Creation and Use of Automated Storm-Based Probabilistic Hazard Information for National Weather Service Forecasting and Warnings

Kristin M. Calhoun[1], Clarice Satrio[2,1], Rebecca Steeves[2,1], Thea Sandmæl[2,1], P. Adrian Campbell[2,1], and Kodi L. Berry[1]

[1]NOAA/OAR/National Severe Storms Laboratory, Norman, OK;
[2]Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, OK

NOAA Technical Report
https://doi.org/10.25923/fpgn-1456
September 2025

U.S. Department of Commerce
National Oceanic and Atmospheric Administration
Oceanic and Atmospheric Research
National Severe Storms Laboratory

Table of Contents

Executive Summary 3

1. Introduction 4
2. Data and Methods 6
    - Underlying Machine Learning Algorithms 6
    - Hazard Identification and Shape 7
    - Storm Tracking and Motion 8
    - Domain Selection and Filtering Methods 9
    - Creation of Automated PHI Plumes 11
    - PHI within GibsonRidge Software 12
    - NWS Training and Feedback 12
3. Results 14
    - Content Analysis and Findings 14
    - Performance Metrics of Automated Severe and Tornado PHI 18
    - Case Study Evaluation and Analysis 20
4. Recommendations for Use and Ongoing Development 28
    - Best Practices for Using PHI in Operations 28
    - Computing Resources Necessary for PHI in Operations 29
    - Archive Case Review and On-Demand Verification 30
    - Integrating Predicted Probabilities 30
    - Warning Recommender within Hazard Services Severe 31
5. Acknowledgements 33
6. References 33

Executive Summary

Select National Weather Service (NWS) forecast offices were given access to Probabilistic Hazard Information (PHI) forecast plumes, derived from machine learning algorithms, for evaluation during real-time severe weather events. Science and Operations Officers (SOOs), along with other operational meteorologists, provided feedback to the development team through *Slack*, online surveys, and Google Meets. Most feedback fell into two categories: 1) system performance during real-time events and any associated errors, and 2) the use of the data in operational decision making and communications with partners. Feedback regarding performance was weighted highly by the development team and was immediately addressed for overall improvements. The primary issue for system performance was regarding storm motion accuracy for new and developing convection. This was addressed through the addition of new calculation methods and a transition from the 13-km Rapid Refresh (RAP) to the 3-km High-Resolution Rapid Refresh (HRRR).

NWS forecasters provided examples on how PHI plumes supported operations and Impact-based Decision Support Services (IDSS). Probability trends were used to quantify evolving risks for partners, both visually and in text-based communication. PHI plumes provided additional confidence in warning decisions for tornadoes and severe weather. The plumes also assisted with warning polygon creation, both in regards to how large of a region to cover relative to the storm or line of storms, and the area covered by the downstream warning. Forecasters also reported that PHI plumes helped them determine whether to extend warnings or allow them to expire.

Automated PHI plumes demonstrated overall skill in supporting warning and situational awareness. Severe PHI performed best at moderate to high probabilities (30–70%), while tornado PHI performed best at higher probabilities (≥50%). Both products exhibited some over forecasting, which forecasters found useful for anticipating storm evolution and supporting decision making. Case studies of Hurricane Beryl (July 2024) and the North Dakota supercell outbreak (June 2025) showed that PHI plumes generally aligned well with NWS warnings and tornado reports, providing early and consistent guidance for storm motion, warning issuance, and downstream communication.

Looking ahead, goals include training field meteorologists on quality control procedures and probability calibration to build PHI expertise. Development is underway to provide access to an expanded archive of events and on-demand verification for specific time periods and regions. Additionally, initial testing is exploring PHI as a warning recommender within NWS software, integrating high-resolution numerical guidance from Warn-on-Forecast, and evaluating methods to visualize and communicate the watch-to-warning gap to partners.

1. Introduction

Experiments in the NOAA Hazardous Weather Testbed (HWT) have shown that storm-based Probabilistic Hazard Information (PHI) can support forecasters in making more effective short-term forecasts and warning decisions (e.g., Karstens et al. 2015, 2018; Calhoun et al. 2024). PHI integrates data from multiple platforms and algorithms to generate storm-based probabilities of individual hazards along with projected storm tracks (Fig. 1). Forecasters have used PHI to assess storm trends, improve situational awareness and confidence during operations, and guide severe thunderstorm and tornado warning decisions. In addition, PHI shows promise for increasing both consistency and accuracy of NWS warnings.

HWT experiments have also demonstrated that PHI can help forecasters produce more reliable probabilistic forecasts with less false alarm area than human-generated forecasts (Karstens et al. 2015). Forecasters also reported greater confidence in their warning decisions when supported by PHI guidance. At the same time, the probability threshold used for severe thunderstorm warning decisions has varied considerably depending on environmental factors, storm modes, and location (50-95%; Karstens et al. 2018). Feedback from emergency managers and other end-users of PHI data have highlighted a parallel need: while they preferred access to the additional lead time that PHI guidance provided, they still relied on deterministic warnings for specific protective actions, such as activating tornado sirens (Karstens et al. 2018).

