Evaluation of a Probabilistic Forecasting Methodology for Severe Convective Weather in the 2014 Hazardous Weather Testbed

CHRISTOPHER D. KARSTENS,*,+ GREG STUMPF,*,# CHEN LING,@ LESHENG HUA,&
DARREL KINGFIELD,*,+ TRAVIS M. SMITH,*,+ JAMES CORREIA JR.,*,** KRISTIN CALHOUN,*,+
KIEL ORTEGA,*,+ CHRIS MELICK,*,** AND LANS P. ROTHFUSZ+

* *Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma*
+ *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
# *NOAA/National Weather Service/Meteorological Development Laboratory, Silver Spring, Maryland*
@ *Department of Mechanical Engineering, University of Akron, Akron, Ohio*
& *School of Industrial and Systems Engineering, University of Oklahoma, Norman, Oklahoma*
** *NOAA/National Weather Service/Storm Prediction Center, Norman, Oklahoma*

(Manuscript received 18 December 2014, in final form 6 July 2015)

ABSTRACT

A proposed new method for hazard identification and prediction was evaluated with forecasters in the National Oceanic and Atmospheric Administration Hazardous Weather Testbed during 2014. This method combines hazard-following objects with forecaster-issued trends of exceedance probabilities to produce probabilistic hazard information, as opposed to the static, deterministic polygon and attendant text product methodology presently employed by the National Weather Service to issue severe thunderstorm and tornado warnings. Three components of the test bed activities are discussed: usage of the new tools, verification of storm-based warnings and probabilistic forecasts from a control–test experiment, and subjective feedback on the proposed paradigm change. Forecasters were able to quickly adapt to the new tools and concepts and ultimately produced probabilistic hazard information in a timely manner. The probabilistic forecasts from two severe hail events tested in a control–test experiment were more skillful than storm-based warnings and were found to have reliability in the low-probability spectrum. False alarm area decreased while the traditional verification metrics degraded with increasing probability thresholds. The latter finding is attributable to a limitation in applying the current verification methodology to probabilistic forecasts. Relaxation of on-the-fence decisions exposed a need to provide information for hazard areas below the decision-point thresholds of current warnings. Automated guidance information was helpful in combating potential workload issues, and forecasters raised a need for improved guidance and training to inform consistent and reliable forecasts.

1. Introduction

For the past approximately 50 years, the National Weather Service (NWS) has issued warnings for severe convective weather events occurring in the United States (e.g., NOAA 2005; Coleman et al. 2011; Brotzge and Donner 2013). These include flash flood warnings (FFW product), severe thunderstorm warnings (SVR product) for hail and/or wind, and tornado warnings (TOR product) for tornadoes and any attendant hail and/or damaging wind. From inception, forecasters issued warnings on a county basis with a text product that is still part of the current warning system. The text product contains information about the hazard location and movement, counties included in the warning, and call-to-action statements. In 2007, the NWS modified its policy and software to allow for storm-based warning (SBW) polygons encompassing parts of one or more counties concurrently (Ferree 2006; NOAA 2007; Sutter and Erickson 2010). Polygon information is transmitted in the text product as a set of latitude–longitude pairs, along with the forecaster-determined hazard centroid and motion vector [time/motion/location (TIME/MOT/LOC)]. Additionally, SBWs can be updated to reduce (but not expand) the polygon area by issuing a severe weather statement (SVS) product.

As real-time detection technology (e.g., Crum and Alberty 1993; Simmons and Sutter 2005; Torres and

Corresponding author address: Christopher D. Karstens, NSSL/WRDD, 120 David L. Boren Blvd., Norman, OK 73072.
E-mail: chris.karstens@noaa.gov

DOI: 10.1175/WAF-D-14-00163.1

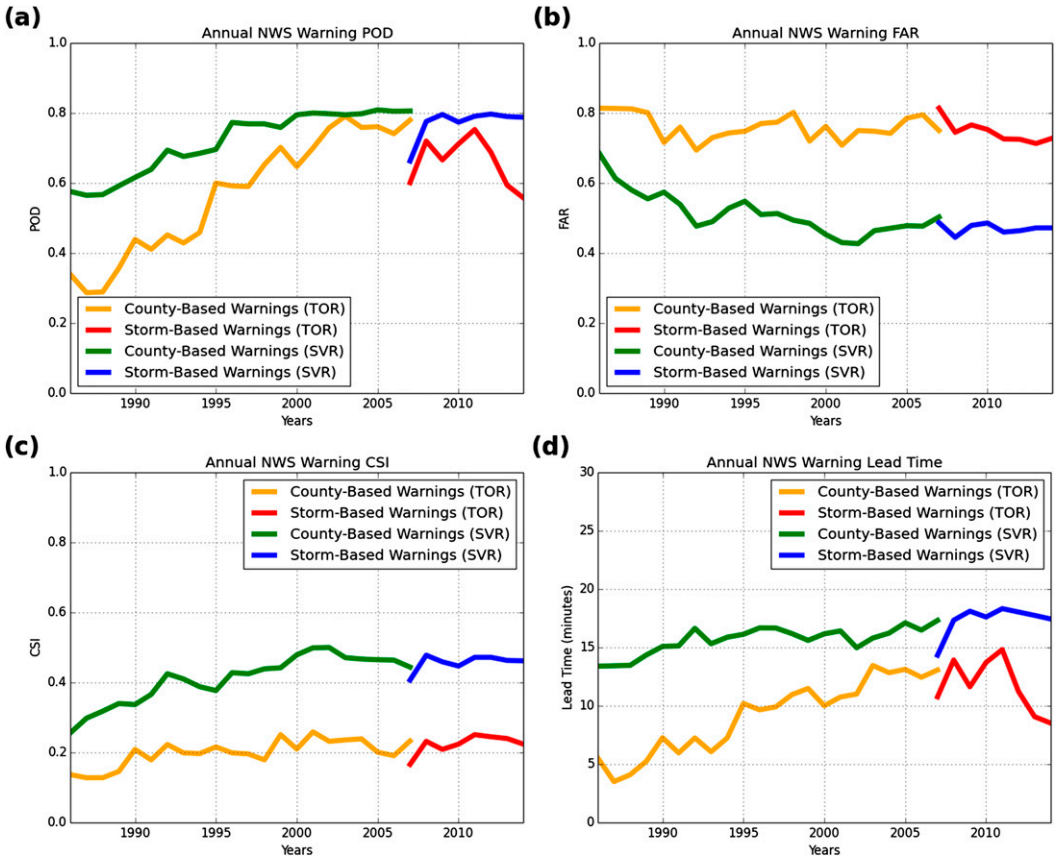

FIG. 1. Trends in annual mean (a) POD, (b) FAR, (c) CSI, and (d) lead time for tornado and severe thunderstorm warnings (1986–2014). The transition from county- to storm-based warnings on 1 Oct 2007 is noted by the switch in line colors (from green to blue and from orange to red). Statistics for 2014 are as of 9 Apr 2015. Data were obtained from the NWS performance management website (NOAA 2015).

Curtis 2007), our conceptual understanding of severe storms (e.g., Lee and Wilhelmson 1997; Markowski et al. 1998) and severe storm environments (e.g., Thompson et al. 2003; Markowski and Richardson 2014), and warning decision-making and training (e.g., Andra et al. 2002) have improved, so have the performance metrics of NWS warnings. Nevertheless, NWS warning performance metrics are dependent upon the collocation in time and space of a local storm report (LSR) within the warning polygon, where the magnitude of the LSR must exceed a threshold defined as severe. Severe thresholds include any tornado, wind exceeding 50 knots (kt; where 1 kt = 0.51 m s^{-1}), and hail exceeding 1 in. (0.75 in. prior to 5 January 2010). Figure 1 presents the annual trend in the traditional verification metrics for warnings, including probability of detection (POD), false alarm ratio (FAR), critical success index (CSI), and lead time, for tornado and severe thunderstorm warnings from 1986 to 2014. During this time period, POD for tornadoes increased from 0.3 to near 0.8, FAR remained nearly the same at around 0.8, CSI increased from 0.15 to near 0.25, and lead time increased from 5 to near 15 min. Similarly, the POD for severe thunderstorms increased from 0.6 to 0.8, FAR decreased from 0.7 to 0.5, CSI increased from 0.25 to 0.45, and lead time increased from 13 to 17 min. A substantial portion of these changes in performance metrics occurred between the mid-1980s and the early 2000s, with little change observed in the last 10–15 yr. The switch to SBWs in 2007 appears to have had a minimal effect on the warning performance metrics, other than perhaps small decreases in POD and lead time. The metrics for 2012, 2013, and 2014 appear slightly worse compared to previous years, which may be attributable to a lack of large outbreaks of significant severe weather (which can boost the overall annual performance).

Despite improvements in warning performance metrics during the past two decades, there is anecdotal evidence that highlights several technological limitations associated with the SBW system. First, warning polygons are static, while the hazard areas are highly dynamic. The SVS product, when issued, cannot add

additional areas to the original warning polygon, only remove areas no longer needing a warning. Thus, forecasters must issue new warnings if hazardous weather moves outside the bounds of the original warning polygon before that warning expires. Second, portions of the SBW polygon are manually and/or automatically removed prior to issuance through a process known as county clipping (i.e., intersection of the warning polygon with geopolitical boundaries) to accommodate county-based alerting systems and reduce overalerting. However, the reduced polygon areas are not always representative of the time-to-space conversion associated with the motion vector, and the amount of automatic county clipping is a setting that varies among weather forecast offices (WFOs) and individual forecasters. The resulting polygons end up being a reflection of the pre-2007 county-based warning system. Third, warnings are verified with a single LSR (i.e., point), yet the hazards occur in swaths (i.e., areas). Uncertainty in receiving at least one LSR for verification can promote the issuance of large and/or overlapping warning polygons, and thus leads to large false alarm areas, and may introduce bias in warning issuance collocated with known population centers. However, given that this discussion is largely anecdotal, it would be advantageous to phrase the aforementioned points as questions to drive research and better understand the limitations of SBWs.

Limitations in the SBW system have also been noted in recent service assessments of the Joplin, Missouri, tornado of 22 May 2011 and the tornado outbreak on 27 April 2011 (NOAA 2011, 2012). A common observation was that people sought additional information beyond initial receipt of the warning, such as visual confirmation, before taking protective action. Inaction to warning messages was attributed to high FAR for tornado warnings, noted in the Joplin report as 76% nationally, and to confusion resulting from inconsistency in the communication of hazardous weather information among the NWS and its partners (e.g., local television and emergency management). Additionally, it was noted that people wanted very specific information about the forecast paths of tornadoes and believed this capability currently exists. With these limitations in mind, it was recommended that the NWS continue to explore evolving the warning system to better support effective decision-making.

The aforementioned recommendation is supported by the National Oceanic and Atmospheric Administration (NOAA) 5-yr research and development plan (NOAA 2014b), which includes an objective for improved decision support tools, with specific strategies that include (i) prototyping warning methodologies that capitalize on future output from storm-scale models, (ii) evaluating experimental products to extend tornado warning lead times to 1 h or greater, and (iii) deploying a unified public warning tool into operations. Realization of these strategies is needed to allow forecasters to take full advantage of the guidance to be generated by projects like Warn-on-Forecast (Stensrud et al. 2009, 2013). Additionally, these strategies are at the core of Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2014), a unifying vision to evolve the current deterministic, product-centric watch and warning paradigm toward a continuum of information via probabilistic hazard information (PHI). To begin addressing the possibility of creating a new watch and warning paradigm, the first of many planned experiments was conducted in the NOAA Hazardous Weather Testbed (HWT; e.g., Clark et al. 2012; Calhoun et al. 2014) during 2014, with several motivating questions (discussed in the next section).

The purpose of this paper is to introduce a prototype system for hazard identification and prediction using PHI, and to present the information that was learned from the 2014 HWT PHI experiment to address issues with the SBW system. Section 2 explains the methodology behind the experiment design and the development of tools that were used by forecasters in a prototype PHI tool. Section 3 presents results from the experiment, including objective measures of forecaster interaction with the prototype PHI tool, verification metrics to ascertain the performance of SBWs versus the probabilistic forecasting methodology, and subjective feedback from the forecasters. The paper closes with discussion and conclusions in section 4.

