forecasting decision. *Appl. Cognit. Psychol.*, **23**, 55–72, https://doi.org/10.1002/acp.1449.

Kahneman, D., and A. Tversky, 1972: Subjective probability: A judgment of representativeness. *Cognit. Psychol.*, **3**, 430–454, https://doi.org/10.1016/0010-0285(72)90016-3.

——, and G. Klein, 2009: Conditions for intuitive expertise: A failure to disagree. *Amer. Psychol.*, **64**, 515–526, https://doi.org/10.1037/a0016755.

Kamberelis, G., and G. Dimitriadis, 2005: Focus groups. *SAGE Handbook of Qualitative Research*, 3rd ed. N. K. Denzin and Y. S. Lincoln, Eds., Sage, 887–907.

Karstens, C. D., and Coauthors, 2015: Evaluation of a probabilistic forecasting methodology for severe convective weather in the 2014 Hazardous Weather Testbed. *Wea. Forecasting*, **30**, 1551–1570, https://doi.org/10.1175/WAF-D-14-00163.1.

——, and Coauthors, 2016: Evaluation of near real-time preliminary tornado damage paths. *J. Oper. Meteor.*, **4**, 132–141, https://doi.org/10.15191/nwajom.2016.0410.

Klein, G., 2000: Can information technology reduce expertise? *Proc. Human Performance, Situation Awareness and Automation Conf.,* Savannah, GA, Human Factors and Ergonomics Society, 226.

LaDue, D. S., P. L. Heinselman, and J. F. Newman, 2010: Strengths and limitations of current radar systems for two stakeholder groups in the southern plains. *Bull. Amer. Meteor. Soc.*, **91**, 899–910, https://doi.org/10.1175/2009BAMS2830.1.

——, S. Ernst, C. D. Karstens, J. Correia Jr., J. E. Hocker, and J. P. Wolfe, 2016: Co-creating the form and function of the prototype probabilistic hazard information (PHI) to meet emergency manager user needs. *11th Symp. on Societal Applications: Policy, Research, and Practice*, New Orleans, LA, Amer. Meteor. Soc., 7.4, https://ams.confex.com/ams/96Annual/webprogram/Paper280331.html.

——, and Coauthors, 2017: Temporal and spatial aspects of emergency manager use of prototype Probabilistic Hazard Information. *Fifth Symp. on Building a Weather-Ready Nation: Enhancing Our Nation's Readiness, Responsiveness, and Resilience to High Impact Weather Events*, Seattle, WA, Amer. Meteor. Soc., 896, https://ams.confex.com/ams/97Annual/webprogram/Paper312923.html.

Lagerquist, R. A., A. McGovern, and T. M. Smith, 2017: Machine learning for real-time prediction of damaging straight-line convective wind. *Wea. Forecasting*, **32**, 2175–2193, https://doi.org/10.1175/WAF-D-17-0038.1.

Lakshmanan, V., and T. Smith, 2010: An objective method of evaluating and devising storm-tracking algorithms. *Wea. Forecasting*, **25**, 701–709, https://doi.org/10.1175/2009WAF2222330.1.

——, ——, G. Stumpf, and K. Hondl, 2007: The Warning Decision Support System–Integrated Information. *Wea. Forecasting*, **22**, 596–612, https://doi.org/10.1175/WAF1009.1.

——, K. Hondl, and R. Rabin, 2009: An efficient, general-purpose technique for identifying storm cells in geospatial images. *J. Atmos. Oceanic Technol.*, **26**, 523–537, https://doi.org/10.1175/2008JTECHA1153.1.

——, B. Herzog, and D. Kingfield, 2015: A method for extracting postevent storm tracks. *J. Appl. Meteor. Climatol.*, **54**, 451–462, https://doi.org/10.1175/JAMC-D-14-0132.1.

Lindell, M. K., S.-K. Huang, H.-L. Wei, and C. D. Samuelson, 2016: Perceptions and expected immediate reactions to tornado warning polygons. *Nat. Hazards*, **80**, 683–707, https://doi.org/10.1007/s11069-015-1990-5.

McDonald, J., and J. Correia Jr., 2016: Insights into predicting tornado development using NEWS-e vorticity forecasts. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 15B.3, https://ams.confex.com/ams/28SLS/webprogram/Paper300761.html.

McGovern, A., K. L. Elmore, D. J. Gagne II, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. M. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

——, E. Jergensen, C. D. Karstens, H. Obermeier, and T. Smith, 2018: Real-time and climatological storm classification using machine learning. *17th Conf. on Artificial and Computational*

Intelligence and its Applications to the Environmental Sciences, Austin, TX, Amer. Meteor. Soc., 1.1, https://ams.confex.com/ams/98Annual/webprogram/Paper326198.html.

Menzel, W. P., and J. F. W. Purdom, 1994: Introducing GOES-I: The first of a new generation of Geostationary Operational Environmental Satellites. *Bull. Amer. Meteor. Soc.*, **75**, 757–781, https://doi.org/10.1175/1520-0477(1994)075<0757:IGITFO>2.0.CO;2.

Meyer, T. C., K. M. Kuhlman, D. M. Kingfield, and D. J. Gagne II, 2016: Using random forest technique to create cloud-to-ground lightning probabilities. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 146, https://ams.confex.com/ams/28SLS/webprogram/Paper301841.html.

Militello, L. G., and R. J. B. Hutton, 1998: Applied cognitive task analysis (ACTA): A practitioner's toolkit for understanding cognitive task demands. *Ergonomics*, **41**, 1618–1641, https://doi.org/10.1080/001401398186108.

Mitchell, E. D., 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.