The 2014–2017 HWT PHI experiments used both initial and early development versions of the ProbSevere algorithm as a first guess of severe guidance (Cintineo et al. 2014, 2018). Other algorithms were tested to estimate tornado likelihood, but these did not directly generate outlines of hazard locations (also referred to as hazard objects) or probabilities like ProbSevere did for severe PHI. These algorithms required deeper forecaster interpretation and therefore saw limited use (Karstens et al. 2018). Beginning in 2021, however, the Tornado Probability Algorithm (TORP; Sandmæl et al. 2023) offered a direct first guess of tornado likelihood and location for individual storms. Because the algorithms for severe hazards and tornadoes each generated their own threat locations and probabilities, forecasters could compare initial probability values and make more informed decisions about PHI communication for Impact-Based Decision Support Services (IDSS) and warning issuance. In practice during more recent HWT experiments (2021-2024), forecasters created PHI plumes for IDSS needs at lower thresholds, typically between 20–50% for severe storms and 10–40% for tornadoes. As expected, warnings were issued at higher probability thresholds, though some overlap existed: tornado warnings were typically issued at 25–60%, while severe thunderstorm warnings used thresholds of 40–75% (Berry et al. 2024; Calhoun et al. 2024). The consistently lower thresholds for tornadoes, for both warnings issuance and in communication, likely reflects a lower risk tolerance (or higher risk aversion) for missed tornado events, given their potential impact.

The communication of probabilistic risk has the advantage of quantifying the uncertainty of an event. End-users of weather information commonly understand that forecasts and weather warnings in particular come with some degree of uncertainty (Joslyn and Savelli 2010; Joslyn and LeClerc 2012; Kox et al. 2015). Additionally, while those making forecasts may often prefer to convey uncertainty using words (e.g. 'likely'), users prefer numeric probabilities when making consequential decisions (Dhami and Mandel 2022) and probability information generally improves decision quality (Ripberger et al. 2022). One initial concern of early prototypes of PHI was the calibration of probabilities; PHI was initially fully human-derived and different forecasters often chose different probabilities for the same storm (Kuhlman et al. 2010). However, the development and incorporation of machine learning algorithms now provides an initial first guess for probabilities for the forecaster and spread between forecasters for the same case is much lower than during initial experiments without guidance.

The goal of creating automated PHI is to provide meaningful quantification of hazard probabilities and help with more consistent communication as well as warning decisions. Additionally, storm-based PHI is intended to help fill gaps between the watch and warnings as well as provide information for storms that may not meet full warning criteria.