2. Methodology

a. Experiment design

The objectives of the HWT PHI experiment included the following:

- document forecasters creating, issuing, and updating probabilistic, feature-following objects (i.e., PHI objects) for a variety of real-time and displaced real-time severe weather events with minimal training;
- analyze warnings issued using warning generation software (WarnGEN; operational software for issuing SBWs) and the Advanced Weather Interactive Processing System, version 2 (AWIPS II), versus probabilistic forecasts via PHI objects using the prototype PHI tool for two displaced real-time severe hail events;
- understand forecasters' thoughts on the paradigm change from deterministic watches and warnings to probabilistic forecasts;

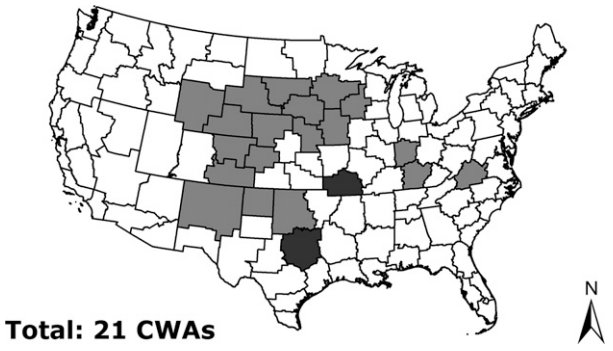

FIG. 2. Map of CWAs in which forecasters operated during the HWT PHI experiment (shaded gray). CWAs used for the two control–test cases (FWD and SGF) are denoted by darker gray shading.

- understand how forecasters use the PHI system, and improve the system design based on forecaster feedback, daily observations, and results from the displaced real-time events; and
- collaborate with the HWT Experimental Forecast Program (EFP) on short-term, regional probabilistic forecasts for individual hazards.

The HWT PHI experiment was conducted during the weeks of 5–9 May, 19–23 May, and 2–6 June 2014. The weeks between operations (12–16 and 26–30 May) allowed for the rapid prototyping of ideas and suggestions from forecasters during the previous week to improve the prototype PHI tool's functionality (e.g., placement of buttons, colors and drop shadows, and probability trend drawing tools). Each week, two forecasters were provided a workstation equipped with the Google Chrome web browser for generating PHI objects using the prototype PHI tool and AWIPS II for display of traditional data products and data interrogation. All sessions were recorded using screen-capturing software, and video and audio recorders. The six forecasters included four males and two females with work experience ranging from a few years to 20+ years and represented NWS offices from four NWS regions (experience from five NWS regions). There were a total of 12 operating days, and forecasters used the prototype PHI tool in a total of 21 NWS county warning areas (CWAs; Fig. 2). Forecasters issued a total of 1213 probabilistic forecasts for a variety of convective modes, including discrete (supercells and isolated cells), mixed (linear systems with isolated cells), and linear (bow echo) modes.

The weekly schedule comprised several activities designed to introduce and test the prototype PHI tool with the forecasters. Day 1 included a short presentation with training material and a hands-on demo of the prototype PHI tool with either a real-time or a displaced real-time severe convective weather event. In addition, a video was distributed to the forecasters to peruse prior to their participation in the experiment. The intent was to give the forecasters enough training material to feel comfortable engaging the prototype software without significantly influencing their conceptual approach to issuing warnings. From this approach, the authors gained an understanding of how forecasters approached and made use of the concepts within the tool, based on aspects of warning decision-making found to be important [e.g., prior conceptual models and experience per Heinselman et al. (2012, 2015)]. Insights gained were combined with results from the control–test experiments to inform future development and experimentation. Days 2 and 3 began with control–test experiments using WarnGEN and the prototype PHI tool, respectively (discussed in section 3b). The remainder of days 2 and 3 was spent evaluating the prototype PHI tool in a relaxed operating period with real-time data for collecting general feedback. On day 4, the forecasters used the prototype PHI tool with a real-time severe weather event in an intensive operating period while testing a variety of geographical sectoring techniques (i.e., split, simultaneous, and multiple CWAs). Forecasters began this day by joining the HWT Spring Forecast Experiment (SFE; e.g., Clark et al. 2012) daily briefing to acquire situational awareness while discussing experimental probabilistic forecasts issued by the SFE for environments supporting severe convection, and immediately followed that with a personal briefing from an SFE representative (also updated later in the day). Day 5 was spent interviewing the forecasters to collect feedback on their experience using the prototype PHI tool and to hear their thoughts on enacting a paradigm shift proposed by FACETs (discussed in section 3c).

b. *Prototype PHI tool strategies*

The prototype PHI tool (Karstens et al. 2014; Fig. 3) is a geospatial web application designed to work toward obtaining forecast goodness through consistency, quality, and value, as discussed in Murphy (1993). The following strategies were developed to accomplish this goal:

1) incorporate tools for generating PHI for individual hazards from severe convective events and for environments that support severe convective events;
2) allow for rapid code development of new or refined PHI generation methods, guidance sources, and interface design; and
3) mimic the layout of the Hazard Services (Hansen et al. 2010) software for AWIPS II.

This study discusses the results from PHI generated explicitly for individual hazards (tornado, wind, and

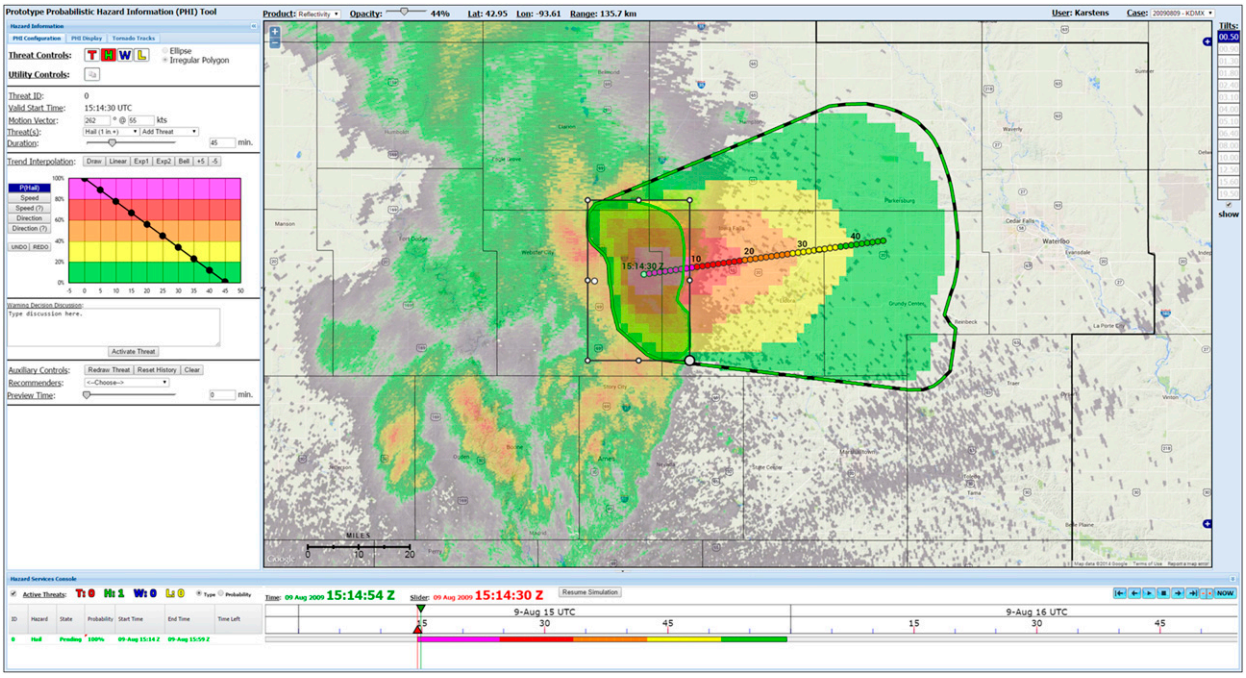

FIG. 3. Screen capture of the web-based prototype PHI tool. The presentation of the tool has been designed to mimic the Hazard Services layout and functionality. The configuration on the left-hand side includes various widgets for adjusting metadata associated with a PHI object, as displayed on the map. The temporal controls for showing time-based information on the map, as well as a table with active or pending PHI objects listed as records are given below. Functionality along the edges above and to the right controls the display of radar information on the map.

hail) from severe convective events. To accomplish the aforementioned task, a methodology was developed for probabilistically denoting the current and predicted spatial extent of individual hazards. This method is based on fundamental knowledge of uncertainty forecasting (Doswell 2004), findings from previous development for denoting and tracking radar-based threat areas (Lakshmanan et al. 2005; Ortega 2008), and evaluation in the HWT (Kuhlman et al. 2008; Stumpf et al. 2008). Furthermore, the current method is empowered by the conceptual versatility of objects, both geographically (polygons drawn by the forecaster or provided as guidance) and programmatically (through object serialization and self-describing attributes). In addition to current and predicted hazard demarcation, an object's spatial position is time derived, allowing it to track along with the hazards, according to an assigned time-based motion vector, and expand, according to an assigned time-based motion uncertainty, throughout the lifetime of a given hazard. This concept is referred to as Threats-in-Motion (TIM; Ortega 2008; Stumpf 2012) and allows for dynamic hazard warning and cancellation.

A critical aspect of this methodology is constructing tools that free forecasters of accumulating evidence to make a warning decision, and give them the ability to quickly convey the probability of the hazard occurrence based on the evidence currently available. Forecasters interactively drew a time-based probability of threshold exceedance trend for each object. This trend is intended to be a manifestation of a forecaster's uncertainty (diagnostic and prognostic) extending through the predictability limit of the hazard (i.e., duration). The probabilities from this trend are mapped to the earth's surface using a combination of a time-to-space conversion and a two-dimensional Gaussian distribution. For this study, the NWS severe thresholds (≥1-in. hail, ≥50-kt wind, and any tornado) were used to generate probabilities of exceedance. After issuance, PHI can be mapped to fine-resolution grids of any spatial and temporal specified dimension (e.g., 1 km^2 and 1 min) and produce location-specific information (time range of arrival, departure, and duration) using a time-to-space conversion. A conceptual mock-up of this output is provided in Fig. 4 for the 20 May 2013 tornado event that tracked through the cities of Newcastle, Oklahoma City, and Moore, Oklahoma (Atkins et al. 2014; Burgess et al. 2014).