Moller, A. R., C. A. Doswell III, M. P. Foster, and G. R. Woodall, 1994: The operational recognition of supercell thunderstorm environments and storm structures. *Wea. Forecasting*, **9**, 327–347, https://doi.org/10.1175/1520-0434(1994)009<0327:TOROST>2.0.CO;2.

Morss, R. E., J. L. Demuth, and J. K. Lazo, 2008: Communicating uncertainty in weather forecasts: A survey of the U.S. public. *Wea. Forecasting*, **23**, 974–991, https://doi.org/10.1175/2008WAF2007088.1.

National Academy of Science, 2000: *From Research to Operations in Weather Satellites and Numerical Weather Prediction: Crossing the Valley of Death*. National Academies Press, 96 pp.

National Research Council, 2014: *Complex Operational Decision Making in Networked Systems of Humans and Machines: A Multidisciplinary Approach*. National Academies Press, 88 pp.

Nemunaitis-Berry, K. L., H. Obermeier, S. A. Jasko, D. LaDue, C. D. Karstens, G. M. Eosco, A. Gerard, and L. Rothfusz, 2017: Broadcast meteorologist decision-making in the 2016 Hazardous Weather Testbed Probabilistic Hazard Information Project. *Fourth Conf. on Weather Warnings and Communication*, Kansas City, MO, Amer., Meteor. Soc., 1.4, https://ams.confex.com/ams/45BC4WXCOMM/webprogram/Paper318190.html.

NOAA/NWS, 2016: NWSChat. NOAA/National Weather Service, https://nwschat.weather.gov/.

Obermeier, H., K. L. Nemunaitis-Berry, S. A. Jasko, D. LaDue, C. D. Karstens, G. M. Eosco, A. Gerard, and L. Rothfusz, 2017: Broadcast meteorologist decision making in the 2016 Hazardous Weather Testbed Probabilistic Hazard Information Project. *12th Symp. on Societal Applications: Policy, Research, and Practice*, Seattle, WA, Amer. Meteor. Soc., 4.3, https://ams.confex.com/ams/97Annual/webprogram/Paper314239.html.

Perry, S. C., and Coauthors, 2016: Get your science used—Six guidelines to improve your products. USGS Circular 1419, 46 pp., https://pubs.er.usgs.gov/publication/cir1419.

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, https://doi.org/10.1175/1520-0477(1994)075<0203:NWSWPB>2.0.CO;2.

Rothfusz, L., C. D. Karstens, and D. Hilderbrand, 2014: Forecasting a continuum of environmental threats: Exploring next-generation forecasting of high impact weather. *Eos, Trans. Amer. Geophys. Union*, **95**, 325–326, https://doi.org/10.1002/2014EO360001.

Schumann, R. L., K. D. Ash, and G. C. Bowser, 2017: Tornado warning perception and response: Integrating the roles of visual design, demographics, and hazard experience. *Risk Anal.*, **38**, 311–322, https://doi.org/10.1111/risa.12837.

Sheridan, T. B., and W. Verplank, 1978: Human and computer control of undersea teleoperators. Man–Machine Systems Laboratory Tech. Rep., Department of Mechanical Engineering, Massachusetts Institute of Technology, 186 pp., http://www.dtic.mil/dtic/tr/fulltext/u2/a057655.pdf.

Smith, T. M., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630, https://doi.org/10.1175/BAMS-D-14-00173.1.

Snellman, L. W., 1977: Operational forecasting using automated guidance. *Bull. Amer. Meteor. Soc.*, **58**, 1036–1044, https://doi.org/10.1175/1520-0477(1977)058<1036:OFUAG>2.0.CO;2.

Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499, https://doi.org/10.1175/2009BAMS2795.1.

Stuart, N. A., and Coauthors, 2006: The future of humans in an increasingly automated forecast process. *Bull. Amer. Meteor. Soc.*, **87**, 1497–1502, https://doi.org/10.1175/BAMS-87-11-1497.

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.

Thompson, R. L., and Coauthors, 2017: Tornado damage rating probabilities derived from WSR-88D data. *Wea. Forecasting*, **32**, 1509–1528, https://doi.org/10.1175/WAF-D-17-0004.1.

Torres, S. M., and C. D. Curtis, 2007: Initial implementation of super-resolution data on the NEXRAD network. *23rd Conf. on International Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, San Antonio, TX, Amer. Meteor. Soc., 5B.10, https://ams.confex.com/ams/87ANNUAL/techprogram/paper_116240.htm.

UCMP, 2017: A blueprint for scientific investigations. University of California Museum of Paleontology, University of California, Berkeley, https://undsci.berkeley.edu/article/0_0_0/howscienceworks_03.

Wheatley, D. M., K. H. Knopfmeier, T. A. Jones, and G. J. Creager, 2015: Storm-scale data assimilation and ensemble forecasting with the NSSL Experimental Warn-on-Forecast System. Part I: Radar data experiments. *Wea. Forecasting*, **30**, 1795–1817, https://doi.org/10.1175/WAF-D-15-0043.1.

Wilson, K. A., P. L. Heinselman, C. M. Kuster, D. M. Kingfield, and Z. Kang, 2017: Forecaster performance and workload: Does radar update time matter? *Wea. Forecasting*, **32**, 253–274, https://doi.org/10.1175/WAF-D-16-0157.1.

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. Mitchell, and K. W. Thomas, 1998: An enhanced hail detection algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 286–303, https://doi.org/10.1175/1520-0434(1998)013<0286:AEHDAF>2.0.CO;2.

Wolfe, J. P., 2014: An open source approach to communicating weather risks. *10th Free and Open Source (FOSS4G) Conf.*, Portland, OR, Open Source Geospatial Foundation, https://av.tib.eu/media/31617.