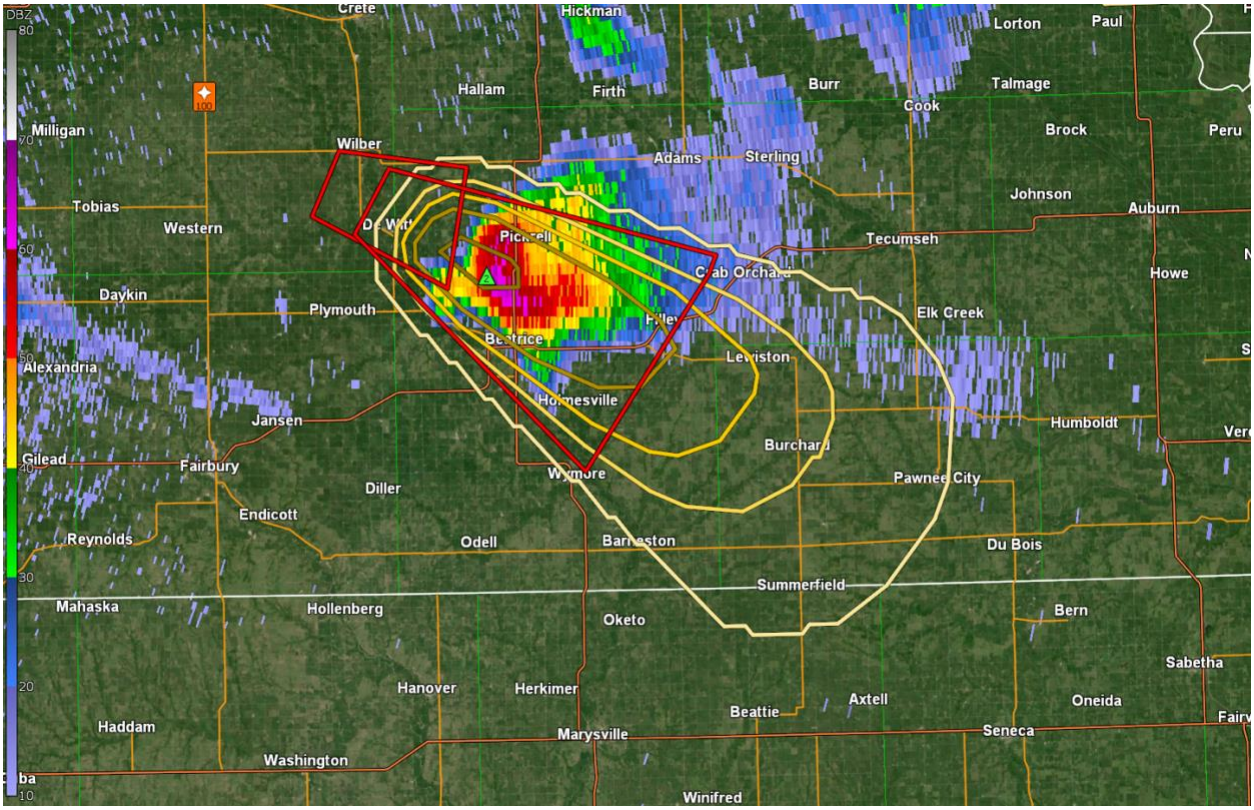

Figure 1: *Automated severe PHI plume (yellow contours, every 20%) and NWS severe thunderstorm warnings (red polygons) for an isolated storm near Beatrice, Nebraska overlaid on 0.5° reflectivity from KUEX on 11 Aug 2023 in GibsonRidge software.*

2. Data and Methods

The creation of fully-automated PHI depends on multiple underlying algorithms as well as multiple quality control methods. Machine learning algorithms provide the basis for individual storm hazard probabilities, but these are not sufficient enough on their own to provide the consistent tracking and motion of real storms (e.g., Karstens et al. 2018). The combination of the methods described herein ideally provides consistent storm-scale guidance on severe and tornadic potential that is reliable across multiple storm environments. As this guidance is used by a growing number of forecasters across a wide range of cases, we receive feedback on examples where improvement is needed. This feedback loop is critical to improve overall performance across a wide range of storm environments and modes.

Underlying Machine Learning Algorithms

Severe PHI plumes use probabilities from ProbSevere version 3 (PSv3; Cintineo et al. 2024). This machine learning model uses a combination of data from multiple sources to provide holistic probabilistic guidance using a data fusion approach. Environmental information is sourced from the High-Resolution Rapid Refresh (HRRR; Dowel et al. 2022). The Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) system provides the radar-derived information, such as merged composite reflectivity, the Maximum Expected Size of Hail (MESH), and merged Azimuthal Shear (AzShear; Mahalik et al. 2017). Additionally, geostationary satellite data provides details on the satellite growth rate and intense convection probability (Cintineo et al. 2020). Lightning data and trends are included from both the Geostationary Lightning Mapper (Rudlosky and Virts 2021) and Earth Networks Total Lightning Data (Zhu et al. 2022). The PSv3 model uses gradient boosted decision trees which uses sequentially trained decision trees to make a probabilistic prediction of each severe hazard (hail, thunderstorm wind, and tornado) as well as a merged probability of any severe occurring over the next hour. The automated severe PHI uses this merged probability as the initial probability for any severe PHI plume.

Tornado probabilities are derived from the Tornado Probability Algorithm (TORP) developed by Sandmæl et al. (2023). Unlike ProbSevere which uses MRMS and a combination of other data sources, TORP is calculated using the 0.5° tilt from WSR-88D radar data and near-storm environmental information sampled from the Rapid Refresh model (RAP; Benjamin et al. 2016). TORP finds potential areas of rotation by first identifying locations with AzShear greater than 0.006 s^{-1} and then sampling additional radar variables (including radial velocity, velocity spectrum width, horizontal and differential reflectivity, specific differential phase, correlation coefficient and the respective gradients of each) within 2.5 km of the AzShear detection. A random forest model is then used to calculate the probability of a tornado using both the radar and environmental properties.

Hazard Identification and Shape

Output from the underlying machine algorithms is used to create hazard objects (or objects). These objects are polygons that outline the geographic area of potential severe storms or tornadoes. However, several additional post-processing steps are required to generate an automated PHI plume with the necessary stability for consistent communication. This is primarily due to the nature of hazard identification and tracking which often result in irregular shapes (particularly when tracked on reflectivity) and erratic behaviors. This was a recurring theme from forecaster feedback in HWT experiments and needed to be addressed.