A limited set of first-guess numerical weather guidance products [known as a recommender in the Hazard Services system; Hansen et al. (2010)] was provided to the forecasters within the prototype PHI tool. This recommender included an implementation of an object identification algorithm (Lakshmanan et al. 2009) to

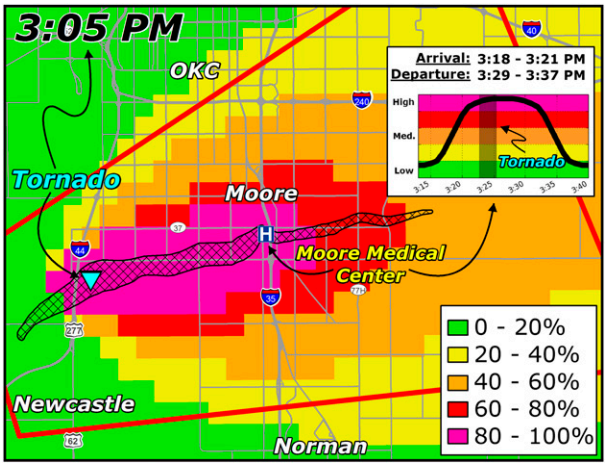

FIG. 4. Conceptual mock-up of end-user PHI (color-filled grid cells of tornado probability with 1-km^2 grid spacing) compared to the current SBW methodology (red polygon). The inset on the upper right-hand side shows a conceptual probability time series associated with this particular PHI-object forecast, as well as the forecast time of arrival and departure (using a 35% probability threshold) for the Moore Medical Center compared to the estimated observed Newcastle–Oklahoma City (OKC)–Moore tornado duration [using terminal Doppler weather radar (TDWR) and damage width; Burgess et al. (2014)].

identify potential hail-producing storms. This algorithm uses an advanced watershed technique to segment local maxima in the −10°C reflectivity isosurface and identify objects on four spatial scales [20, 200, 800, and 2000 km^2; Humphrey et al. (2014)]. For simplicity, the *k*-means elliptical fits to the identified radar objects were provided to the forecasters, as opposed to an irregular radar-derived isosurface (e.g., Cintineo et al. 2014). The maximum value of maximum estimated size of hail (MESH) within each cluster was used to derive a recommended probability of severe hail (POSH). The POSH value was derived by calculating the probability of observed hail exceeding 1 in., using reports obtained from the Severe Hazards Analysis and Verification Experiment (SHAVE; Ortega et al. 2009), given a value of MESH (maximum value in 5-km search radius) from 101 storms (1339 volumes). The recommender guidance was provided to the forecaster in real time and for the two displaced real-time cases in the control–test experiment. It is important to note that the guidance probability automatically populated the PHI-object trend with an assumed decay to zero probability over 45 min; however, this information could be modified by the forecaster (discussed further in section 3a).

c. *Prototype PHI tool tactics*

For this experiment, the creation of PHI was accomplished by the forecaster in two ways: manual creation or incorporation of recommender guidance. This section illustrates how each of these methods is performed, beginning with the manual creation method:

1a—Forecaster engages a hazard-specific object creation tool (Fig. 5a) and draws a polygon (ellipse or irregular shape) encompassing the radar-indicated hazard area (Fig. 5b).
2a—Forecaster computes the swath area (area swept out by the moving object through its duration) by toggling to previous radar scans of the hazard and adjusting the position of the object to match the hazard location at these previous times (Figs. 5c–e). The swath area is internally computed using a cascaded union of all time- and space-interpolated object positions through the duration period.
3a—As step 2a is performed, a mean motion vector is computed from the object's location history as well as vector error statistics. These vector error statistics are used to compute the object's motion uncertainty. The forecaster can optionally override the mean vector and vector error statistics by applying experimental object motion recommenders (e.g., 75R30 rule used by operational forecasters to describe observed motion patterns of maturing supercell thunderstorms turning to the right by 30° and slowing down to 75% of the original speed; e.g., Fig. 5f).
4a—Forecaster interactively draws a probability of exceedance trend and assigns a duration to the object. The probability trend begins at the current radar scan time and ends at the current radar scan time plus the assigned duration (Fig. 5f). The geospatial grid may be previewed at this step.
5a—Forecaster optionally adds metadata to the object. Currently, this is accomplished by typing a brief discussion describing the meteorological significance of the hazard area (Fig. 5f).
6a—Forecaster issues their probabilistic forecast. A background process is triggered to compute hazard information in a variety of geospatial formats [e.g., shapefile, Keyhole Markup Language (KML), Network Common Data Form (netCDF), Gridded Binary second edition (GRIB2), and Geographic JavaScript Object Notation (GeoJSON; http://geojson.org/geojson-spec.html)] for customized display (e.g., Fig. 4).

Use of recommender guidance:

1b—Forecaster selects a recommender object (Fig. 6a). A pop-up dialog displays the object's attribute trends (e.g., MESH and POSH) for analysis (Fig. 6b).
2b—Forecaster engages the recommender (Fig. 6c). The object is displayed along with a swath derived from the object's motion vector and motion

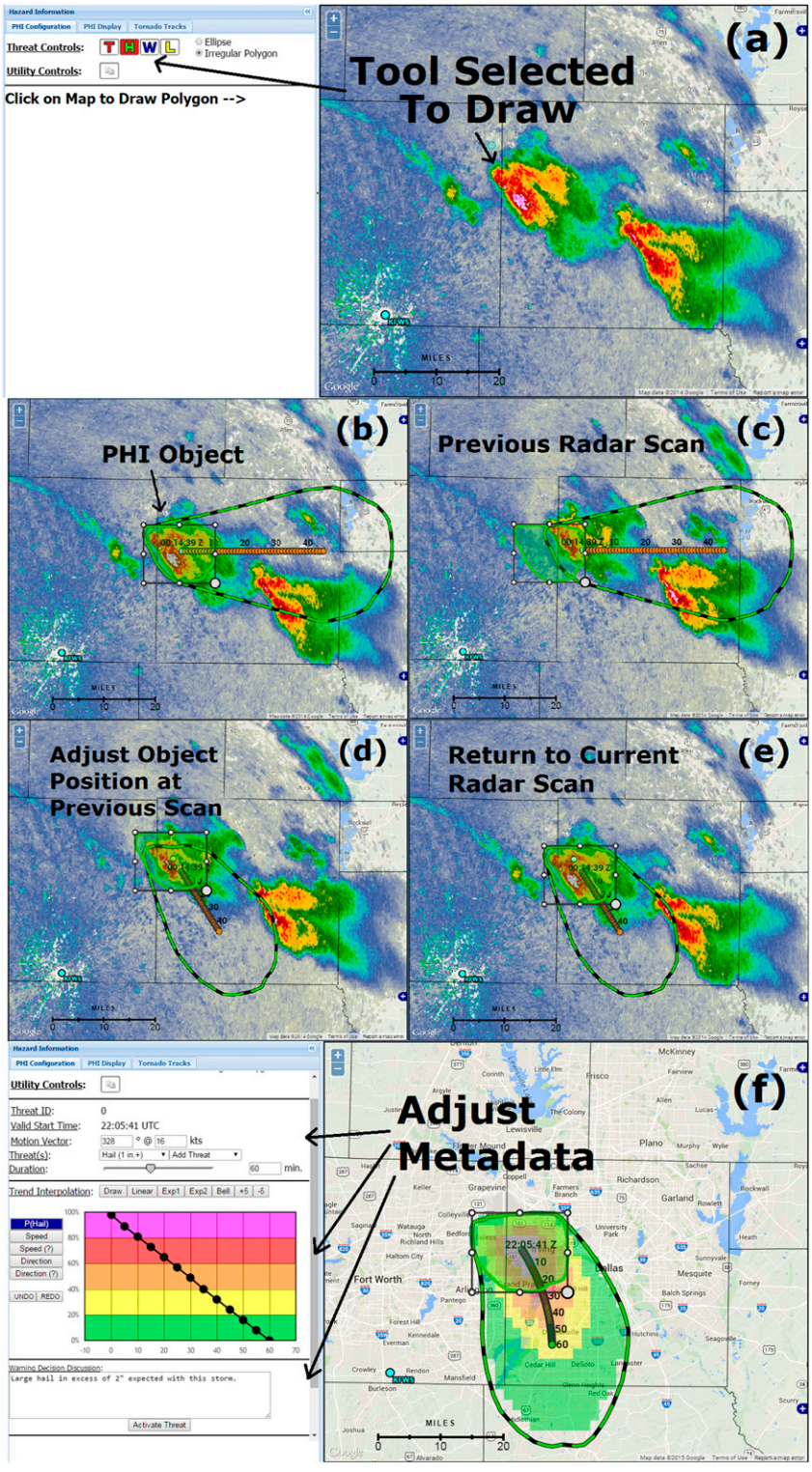

FIG. 5. Illustration of how forecasters manually created a PHI object. (a) The forecaster selects a hazard-specific drawing tool (left) and encompasses the radar-indicated hazard area with a polygon, (b) default attributes are assigned to the object to create a swath, (c) the forecaster steps back previous radar scans while the object moves accordingly, (d) the forecaster adjusts the position of the object to geospatially match the location of the hazard area, (e) the forecaster returns to the current radar scan, and (f) the forecaster adjusts the probability of exceedance trend and adds metadata to the object prior to issuance.

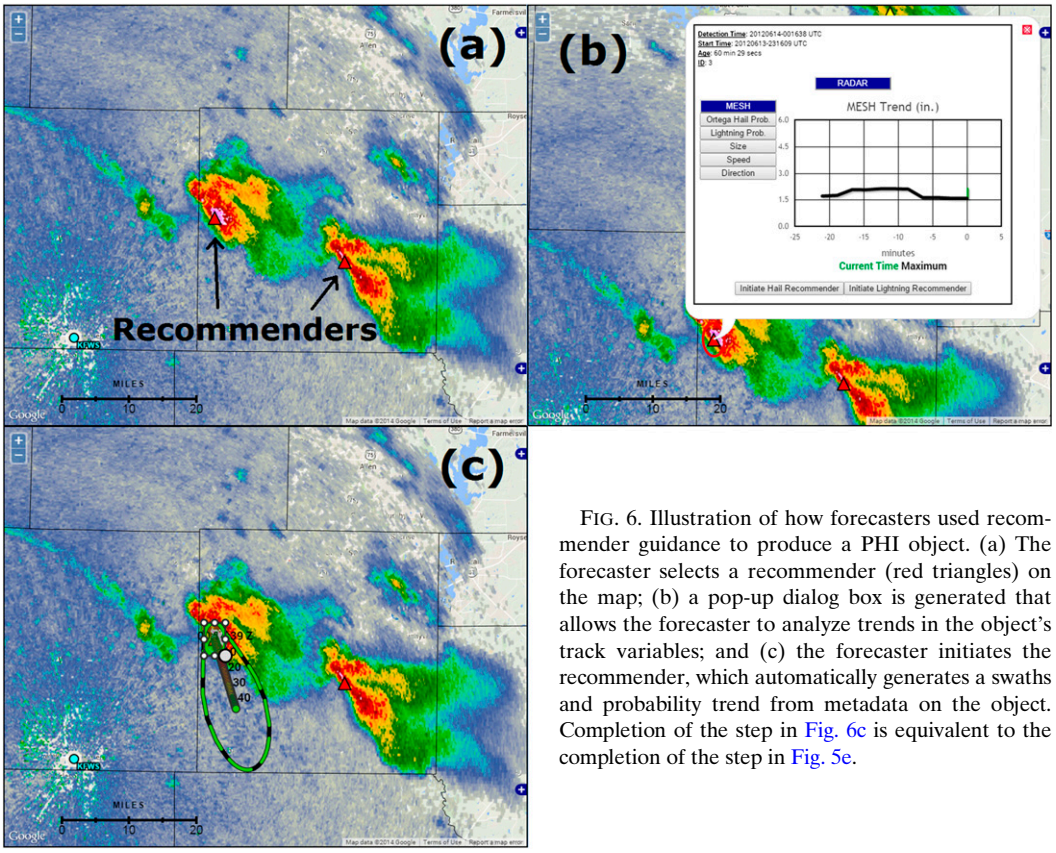

FIG. 6. Illustration of how forecasters used recommender guidance to produce a PHI object. (a) The forecaster selects a recommender (red triangles) on the map; (b) a pop-up dialog box is generated that allows the forecaster to analyze trends in the object's track variables; and (c) the forecaster initiates the recommender, which automatically generates a swaths and probability trend from metadata on the object. Completion of the step in Fig. 6c is equivalent to the completion of the step in Fig. 5e.

uncertainty (provided via the *k*-means algorithm). The probability trend is automatically populated (see discussion in section 2b).

3b—Forecaster optionally overrides the automatically generated swath and/or probability trend by performing steps 2a–4a from the manual creation method.

4b—Forecaster performs steps 5a and 6a from the manual creation method.

After the probabilistic forecast has been issued, the object moves and expands in time according to the time-based motion vector and motion uncertainty assigned to it. Thus, as new radar scans arrive, the object moves along with the projected motion of the hazard area. Over time the geospatial extent of the hazard and its intensity will evolve and achieve inconsistency with the moving object and the exceedance probabilities, thus requiring an update. An update may be accomplished as follows:

1c—Forecaster selects a previously issued object and chooses to update the object (Figs. 7a,b).