For severe PHI, the PSv3 detection based on MRMS composite reflectivity serves as the initial object. As mentioned, the use of reflectivity for hazard identification and tracking often results in irregular shapes. Thus, in order to apply uniformity and stability, a confidence interval is calculated surrounding the base object to fit an ellipse to it. This ensures that the ellipse remains aligned with PSv3 while adopting a more consistent shape. The elliptical shape minimizes significant variations in shape and size, which in turn affects PHI plume geographic coverage. However, inconsistencies across time steps can still impact the plume. For example, a notable issue with ProbSevere is track breakages, especially how it handles continuity along linear convective systems, such as when it segments one object into multiple along a line, then merges them again.

Unlike severe PHI, the initial object from tornado PHI is derived from a point location associated with the peak AzShear value as detected by TORP. This is expanded to a 7.5 km radius to define a coverage area relative to the peak probability within the plume. This radius value was chosen through validation of both HWT forecaster-created tornado PHI plumes and practically perfect representations of PHI plumes from historical tornado paths by Gesell (2020). This expansion to 7.5 km more accurately represented the risk area compared to smaller radii values due to a number of factors including: 1) addressing variations of the radar-location placement of the mesocyclone versus tornado damage paths, 2) resolving common tornado deviant-motion possibilities such as the occlusion process, and 3) adequately communicating the appropriate risk to those close to the tornado hazard.

To reduce variability between time steps in terms of both object position and shape, smoothing is applied to the objects (ellipses for severe PHI and point locations for tornado PHI) due to inconsistencies in identification and tracking. This smoothing utilizes a moving average of previous time steps, which results in a spatial lag of the resultant ellipse or point location from the base object. To counteract this lag, a future projection interpolates the storm's position a specified number of minutes forward, based on the Quality Controlled (QC'd) storm motion (discussed later).

Storm Tracking and Motion

Originally storm motion calculations relied on attributes from the underlying machine learning algorithms, but forecaster feedback from HWT experiments indicated that more accurate and consistent storm motions were needed (Karstens et al. 2018). To fully automate the system, without requiring forecasters to correct errors in real-time, improved quality control (QC) was necessary to address common issues such as poor initial motion estimates, offsets due to storm mergers, and other deviant motions. Refining storm motion QC has therefore been a central focus over the course of this project. The current QC approach is described below, with the rationale and timing of improvements discussed in the Results section.

For the initial time steps of a severe object, an initial storm motion is determined through either using adjacent or overlapping severe objects (primary) or a blend of data from the HRRR (secondary). Using motion from adjacent or overlapping severe objects is helpful to avoid having new objects reset to HRRR motion when there is a more established motion estimate available from existing, longer-lived objects. When using HRRR data for an initial estimate, the Supercell Composite Parameter (SCP; Thompson et al. 2004) is approximated to rudimentarily distinguish between storm environments, which then dictates the field to be used. If the SCP is less than two, 0-6 km mean motion is used; otherwise, if the SCP is two or greater, Bunkers Right (Bunkers et al. 2000) is used. We note that the use of SCP is a recent update and is still being refined to handle a range of storm environments. Possible changes might include integrating other fields in addition to SCP.

Over subsequent time steps, the QC process gradually shifts from relying solely on this initial estimate (from either adjacent or overlapping severe objects or HRRR data) to using a centroid tracking approach. A weighted average is applied during this transition period to prevent large or erratic changes in storm motion. This phased approach facilitates a smooth transition between the storm motions while allowing sufficient time for centroid tracking to stabilize and become more dependable. The number of time steps for this transition varies by hazard type; severe objects require a larger number of time steps due to their longer average durations compared to tornado objects. This is also the justification for only using adjacent or overlapping severe objects, rather than tornado objects or both.

Additional QC steps are performed to prevent abrupt changes in storm motion from one time step to the next, especially once transition occurs to centroid tracking. Storm motion is transformed into its zonal and meridional components and each undergoes both smoothing using a moving average over the entirety of its lifetime and through a three-stage filtering process. The first filter eliminates any data points that significantly deviate from previous data points. The second is a Savitzky-Golay filter which is applied to smooth data while maintaining trend information and

improving the signal-to-noise ratio. Finally, a Kalman filter is applied which minimizes variations between consecutive time steps. These additional QC steps are necessary due to the nature of centroid tracking and inherent limitations of the underlying algorithms. For example, if there are track breakages within ProbSevere, the centroid would drastically shift, falsely indicating movement of the storm in that direction. As a final control, if the storm motion exceeds 38.6 m s^{-1}, it is capped at this value as an upper limit. This threshold was found to be an upper bound storm speed (above the 99.9th percentile) within both the severe and tornado PHI datasets.