2c—A clone of the original polygon is interpolated to a location consistent with the object's motion vector according to the current radar scan time. The attribute trends, including speed, direction, speed uncertainty, direction uncertainty, and probability, are truncated to remove values no longer valid (Fig. 7c).

3c—If the forecaster used a recommender object previously, the object attribute trends from step 2c are automatically updated.

4c—Forecaster optionally overrides automated updates or prior attribute information by performing steps 2a–4a from the manual creation method (Fig. 7d).

5c—Forecaster performs steps 5a and 6a from the manual creation method.

It is important to note that these tactics are a first-guess implementation toward achieving the strategies outlined in the previous section and serve as the basis for understanding methods that worked well versus those that need improvement through testing and iterative development, as is discussed in section 3.

3. Results and discussion

a. Prototype PHI tool usage

Forecast duration is part of both the SBW system and the prototype system; however, an additional

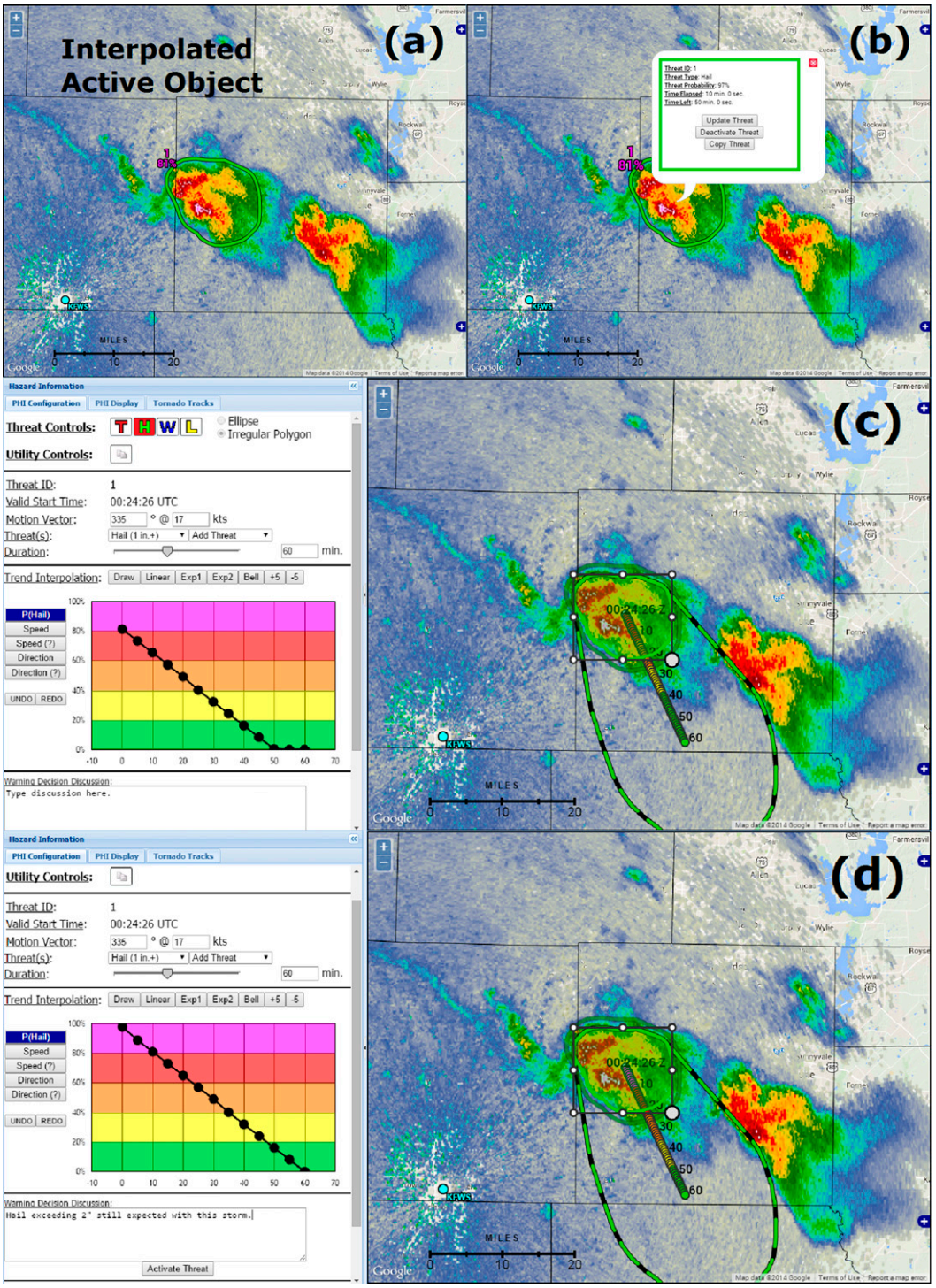

FIG. 7. Illustration of how a forecaster updated a PHI object. (a) The forecaster interpolated the position of the object created in Fig. 5 after two additional radar scans (approximately 9 min later) via TIM; (b) the forecaster selected the object, generating a pop-up dialog box with options to update, copy, or deactivate the object; (c) the forecaster chose to update the object; (d) which allowed the object and metadata to be updated. The original time-based motion and probability attributes are truncated (removal of the first approximately 9 min) in (c).

component in the prototype system is the ability to forecast a probabilistic trend representing the diagnostic and prognostic uncertainty through a duration extending through the predictability limit of the hazard. This added component of hazard prediction raises the question, will forecasters embrace the ability to extend the duration of forecasts past times typical of NWS SBWs with low probabilities? A compilation of all probabilistic forecast trends issued by the forecasters during the experiment (1213 total) is provided in Fig. 8a. First, a distinct cutoff in several of the probability trend lines (75% of the lines) occurs at the 45-min mark. Within the prototype PHI tool, the default duration assigned to all PHI objects was 45 min (as is a common default in the NWS WarnGEN application), and thus it is perhaps not coincidental that so many of the probability trends abruptly end at the 45-min mark. Another interesting aspect of many of these trend lines is the relatively high values that remain near the end of the probability trend line. This effect raises a question about the relationship between the shape and decay rate of these curves to hazard prediction. The effect may imply that, based on the evidence at hand, forecasters indeed felt confident that the hazard would persist through the duration issued, but did not extend the duration to appropriately reflect their confidence, or that forecasters felt confident in the duration, but were unsure how to probabilistically forecast the hazard persisting through the duration. These implications may be indicative of an apparent prognostic limitation (i.e., binary aspect of warnings) associated with the current SBW system.

To investigate this effect further, the daily distributions of the ending values from the probabilistic trends from all forecasters, along with the total distribution, are provided in Fig. 8b. The total distribution is skewed toward the low-probability spectrum, with the median value near 30%, although approximately 25% of the distribution exceeds a probability value of 50%. From the daily distributions, it appears that a majority of this 25% is composed of trends drawn early on in the week. Additionally, the median values in the daily distributions decrease throughout the week. The aforementioned discussion supports the notion that, early on, forecasters relied primarily on prior knowledge to issue probabilistic forecasts in the prototype PHI tool, but as the week progressed, forecasters became more familiar with the probabilistic forecasting concepts and adjusted their probabilistic trends accordingly. This finding suggests that more research is needed to identify an ideal shape and decay rate to assign to the forecast probability trends, and the findings of such studies should feed back to the forecaster through training.

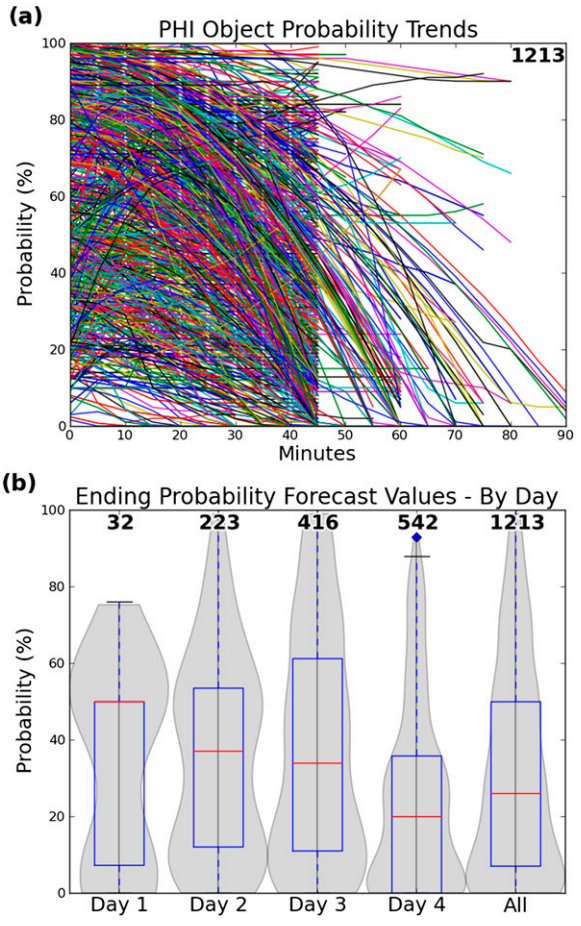

FIG. 8. (a) All probability trends drawn by the forecasters during the experiment and (b) daily distributions [violin plots (http://matplotlib.org/api/pyplot_api.html#matplotlib.pyplot.violinplot) with box-and-whisker diagrams, diamonds are outliers beyond the whisker lengths of Q1 − (1.5 × IQR) and Q3 + (1.5 × IQR); where Q1 indicates the first quantile and Q3 indicates the third quantile] of the ending probability value from each trend drawn by the forecasters. Note that the day 1 median is approximately 50% in (b). Counts of the number of forecasts for each statistic are provided.

Another critical aspect of the forecasts that was analyzed was the amount of time it took forecasters to generate their probabilistic forecasts. The creation time is important, as it can present a potential bottleneck in the information flow. Does the creation of PHI increase mental workload for the forecaster and, if so, is it worthwhile? The creation time distributions for each forecaster, as well as the total distribution for all forecasters, from all initial object creations (excluding updates; 545 total) are given in Fig. 9a. For five of the six forecasters, the interquartile ranges (IQRs) of the creation time distributions are under 2 min, with median values near 1 min. The exception is forecaster E, who had an IQR of 2–3 min and a median value near 2.5 min. Thus, it appears that the

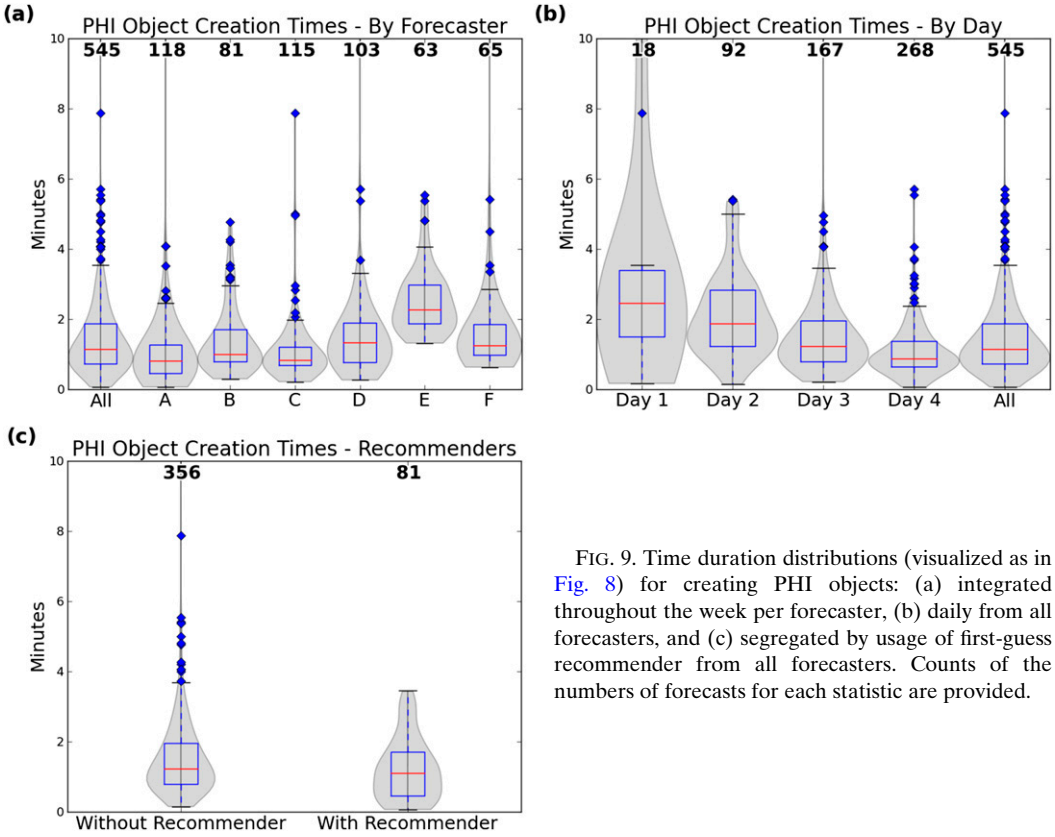

FIG. 9. Time duration distributions (visualized as in Fig. 8) for creating PHI objects: (a) integrated throughout the week per forecaster, (b) daily from all forecasters, and (c) segregated by usage of first-guess recommender from all forecasters. Counts of the numbers of forecasts for each statistic are provided.