Domain Selection and Filtering Methods

This demonstration utilizes a dynamic, data-driven domain to determine where PHI plumes are generated. The original design relied solely on a floater domain, repositioned daily to align with the highest categorical outlook in the SPC Day 1 Convective Outlook. This approach ensured that PHI products were focused on the most likely convective areas while mitigating computational constraints. However, feedback from participating offices highlighted the need for consistent and reliable data feeds within their regions, particularly when convective activity occurred outside the floater domain. To address this, default domains were established and tailored to partner regions. This was done first for Southern Region to cover participating offices and later expanded to Eastern and Central Regions (Fig. 2) with plans to extend to Western Region as offices join the demonstration. Within these default domains, severe PHI plumes are continuously generated to support operations. The domains for tornado PHI plumes remain repositioned each day (at approximately 0610 and 1640 UTC) according to the SPC outlook. Tornado PHI plumes are limited to detections from 10 WSR-88D radars within and surrounding the maximum risk area, consistent with processing constraints. When necessary, a floater domain is appended to ensure that any portion of the tornado domain lying outside the default severe domain is included.
Automated PHI plumes are generated within the active domain every even minute. This yields a consistent update rate for a dataset that is derived from various underlying machine learning algorithms, each with their own update frequencies. To generate PHI, the first step is to search through all of the identified hazard objects to determine which objects are valid for each even minute mark of interest.

A variable probability threshold is used for preparing PSv3 objects for PHI, which in turn affects automated severe PHI plume production. This threshold depends on server demands and is applied to ensure critical data are consistently being distributed, even during extreme cases of wide coverage or other machine limitations. The default PSv3 probability used to process severe objects is 5%. As demand on the computing system increases, this value can be increased to prevent long delays in processing of the real-time data. In very extreme cases, this may increase to above 30%, but thresholds of this magnitude are quite rare.

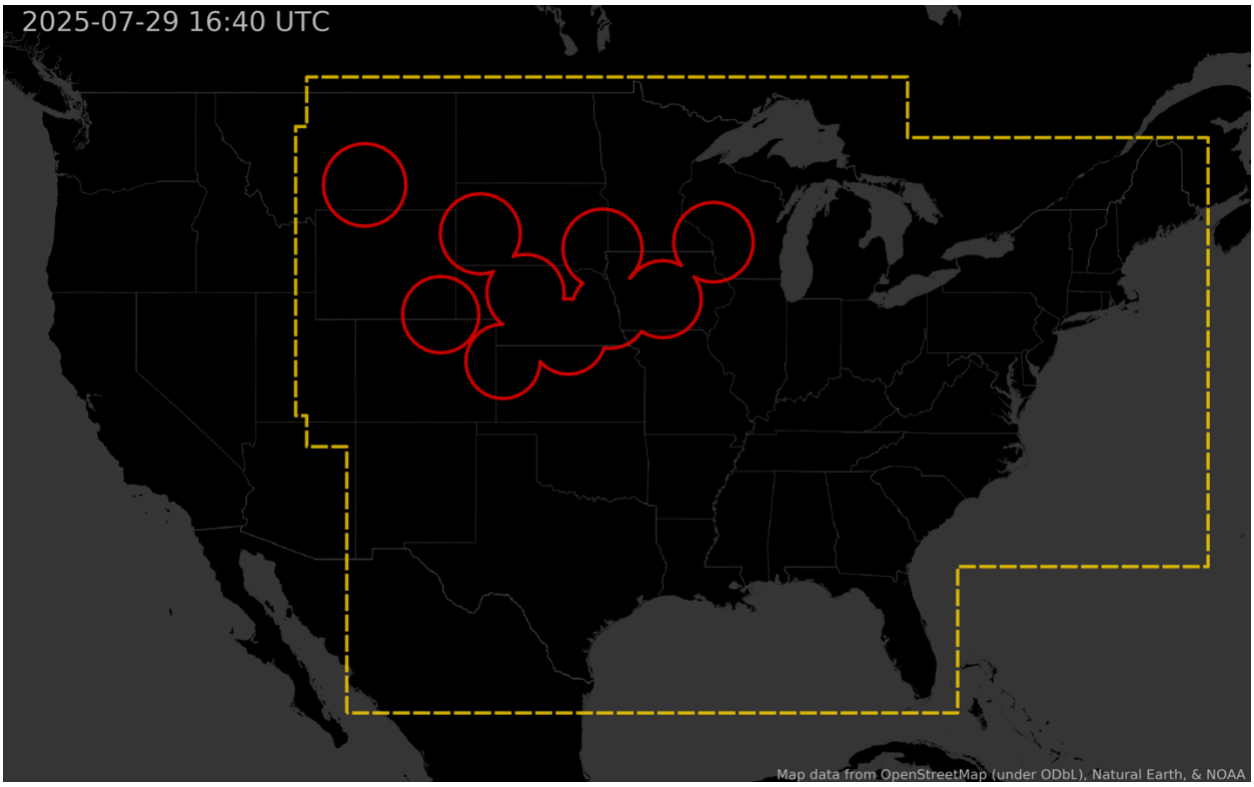

Figure 2: *Example of PHI domains for both severe (yellow) and tornado (red, limited to 10 individual radars) on 29 July 2025.*

For tornado PHI, the underlying machine learning algorithm is single-radar based. Each radar has varying refresh cycles depending on the selected Volume Coverage Pattern (VCP), which leads to different update times for TORP objects. To mitigate discontinuities caused by these asynchronous updates, all TORP objects from the preceding eight minutes are selected and grouped by identification number. From this selection, only initial TORP detections and the latest detections from each group are retained, ensuring greater temporal consistency in plume generation. These objects are then processed through a series of QC filters designed to reduce false alarms. Some of these filters are described in more detail by Sandmæl et al. (2023), while others have been developed during the course of this demonstration in response to forecaster feedback to address common artifacts such as detections near wind farms. The filters are briefly described below in order of their application.

First, a range filter is applied to remove tornado objects with probability values of 60% or below at distances farther than 160 km from the radar. Second, a near-radar filter is applied that aims to remove tornado objects from possible ground clutter detections within 30 km of the radar. To further reduce possible ground clutter or wind farm detections, an additional filter was recently incorporated, where objects are removed if TORP flagged them as matching with known wind-farm locations and historical ground clutter locations. This is followed by a more aggressive

reflectivity filter than the initial 20-dBZ mask used when TORP makes detections; this filter includes the requirement that the reflectivity value of the object center is at least 30 dBZ for several of its first few detections while a tracked object has a short duration. Next, a base 30% probability is required for the QC feed of tornado PHI, based on verification of TORP algorithm performance in order to reduce false alarms due to additional radar artifacts. However, if a tornado object has reached this threshold in the past, it will continue to the next filter to help with continuity so that storm trends can be assessed. Finally, an overlap filter is applied to account for objects identified from different radars for the same circulation. In cases where there is 20% overlap or more between the 7.5 km radius objects, the object with the higher probability and longer duration is kept.

Creation of Automated PHI Plumes

PHI plumes are generated every even minute by using a combination of the underlying machine learning algorithms' probability prediction, QC hazard position and shape, and QC storm motion estimates. For a given time step, each object found from the hazard object selection process (outlined in the Domain Selection and Filter Method section) is spatially interpolated to the even minute mark and then interpolated into the future at one-minute intervals using its respective properties for storm motion until the duration of the object is met (60 min for severe; 30 min for tornado). A direction and speed uncertainty are applied to the object shape every minute, increasing the size of the object at each time step (Fig. 3a). The probability at each minute is extracted from the hazard trend line. For severe, this hazard trend line is created using a