forecasters' knowledge and experience in some ways dictate the creation time; however, it is encouraging that the creation times are seemingly quick, at least compared to the typical volume update frequency of 4–5 min from the Weather Surveillance Radar-1988 Doppler (WSR-88D; Crum and Alberty 1993). In addition to the bulk creation time distributions from each forecaster, the combined distributions by day are given in Fig. 9b. A decrease in the spread of the distributions can be noted from days 1 through 4, as well as a decrease in the median creation time from near 2.5 min on day 1 to near 1 min on day 4. This decreasing trend suggests that forecasters initially had to learn how to use the tool, thus implying high mental workload. As forecasters became more familiar with the prototype PHI tool, the mental workload seemingly dropped. Thus, the prototype system appears to be intuitive since it took only a few days for them to become comfortable with issuing probabilistic forecasts.

A comparison of the creation times between PHI objects initially generated without the use of recommender guidance, versus those with recommender guidance, for hail only (437 total) is given in Fig. 9c. It is hypothesized that usage of recommender guidance will substantially decrease the creation time. Although the distribution with recommender appears to be more skewed toward a lower creation time value than the distribution without recommender, the difference between these distributions is, for the most part, negligible. It was observed that forecasters would often adjust the shape of the recommended *k*-means object as well as recalculate the motion vector, thus overriding the radar-derived values and consuming a nearly equivalent amount of time to create the probabilistic forecast. Knowing that forecasters used the recommender guidance less than 20% of the time (Fig. 9c) suggests that forecasters did not trust or were not interested in the guidance information. This implication is perhaps not surprising given the minimal amount of training received prior to the experiment, as well as not knowing the skill and reliability of the recommender guidance (shown later).

b. Control–test experiment verification

During the experiment, two 3-h periods were used to conduct a control–test experiment (i.e., simulated operations) with two displaced real-time severe hail events. During the first (control) period, forecasters used AWIPS II and WarnGEN to issue SVR SBWs (for hail only), with three forecasters working WFO Fort Worth, Texas (FWD), from 2200 UTC 13 June to 0020 UTC 14 June 2012, and three forecasters working WFO

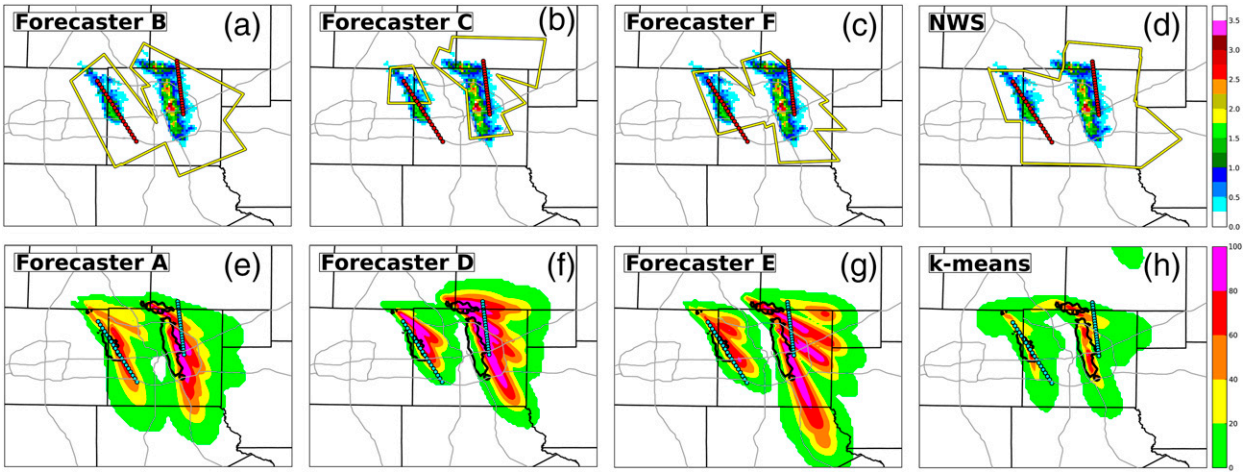

FIG. 10. Comparison of traditional warnings (yellow lines) issued using (a)–(c) WarnGEN vs (e)–(g) probabilistic forecasts (filled contours) issued using the prototype PHI tool in the HWT PHI experiment for the 13 Jun 2012 control–test case. For comparison, (d) the NWS warnings issued during the same time period and (h) the automated probabilistic forecasts generated from the *k*-means objects are provided. MESH observations are provided, denoted by filled contours in (a)–(d), and ≥1 in. as thick black contours in (e)–(h). Additionally, filtered subsets of points (every fifth point) used for the track method of verification are provided for comparisons, denoted as red and blue dots in the top and bottom rows, respectively.

Springfield, Missouri (SGF), from 2300 UTC 5 August to 0120 UTC 6 August 2013 (Fig. 2). During the second (test) period, the forecasters worked the opposite event from the control period and used the prototype PHI tool to issue probabilistic forecasts (for hail only) instead of issuing SVR SBWs. This yielded three sets of SBWs and three sets of probabilistic forecasts to evaluate for each case.

The primary intent of the control–test experiment was to compare probabilistic forecasts and SBWs from the same forecasters, and to document those results such that improvements can be made for future development and experimentation within the context of insights gained from the real-time events. Traditional verification metrics including POD, FAR, CSI, and lead time were computed in addition to producing comparisons of forecast area, false alarm area, and reliability diagrams. However, it is important to acknowledge a few limitations to these comparisons. First, the sample size is notably small and is drawn from two cases of similar convective mode (supercellular) and evolution (right-turning and splitting cells). The durations of the SBWs and probabilistic forecasts differ, though not enough to have a significant impact on the results (not shown). Also, issuance of warnings for hail only could affect the size of the HWT SBW polygons compared to the NWS SBW polygons (issued for hail and wind), and it is unclear how the inclusion (or exclusion) of wind could affect various warning attributes. However, it is noteworthy that no wind reports were obtained for the 13 June 2012 case (though the NWS warning text included wind tags of 50–60 mi h^{-1} for this event), and the hail and wind reports were approximately collocated for the 5 August 2013 case. Thus, the results from the control–test experiment are illustrative of and most applicable to the particular events and hazards examined, as opposed to a generalizable context.

An overview of the time- and space-integrated forecasts from this control–test experiment is provided in Figs. 10 and 11, as well as the NWS SBWs and automatically generated probabilistic forecasts. The automatically generated probabilistic forecasts were produced from objects identified using the *k*-means clustering algorithm by incorporating each object's polygon, motion vector, and MESH-derived POSH value (described in section 2b), in combination with empirically determined default motion uncertainty (8 kt and 15°), duration (45 min), and decay rate for the probabilistic trend (linear) to derive the probabilistic forecast. The corresponding probabilistic forecast areas (per threshold) and SBW areas are shown in Fig. 12. Note that the dichotomous SBW areas are compared to the probabilistic forecasts using lines spanning the probabilistic spectrum (0%–100%), as it is understood that warning decisions are often made with uncertainty about the current and/or future occurrence of the hazard. A similar procedure is carried out in Figs. 13 and 14.

For both cases, the probabilistic forecast areas decrease at an approximate exponential trend with increasing probability thresholds. This is to be expected given the application of a two-dimensional Gaussian distribution to the time-integrated probabilistic forecast

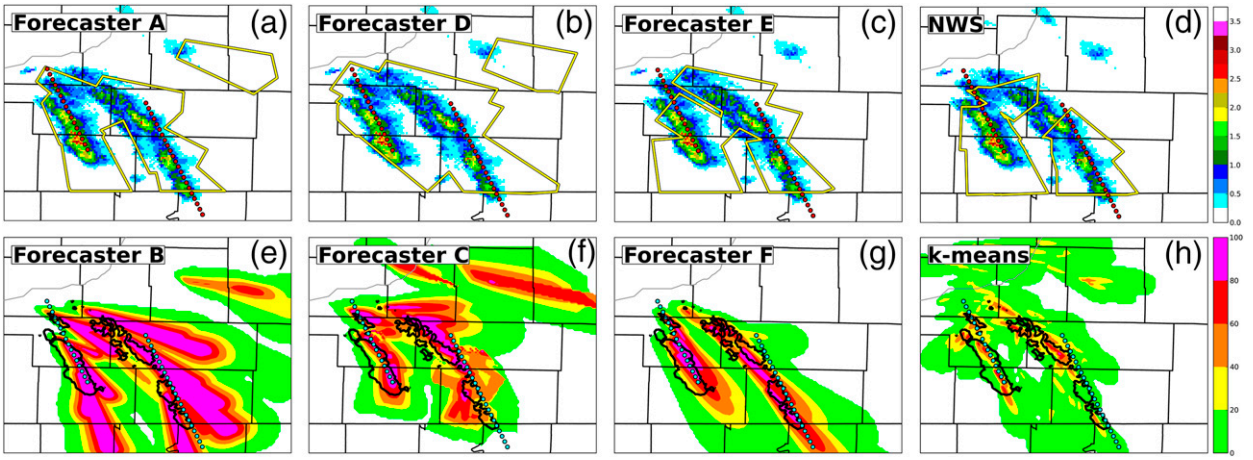

FIG. 11. As in Fig. 10, but for the 5 Aug 2013 control–test case.

swaths. Note that the nonmonotonic decrease of some probabilistic hazard areas in Fig. 12b is attributable to some forecasters using nonelliptical objects to encompass the hazard areas. An intersection between the mean SBW area and mean probabilistic forecast area occurs in the low-probability spectrum (10%–30%). Additionally, the probabilistic forecast areas associated with the automation are almost always less than the corresponding forecaster-generated probabilistic forecast areas. The forecaster-issued SBW areas compare well with the NWS warning area for 5 August 2013, while the forecaster-issued SBW areas are considerably less than the NWS warning area for the 13 June 2012 case. For the latter case, Fig. 10 shows that the NWS SBW warning areas correspond well with geopolitical county boundaries. Geopolitical conformity (i.e., county clipping) of SBWs issued in the HWT was ignored by setting the minimum percentage and area thresholds to zero in WarnGEN.

A comparison of verification metrics (POD, FAR, CSI, and lead time) for the probabilistic forecasts (HWT and automated) to the SBWs (HWT and NWS) is provided in Fig. 13. These statistics were produced using the methodology outlined in the NWS directive for track events (severe hazard traveling over time and space), which relies on a 1-min report interpolation between the first and last LSRs acquired from *Storm Data*. For each case, two track events occurred, as denoted by the two accumulated MESH swaths in Figs. 10 and 11.

For the probabilistic forecasts, the POD decreases and FAR increases with increasing probability thresholds (Fig. 13). The HWT probabilistic forecasts show an increase in POD compared to the HWT SBWs in the low-probability spectrum, but worse FAR overall, for both cases. Compared to the NWS SBWs, the HWT probabilistic forecasts had worse POD for the 13 June 2013 case, better POD for the 5 August 2013 case, and better FAR for both cases. Interestingly, the POD and FAR from the automated probabilistic forecasts were generally worse than the HWT probabilistic forecasts. Consequently, the mean CSI from the HWT probabilistic forecasts shows marginal improvement over the mean CSI from HWT SBWs, applicable in the low-probability spectrum (5%–20%), yet considerably more improvement compared to the NWS SBWs for both cases. The CSIs from the automated probabilistic forecasts are considerably lower than all other forecasts. Intersections between the HWT SBWs and HWT probabilistic forecast means occur for POD and CSI, and this intersection occurs in the low- to midprobability spectrum (less than 40%–50%). No intersection occurred with the mean FARs.

It is important to note that the degradation of the verification metrics in the mid- to high-probability spectrum is likely attributable to the decreasing spatial extent associated with increasing probability thresholds noted in Fig. 12, along with a limitation in using an interpolation between the first and last LSRs from an event. Probabilities in the mid- to high-probability spectrum were typically valid for a portion of the total forecast duration. Meanwhile, interpolated reports were accumulated over the total forecast duration, resulting in points inevitably landing outside of the mid- to high-probability forecast areas. Additionally, subsets of the track points used in the verification are shown in Figs. 10 and 11, highlighting that the interpolation of points between the first and last LSR does not always coincide geospatially with the radar-indicated hazard areas. Thus, the dislocation in time and space of points resulting from the interpolation likely compounds the degradation of verification metrics in the mid- to high-probability spectrum.

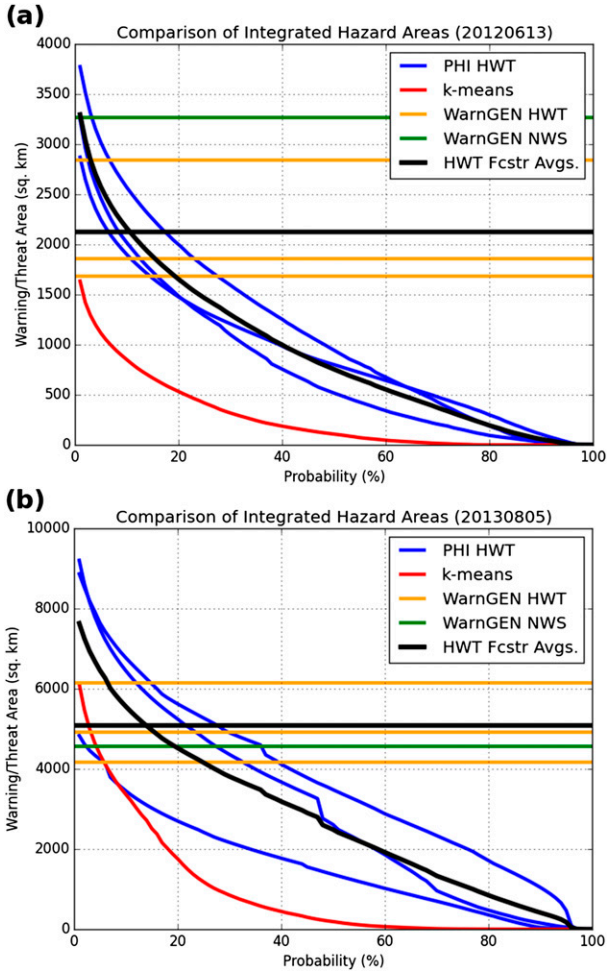

FIG. 12. Comparison of the geospatially integrated area of the WarnGEN-issued warnings vs the geospatially integrated area of the probabilistic forecasts (thresholds 1%–100%) for the (a) 13 Jun 2012 and (b) 5 Aug 2013 control–test experiment cases. The horizontal lines represent the values for the forecasts that do not convey probabilities (WarnGEN NWS and WarnGEN HWT) and are shown as a reference to compare to the range of values for the probabilistic forecasts.

The lead times for the probabilistic forecasts decrease with increasing probability thresholds. Additionally, lead times were longer from all of the probabilistic forecasts, some exceeding 1 h, compared to all of the SBWs in both cases. Interestingly, the intersection point between the mean lead time from the HWT probabilistic forecasts and the HWT SBWs occurs in the high-probability spectrum (75%–80%), thus representing a departure from previously noted intersection points. The substantial increase in lead time is likely attributable to the flexibility in extending the duration of the hazard-following objects combined with low values of probability. Although increases in lead time may give people more time to prepare and make better decisions prior to a natural disaster, the event of 31 May 2013 suggests that providing long lead times, combined with inconsistent messaging and public anxiety, can result in a negative public response (NOAA 2014a). Significant efforts are needed to build on prior research attempting to understand public decision-making with long lead times and PHI for severe convective weather (e.g., Hoekstra et al. 2011; Ash et al. 2014).

In addition to the traditional verification metrics, an alternative metric called false alarm area (FAA) percentage was computed to ascertain the amount of area falsely denoting the location of the hazard. This metric is used to partially address the aforementioned limitation with directly applying the current verification methodology to probabilistic forecasts. The FAA percentage was computed by intersecting MESH areas $\geq$1 in. with (i) the probabilistic forecasts (per threshold) and (ii) the SBW polygons and taking the inverse of the intersection to the total area ratio. Figure 14a shows a decreasing trend in FAA percentage with increasing probability thresholds for most forecasters, and the FAA percentages are generally less than those from SBWs. Additionally, the automatically generated forecasts produce a distinctly lower FAA percentage compared to all other forecasts. Note that the volatility in the false alarm areas at probability values nearing 100% is attributable to a small sample size of forecasts at these large values. Most of the probabilistic FAA percentage values do not depart from SBW FAA percentage values until the mid- to high-probability spectrum.

Given the issues previously noted in section 3a with Fig. 8a with regard to the probability trends, it is hypothesized that the inclusion of high-probability values at long forecast times impacts the results in Fig. 14a. Thus, a systematic operation was performed to generate alternative probabilistic forecasts using the initial value from each forecaster's probability trend forecast and linearly decaying the probability trend to zero through the specified duration. The results of this systematic operation are given in Fig. 14b, and a notable decrease in FAA percentage throughout the forecast probability spectrum is apparent. Although an ideal shape of the probability forecast trends has not been determined, the improvements evident in Fig. 14b are supportive of the aforementioned issues with creating trends noted in section 3a and a need to resolve those issues through continued research and forecaster training. Additionally, the departure of the FAA percentage from automatically generated forecasts compared to the human-generated forecasts emphasizes a need to incorporate the recommender guidance information into the forecast process.

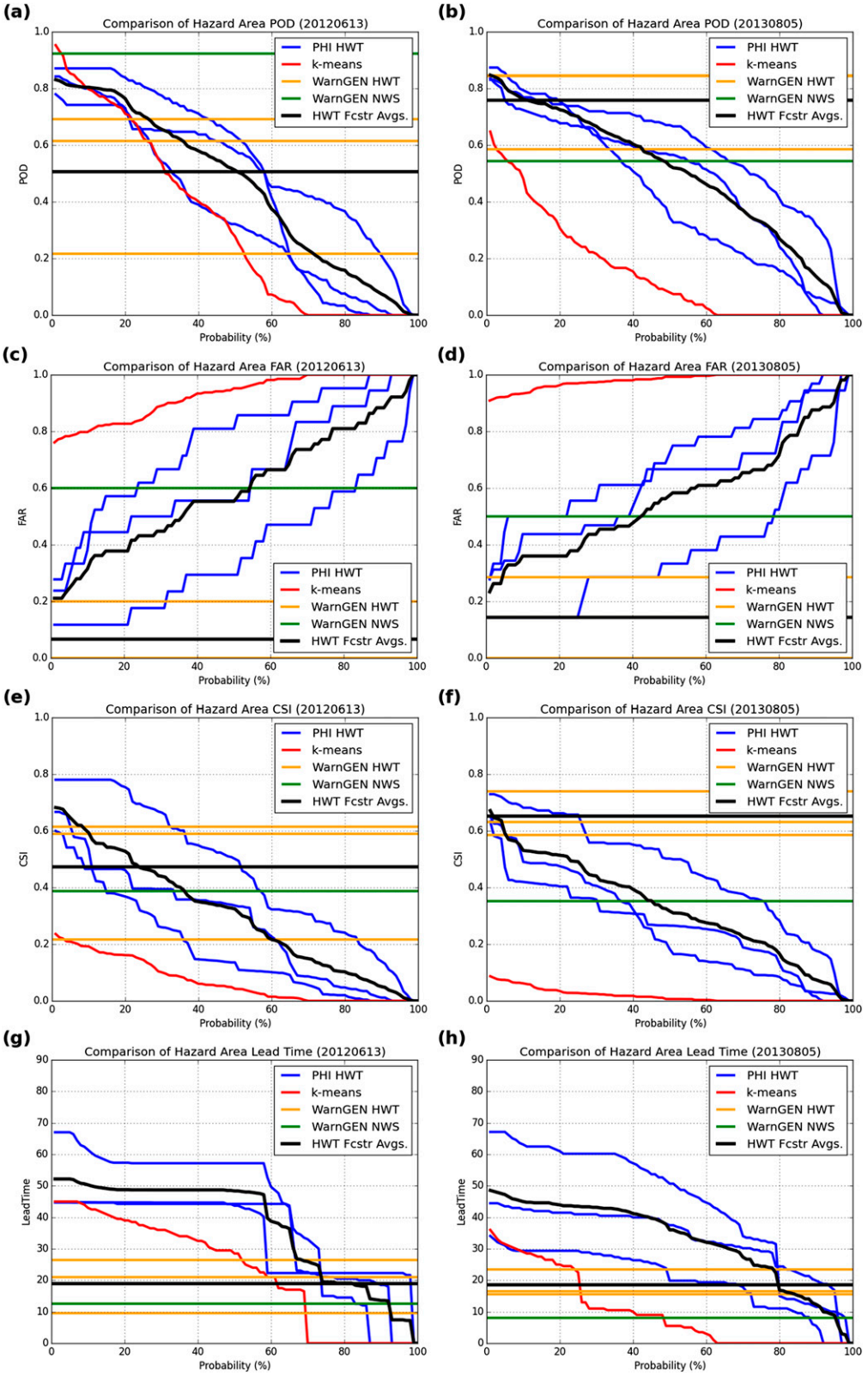

FIG. 13. Comparison of verification metrics (a),(b) POD; (c),(d) FAR; (e),(f) CSI; and (g),(h) lead time computed using the NWS directive method for WarnGEN-issued SBWs and probabilistic forecasts (thresholds 1%–100%) for the (left) 13 Jun 2012 and (right) 5 Aug 2013 control–test cases.

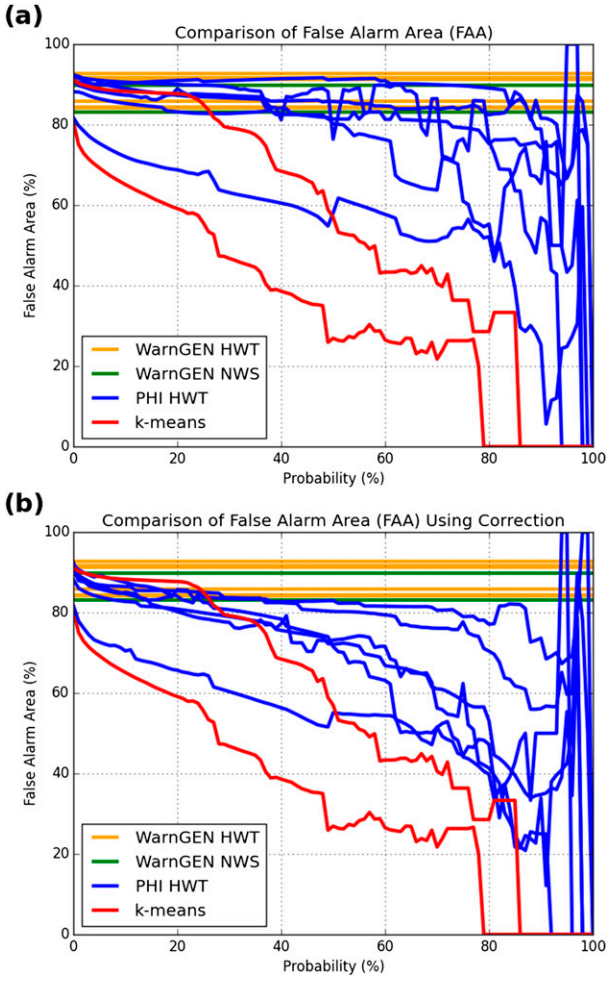

FIG. 14. (a) Mean FAA percentage computed as a function of probability threshold for each human-generated probabilistic forecast compared to SBWs and automated probabilistic forecasts (k-means 20-km^2 length scale) from the 13 Jun 2012 and 5 Aug 2013 control–test cases. (b) As in (a), but for probabilistic forecasts that were systematically adjusted probability trends that linearly decay to zero.