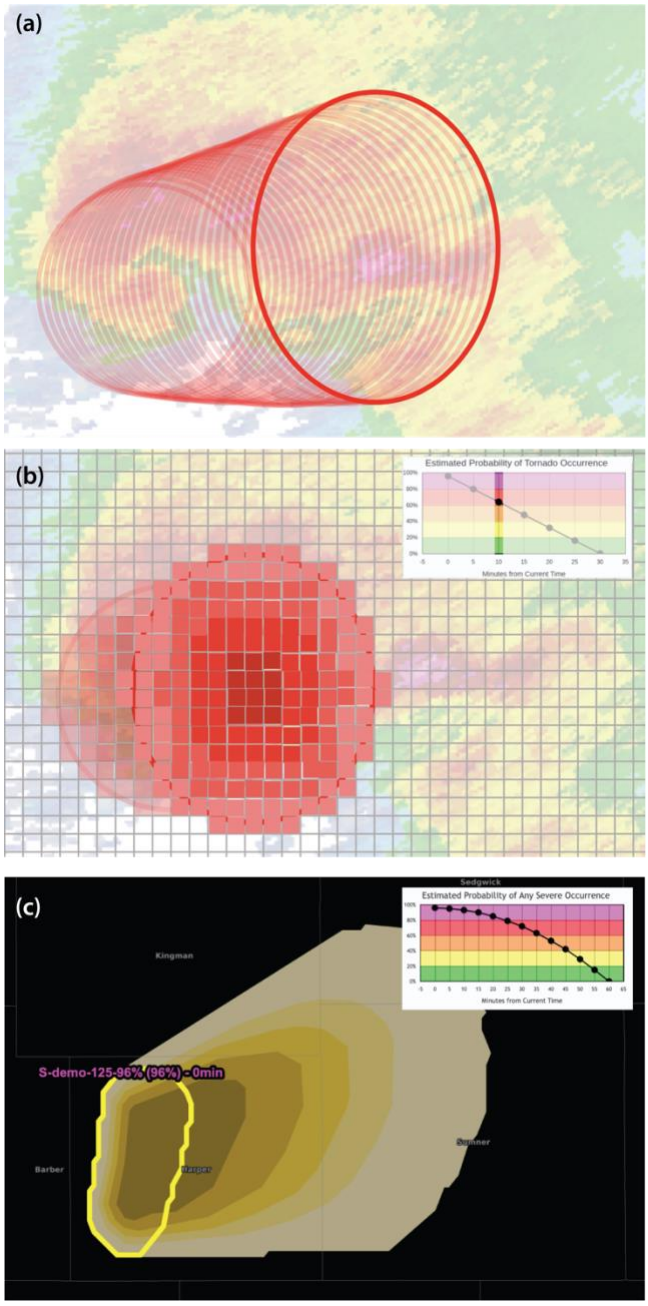

Figure 3: *Steps illustrating PHI plume creation. (a) Example of object projection using speed and direction uncertainty. (b) Example of interpolation of TORP probabilities for tornado PHI at a single time step; Gaussian interpolation is used spatially and linear interpolation is used temporally (inset). (c) Example of a completed severe PHI plume using trend line (inset). Initial probability is from PSv3 then a Gaussian interpolation is used in both time and space from center to edge of identified hazard object.*

Gaussian filter to go from the PSv3 probability at the current time to 0% at 60 min (Fig. 3c, inset). For tornado PHI, the probability decreases linearly from the TORP probability to 0% at 30 min (Fig 3b, inset).

To create spatial grids of probabilities for the object, the initial peak probability is set at the hazard object centroid at the first even minute mark. The spatial probabilities are calculated and gridded using a Gaussian filter to decrease the probability to 0% to the outer edge of the object. This process is repeated for the one-minute interpolated objects where the extracted probabilities are set to the hazard object centroid (Fig. 3b). These initial and one-minute grids are merged into a single grid where the maximum probability value is kept in the case of overlapping values. The resultant grid is then contoured, creating the PHI plume viewed by forecasters (Fig. 3c).

PHI within GibsonRidge Software

In addition to the Advanced Weather Interactive Processing System (AWIPS), NWS forecasters commonly use GibsonRidge 2 (GR2) Analyst software to interrogate WSR-88D radar data during warning operations (e.g., Boettcher et al. 2022). The PHI plumes are converted to placefiles (text files formatted specifically for GR2) that update every two minutes once loaded into the software. Several placefile feeds were created in order to provide a variety of visualization options. Forecasters can display PHI in dynamic loops or static snapshots, with options to customize plumes with or without shading as well as with merged or unmerged probability contours. A depiction of visualization options in GR2 using shaded and contoured tornado PHI and merged contours for severe PHI from a tornadic event over central Tennessee on 9 Dec 2023 is shown in Fig. 4.

NWS Training and Feedback

NWS forecasters participating in the demonstration were asked to provide feedback on the use and applicability of the PHI plumes within operations. Initial participants included only Science Operations Officers (SOOs) at limited NWS forecast offices in the NWS Southern Region; this was later expanded to include forecasters at offices throughout Southern, Eastern and Central Regions. All participants were given a ‘quick guide’ as reference to the key background information on PHI creation, quality control processes, and limitations of the data as well as details on how to load the data into the GR2 software. Additionally, multiple virtual seminars were given to forecasters throughout the demonstration period by the development team. These seminars provided detailed background on the creation of the probabilities and expanded on the information available within the quick guide. One of these seminars was provided through the Warning Decision Training Division Research to Operations and Operations to Research Webinar Series

(ROOTs); this webinar was recorded for future access by new participants (NWS Learning Office 2025).

In order to facilitate a wide array of feedback, feedback could be provided within a shared NWS-NSSL *Slack* channel, directly via email, or within online surveys and semi-regular meetings between the development team and local offices. This multi-platform feedback was later deidentified and tabulated for content analysis by the development team. Emergent categorization of the data was completed to determine the coding themes (e.g., Krippendorff 2023) of the feedback we received from forecasters (i.e., 'performance', 'feature requests', 'external decisions and communication', and 'training/understanding'). These themes will be discussed in the Results below.