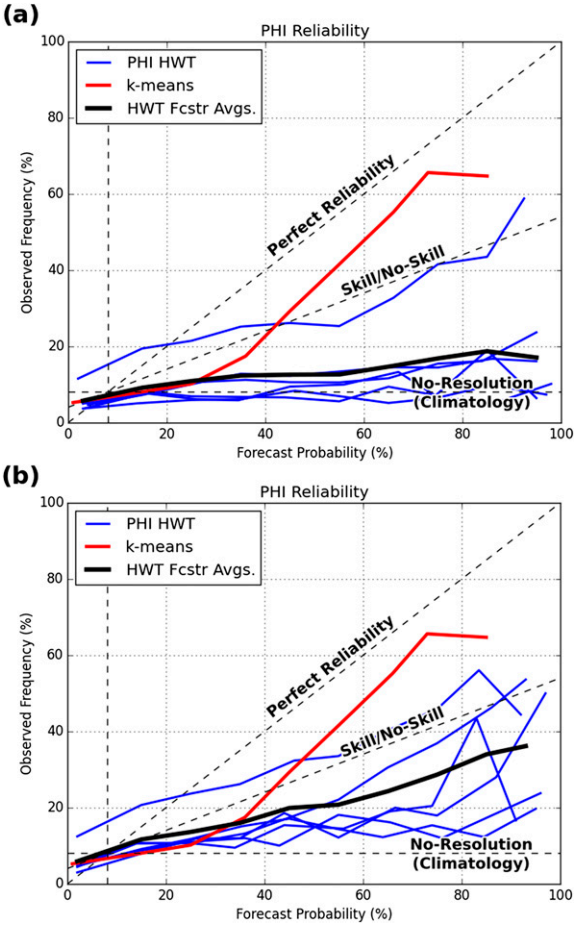

FIG. 15. (a) Reliability diagram displaying the probabilistic forecasts from each forecaster and the mean of all forecasters, verified using MESH values ≥1 in. from the 13 Jun 2012 and 5 Aug 2013 control–test cases. For comparison, probabilistic forecasts using k-means clusters at the 20-km^2 length scale were generated using constant motion uncertainty values (8 kt and 15°) and a linear decay rate for the recommended probability value. (b) As in (a), but for probabilistic forecasts that were systematically adjusted probability trends that linearly decay to zero.

Finally, a comparison of the reliability from the HWT probabilistic forecasts to the automated probabilistic forecasts is provided in Fig. 15a. To compute the reliability, the accumulated MESH values ≥1 in. were treated as observations of severe hail for each probability forecast issued. For each forecaster, the frequency of observed severe hail was tabulated in bins with 10% range (i.e., number of observations compared to number of forecast values within a given range). It is important to note that MESH is used to compute the diagnostic probabilities for the automated probabilistic forecasts, which could result in some correspondence between the analyses of the forecasts and observations. However, the previously noted prognostic and geospatial assumptions used to analyze and generate these forecasts have no direct relationship to the accumulated MESH used as observations throughout the forecast periods.

The HWT forecasts show some reliability in the low-probability spectrum, but quickly deviate toward overforecasting in the mid- to high-probability spectrum, below the no-skill line and close to the no-resolution line. Interestingly, the trends from five of the six forecasters are quite consistent. The one exception, who exhibited more reliability in the mid- to high-probability spectrum but following the skill–no-skill line, was an experienced forecaster quite familiar with the concept of probabilistic forecasting for severe convective hazards. Note that the skill–no-skill line is situated halfway between climatology and perfect reliability.

The systematic operation performed to produce the results in Fig. 14b was extended to Fig. 15b. Compared to Fig. 15a, this simple adjustment improved the forecast reliability in the mid- to high-probability spectrum by 10%–20%, although most of the forecasts in this spectrum still overforecast and are below the skill–no-skill line. The one exception noted in Fig. 15a shows modest improvement above the no-skill line. It can also be noted in Figs. 15a and 15b that the automated probabilistic forecasts from the 20-km^2 length scale objects show considerably more reliability, even though these forecasts showed degraded verification metrics. These results again emphasize a need to incorporate recommender guidance into the forecasts unless the forecaster can be certain of a severe event occurring that is beyond the capability of the radar-derived guidance information. However, it seems that, for all-around improvement within the context of the two cases tested, a balance is needed between the infusion of rapidly updating real-time guidance information and the pattern recognition abilities of well-trained forecasters. This assertion is supported by human factors research suggesting that the incorporation of reliable guidance in uncertainty can elevate the performance of a nonexpert to that of an expert (e.g., Kirschenbaum et al. 2014).

c. *Forecaster feedback*

Subjective feedback was formally gathered throughout the experiment to gain insight into the concepts of the prototype system that worked well, needed improvement, or needed to be rethought. Forecasters consistently thought that by issuing probabilistic forecasts as opposed to deterministic SBWs, they were freed of the decision to issue or not issue their forecast. Relaxation of on-the-fence decisions should allow for low-probability hazard information to be available to the public more readily than in the SBW system. However, inclusion of low-probability hazard information will likely require forecasters to monitor and update more hazard areas and/or potential hazard areas than in the SBW system. During the experiment, forecasters felt comfortable monitoring approximately four or five hazard areas simultaneously, and thereafter the mental workload significantly increased with increasing hazard responsibility. The workload was further complicated when storms underwent a complex evolution (e.g., splitting and merging cells, upscale growth to linear convective system). To combat these workload issues, forecasters suggested improvements in recommender guidance in ways that would more easily allow them to prioritize hazards and update rapidly, versus those that can be automated and/or updated as needed. Other workload mitigation strategies include sectorizing, either by geographic region or by hazard type, or grouping the hazards appropriately, any of which likely have advantages and disadvantages.

During the experiment, several forecasters felt inclined to interrogate the storms and the environments more intently with the prototype system. This feedback suggests that forecasters will have enhanced situational awareness and supports the notion of sectoring forecasters by geographic region, as opposed to hazard type. However, not all political boundaries can be ignored with the prototype system. Specifically, CWAs denote the geographic area of responsibility for each of the 122 WFOs in the NWS, and current policy disallows neighboring offices to issue warnings beyond their CWA boundaries. This restriction may pose a major challenge to operation implementation from a forecast consistency standpoint when hazards reside in and/or are projected to move into two or more CWAs simultaneously. However, this restriction may also present an opportunity to develop and test various handoff procedures and collaboration tools to establish consistency, especially if recommender guidance can be made uniformly available to the WFOs.

4. Conclusions

In this study, the early configuration of a new method was introduced for hazard identification and prediction using PHI. This method was tested in the 2014 HWT PHI experiment. Findings from this experiment were documented, including the prototype PHI tool usage, control–test experiment verification, and forecaster feedback. The collective findings documented herein will be used to inform future improvements and training for the prototype PHI tool and to inform future iterations of development and testing in the HWT.

When forecasters used the prototype PHI tool, selection of the default hazard duration of 45 min occurred approximately 75% of the time. Additionally, a significant number of probability trends ended abruptly at a nonzero value, though these ending probability values decreased throughout the week. These findings highlight potential incomplete aspects of the forecasts, perhaps related to unfamiliarity with how to extend the duration (probabilistic prediction) to match the forecaster's confidence in probabilistic prediction (duration). Forecasting tools should include training that encourages forecasters to connect their forecast durations with their probabilistic forecast trends and vice versa, as well as include tools that engage forecasters in this thought process. In the future, it would be advantageous to develop recommender guidance to the forecaster on storm longevity based on environmental characteristics,

perhaps through convection-allowing storm-scale numerical weather prediction models or though storm-based climatological information from the Multiyear Reanalysis of Remotely Sensed Storms (MYRORSS; Cintineo et al. 2011; Ortega et al. 2012) project.

An encouraging aspect of the prototype PHI tool usage is that forecasters were able to create their probabilistic forecasts quickly. Five of the six forecasters typically took from 30 s to 2 min to create their forecasts, while the remaining forecaster typically took between 2 and 3 min. As forecasters became more familiar with the prototype PHI tool, the creation times decreased and ranged from approximately 45 s to 1.5 min by day 4. These results suggest that forecasters found the tool intuitive to use despite minimal training. Additionally, the usage of recommender guidance did not result in a substantial difference in creation time. As modifications are made to the prototype PHI tool, efforts should be made to maintain these reported values and explore ways to improve times associated with recommender guidance, provided that the guidance is skillful and reliable. It is hypothesized that providing forecasters with a more representative first-guess object shape (e.g., Cintineo et al. 2014), as opposed to ellipses, will improve the performance of incorporating of recommender guidance with the forecaster. On the other hand, the inconsistencies observed between individual forecasters and the lack of baseline workload statistics highlights a need for human factors research with NWS warning operations to understand more about the warning decision-making process (Boustead and Mayes 2014).

The NWS directive verification method for track events was used to compare verification metrics from SBWs issued in the HWT and from the NWS to probabilistic forecasts issued using the prototype PHI tool in the HWT for two severe hail events. In the low spectrum of the probabilistic forecasts, the POD and CSI were marginally higher for the probabilistic forecasts than the SBWs, with a modest reduction in FAR compared to the NWS SBWs (increase compared to the HWT SBWs) and a significant increase in lead time. Additionally, the low spectrum of the probabilistic forecasts was found to be somewhat reliable. These results suggest that, for the events tested, the PHI concepts uphold, and in some areas improve upon, the metrics from SBWs. However, all of the probabilistic forecast verification metrics degrade with increasing probabilistic thresholds. This degradation is likely attributable to the smaller coverage associated with large probability thresholds combined with a dislocation in time and space of points interpolated between the first and last LSRs of the event. These limitations highlight a need for new verification methodologies, such as the practically perfect methodology (Davis and Carr 2000; Hitchens et al. 2013), that are more applicable to probabilistic forecasts. With all of the aforementioned results, it is important to be mindful of the small sample size (two cases and six forecasters). Future experiments should include more control–test case study experiments with a diverse set of convective modes and evolutions from throughout the United States to expand upon the results documented in this study.

Verification metrics from the automatically generated probabilistic forecasts from 20-km^2 k-means clusters were worse than the metrics from the human-generated probabilistic forecasts. However, the FAA percentage and reliability of these forecasts were better than the human-generated forecasts. Systematically adjusting the human-generated probabilistic trends to linearly decay toward zero improved both the FAA percentage and reliability of the forecasts, but still produced more FAA percentage and remained less reliable than the automatically generated forecasts. These results highlight that more research is needed to identify the ideal shape and decay rate for probabilistic trends to inform future training for the forecaster. Additionally, these results suggest a need to integrate the human decision-making and pattern recognition abilities with real-time guidance information, in addition to training.

Efforts are under way to improve the presentation of guidance information within the prototype PHI tool to give forecasters the ability to rapidly select which hazards to closely monitor and update versus those requiring less attention and perhaps leave to automation. It is planned to incorporate guidance information such as real-time near-storm environment information and severe potential (Cintineo et al. 2014), storm-based climatological information from the MYRORSS project, and probabilistic output from a Warn-on-Forecast system. Additionally, verification metrics, such as those presented or proposed herein, will be made available to forecasters during or immediately after an event to reinforce aspects of training that can improve future forecasts. With these changes implemented, we plan to conduct more control–test experiments with events of varying convective modes and evolutions, and begin addressing the potential issues associated with geographic sectoring and CWA handoff procedures.

Acknowledgments. The authors thank Elizabeth Mintmire, Valliappa Lakshmanan, Jim LaDue, Harold Brooks, Patrick Marsh, Mike Magsig, Jack Kain, Adam Clark, Mike Coniglio, Lou Wicker, Tracy Hansen, Kevin Manross, Bryon Lawrence, and Kim Elmore for their helpful guidance and suggestions to this project. Additionally, the authors thank the six NWS forecasters

who participated in the HWT experiment for their diligent efforts and honest feedback aimed at improving the prototype PHI tool and making progress toward the FACETs vision. This manuscript was substantially improved thanks to the constructive comments of Dr. Mathew Bunkers, Phillip Schumacher, and two anonymous reviewers. Partial support for this research was provided by NOAA (Grant NA11OAR320072), NSSL, and MDL. 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, doi: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, doi:10.1175/WCAS-D-13-00021.1.

Atkins, N. T., K. M. Butler, K. R. Flynn, and R. M. Wakimoto, 2014: An integrated damage, visual, and radar analysis of the 2013 Moore, Oklahoma, EF5 tornado. *Bull. Amer. Meteor. Soc.*, **95**, 1549–1561, doi:10.1175/BAMS-D-14-00033.1.

Boustead, J. M., and B. E. Mayes, 2014: The role of the human in issuing severe weather warnings. *Proc. 27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 4B.2. [Available online at https://ams.confex.com/ams/27SLS/webprogram/Paper254547.html.]

Brotzge, J., and W. Donner, 2013: The tornado warning process: A review of current research, challenges, and opportunities. *Bull. Amer. Meteor. Soc.*, **94**, 1715–1733, doi:10.1175/BAMS-D-12-00147.1.

Burgess, D., and Coauthors, 2014: 20 May 2013 Moore, Oklahoma, tornado: Damage survey and analysis. *Wea. Forecasting*, **29**, 1229–1237, doi:10.1175/WAF-D-14-00039.1.

Calhoun, K. M., T. M. Smith, D. M. Kingfield, J. Gao, and D. J. Stensrud, 2014: Forecaster use and evaluation of real-time 3DVAR analyses during severe thunderstorm and tornado warning operations in the Hazardous Weather Testbed. *Wea. Forecasting*, **29**, 601–613, doi:10.1175/WAF-D-13-00107.1.

Cintineo, J. L., T. M. Smith, V. Lakshmanan, and S. Ansari, 2011: An automated system for processing the Multi-Year Reanalysis of Remotely Sensed Storms (MYRORSS). *Proc. 27th Conf. on Interactive Information Processing Systems*, Seattle, WA, Amer. Meteor. Soc., J9.3. [Available online at https://ams.confex.com/ams/91Annual/webprogram/Paper182332.html.]

——, 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, doi:10.1175/WAF-D-13-00113.1.

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, doi:10.1175/BAMS-D-11-00040.1.

Coleman, T. A., K. R. Knupp, J. Spann, J. B. Elliott, and B. E. Peters, 2011: The history (and future) of tornado warning dissemination in the United States. *Bull. Amer. Meteor. Soc.*, **92**, 567–582, doi:10.1175/2010BAMS3062.1.

Crum, T. D., and R. L. Alberty, 1993: The WSR-88D and the WSR-88D Operational Support Facility. *Bull. Amer. Meteor. Soc.*, **74**, 1669–1687, doi:10.1175/1520-0477(1993)074<1669:TWATWO>2.0.CO;2.

Davis, C., and F. Carr, 2000: Summary of the 1998 Workshop on Mesoscale Model Verification. *Bull. Amer. Meteor. Soc.*, **81**, 809–819, doi:10.1175/1520-0477(2000)081<0809:SOTWOM>2.3.CO;2.

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

Ferree, J., 2006: NOAA/National Weather Service's storm-based warnings. Preprints, *23rd Conf. Severe Local Storms*, St. Louis, MO, Amer. Meteor. Soc., P11.6. [Available online at https://ams.confex.com/ams/pdfpapers/115513.pdf.]

Hansen, T. L., and Coauthors, 2010: Hazard information services vision. Preprints, *26th Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Atlanta, GA, Amer. Meteor. Soc., 6B.4. [Available online at https://ams.confex.com/ams/90annual/techprogram/paper_159734.htm.]

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, doi: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, doi:10.1175/WAF-D-14-00042.1.

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

Hoekstra, S., K. Klockow, R. Riley, J. Brotzge, H. Brooks, and S. Erickson, 2011: A preliminary look at the social perspective of Warn-on-Forecast: Preferred tornado warning lead time and the general public's perceptions of weather risks. *Wea. Climate Soc.*, **3**, 128–140, doi:10.1175/2011WCAS1076.1.

Humphrey, T. W., V. Lakshmanan, T. M. Smith, K. L. Ortega, B. T. Smith, and R. L. Thompson, 2014: Automated storm classification for the development of probabilistic hazard information. *Proc. 26th Conf. on Weather Analysis and Forecasting*, Atlanta, GA, Amer. Meteor. Soc., J3.4. [Available online at https://ams.confex.com/ams/94Annual/webprogram/Paper239857.html.]

Karstens, C. D., T. M. Smith, K. M. Calhoun, A. J. Clark, C. Ling, G. J. Stumpf, and L. P. Rothfusz, 2014: Prototype tool development for creating probabilistic hazard information for severe convective weather. *Second Symp. on Building a Weather-Ready Nation: Enhancing Our Nation's Readiness, Responsiveness, and Resilience to High Impact Weather Events*, Atlanta, GA, Amer. Meteor. Soc., 2.2. [Available online at https://ams.confex.com/ams/94Annual/webprogram/Paper241549.html.]

Kirschenbaum, S. S., J. G. Trafton, C. D. Schunn, and S. B. Trickett, 2014: Visualizing uncertainty: The impact on performance. *Hum. Factors: J. Hum. Factors Ergon. Soc.*, **56**, 509–520, doi:10.1177/0018720813498093.

Kuhlman, K. M., T. M. Smith, G. J. Stumpf, K. L. Ortega, and K. L. Manross, 2008: Experimental probabilistic hazard information in practice: Results from the 2008 EWP Spring Program. Preprints, *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., 8A.2. [Available online at https://ams.confex.com/ams/pdfpapers/142027.pdf.]

Lakshmanan, V., I. Adrianto, T. Smith, and G. Stumpf, 2005: A spatiotemporal approach to tornado prediction. *Proc. Int. Joint Conf. on Neural Networks*, Montreal, QC, Canada, IEEE, doi:10.1109/IJCNN.2005.1556125.

——, 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, doi:10.1175/2008JTECHA1153.1.

Lee, B. D., and R. B. Wilhelmson, 1997: The numerical simulation of non-supercell tornadogenesis. Part I: Initiation and evolution of pretornadic misocyclone circulations along a dry outflow boundary. *J. Atmos. Sci.*, **54**, 32–60, doi:10.1175/1520-0469(1997)054<0032:TNSONS>2.0.CO;2.

Markowski, P., and Y. Richardson, 2014: What we know and don't know about tornado formation. *Phys. Today*, **67**, 26–31, doi:10.1063/PT.3.2514.

——, E. N. Rasmussen, and J. M. Straka, 1998: The occurrence of tornadoes in supercells interacting with boundaries during VORTEX-95. *Wea. Forecasting*, **13**, 852–859, doi:10.1175/1520-0434(1998)013<0852:TOOTIS>2.0.CO;2.

Murphy, A. H., 1993: What is a good forecast? An essay on nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

NOAA, 2005: NOAA remembers the Midwest's deadly 1965 Palm Sunday tornado outbreak. [Available online at www.noaanews.noaa.gov/stories2005/s2418.htm.]

——, 2007: Storm-based warnings team report. NOAA Tech. Rep., 45 pp. [Available online at www.nws.noaa.gov/sbwarnings/docs/Polygon_Report_Final.pdf.]

——, 2011: NWS Central Region service assessment: Joplin, Missouri, tornado. National Weather Service, 35 pp. [Available online at www.nws.noaa.gov/om/assessments/pdfs/Joplin_tornado.pdf.]

——, 2012: The historic tornadoes of April 2011. NWS Service Assessment, 76 pp. [Available online at www.nws.noaa.gov/os/assessments/pdfs/historic_tornadoes.pdf.]

——, 2014a: May 2013 Oklahoma tornadoes and flash flooding. NWS Service Assessment, 63 pp. [Available online at www.nws.noaa.gov/os/assessments/pdfs/13oklahoma_tornadoes.pdf.]

——, 2014b: NOAA 5 year research and development plan 2013–2018. [Available online at http://nrc.noaa.gov/CouncilProducts/ResearchPlans/5YearRDPlan/NOAA5YRPHome/StrategicApproachtoRD/Goals,Questions,Objectives,Targets/WeatherReadyNation/WeatherQ1.aspx.]

——, 2015: NWS performance management—StormGen reports. [Available online at https://verification.nws.noaa.gov/stormdat/stormgen/index.aspx.]

Ortega, K. L., 2008: Severe weather warnings and warning verification using threat areas. M.S. thesis, School of Meteorology, University of Oklahoma, 50 pp.

——, T. M. Smith, K. L. Manross, A. G. Kolodziej, K. A. Scharfenberg, A. Witt, and J. J. Gourley, 2009: The Severe Hazards Analysis and Verification Experiment. *Bull. Amer. Meteor. Soc.*, **90**, 1519–1530, doi:10.1175/2009BAMS2815.1.

——, ——, J. Zhang, C. Langston, Y. Qi, S. E. Stevens, and J. E. Tate, 2012: The Multi-Year Reanalysis of Remotely Sensed Storms (MYRORSS) project. *Proc. 26th Conf. on Severe Local Storms*, Nashville, TN, Amer. Meteor. Soc., P4.74. [Available online at https://ams.confex.com/ams/26SLS/webprogram/Paper211413.html.]

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, doi:10.1002/2014EO360001.

Simmons, K. M., and D. Sutter, 2005: WSR-88D radar, tornado warnings, and tornado casualties. *Wea. Forecasting*, **20**, 301–310, doi:10.1175/WAF857.1.

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

——, and Coauthors, 2013: Progress and challenges with warn-on-forecast. *Atmos. Res.*, **123**, 2–16, doi:10.1016/j.atmosres.2012.04.004.

Stumpf, G. J., 2012: The Experimental Warning Program at the NOAA Hazardous Weather Testbed: Experimenting with new warning verification and service techniques. NWA Professional Development Committee Feature Article 2012-1, 5 pp. [Available online at http://www.nwas.org/committees/professionaldevelopment/NWAPD-2012-1-Stumpf.pdf.]

——, T. M. Smith, K. Manross, and D. L. Andra, 2008: The Experimental Warning Program 2008 Spring Experiment at the NOAA Hazardous Weather Testbed. *Extended Abstracts, 24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., 8A.1. [Available online at https://ams.confex.com/ams/pdfpapers/141712.pdf.]

Sutter, D., and S. Erickson, 2010: The time cost of tornado warnings and the savings with storm-based warnings. *Wea. Climate Soc.*, **2**, 103–112, doi:10.1175/2009WCAS1011.1.

Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243–1261, doi:10.1175/1520-0434(2003)018<1243:CPSWSE>2.0.CO;2.

Torres, S. M., and C. D. Curtis, 2007: Initial implementation of super-resolution data on the NEXRAD network. Preprints, *23rd Conf. on International Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, San Antonio, TX, Amer. Meteor. Soc., 5B.10. [Available online at https://ams.confex.com/ams/pdfpapers/116240.pdf.]