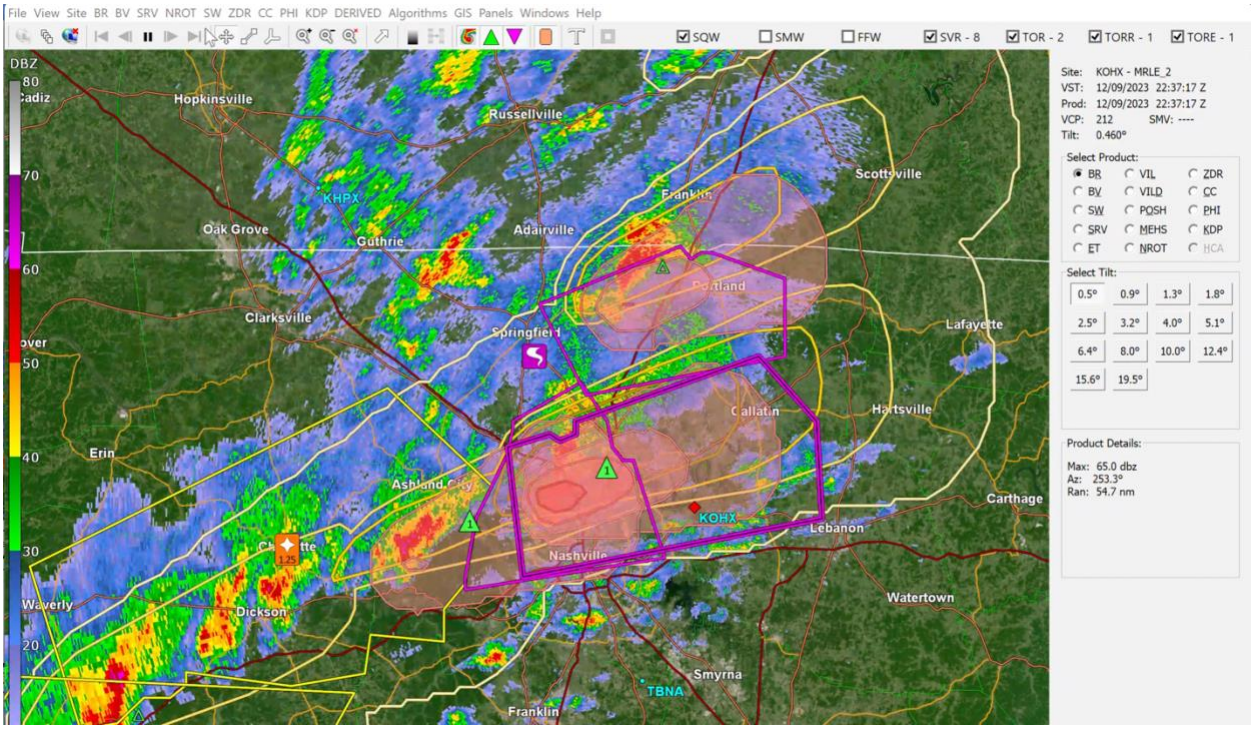

Figure 4: *A screen capture of PHI plumes in GR2 from 9 December 2023 at 2237 UTC in central Tennessee. Severe and tornado PHI are shown as yellow contours and red shaded contours, respectively, with contours every 20% probability. Overlaid are NWS severe thunderstorm (yellow polygons) and tornado (magenta polygons) warnings on KOHX 0.5° reflectivity.*

3. Results

Content Analysis and Findings

Roughly 60% of forecaster feedback addressed the performance of PHI plumes during specific weather events. Early in the project, performance issues were the primary theme and consisted of greater than 70% of the feedback, but the proportion of comments on this topic has declined as iterative development addressed many of the initial challenges. The feedback provides a clear picture of how forecasters perceived both the strengths and the limitations of severe and tornado PHI. These themes differed somewhat between severe and tornado PHI, reflecting differences in storm characteristics and algorithm behavior.

For **severe PHI**, positive feedback most often highlighted:

- Probability (~40% of positive severe comments) – Forecasters noted realistic and meaningful probability trends that provided confidence in decision making.

  "Another good performance…producing golf ball- to baseball-size hail…PHI fields were focused on the storms that verified."

- Coverage and identification (~30%) – Many comments pointed to good representation of threat areas and effective identification of objects.

  "PHI was good reinforcement of the highest risk along a broken line of storms"

- Motion (~20%) – When motion was handled well, forecasters valued the reliability of storm tracks and used them to guide warning polygons.

  "We relied on the PHI plume to anticipate broader storm motion and drew a polygon that worked out really well."

- Overall strong performance (~5%) – Forecasters highlighted examples where the product as a whole performed well.

  "PHI Severe did a good job with storms in the Texas Hill Country yesterday evening…hail reports were well represented within the contours."

Negative feedback for **severe PHI** related to:

- Motion (~75% of negative severe comments) – The vast majority of complaints cited unreliable or unrealistic storm motion, especially early in the storm lifetime and with weaker or complex storms.

  "Watching storms in KS this morning saw this behavior with plumes directed 180 degrees from the mean storm motion. It seemed to reoccur with new cells being identified on the tail end of the convection then losing them and reidentifying cells again."

- Coverage (~10%) – Some forecasters described difficulty maintaining full coverage of linear systems or merging and splitting storms.

  "The convective mode morphed from supercellular to multicell to linear, and during this time, we [had] warnings on two cells… The PHI severe plume "bullseye" was directly between the two warnings. The northern cell was more of a wind threat; the southern was a hail threat."

- Probability (~10%) – A few comments suggested possible low biases in bowing segments or mesoscale convective systems in relation to severe wind.

  "Severe PHI [was] struggling with the MCS…it looks like it's got a bit of a low bias."

- Other (~5%) – criticisms occasionally mentioned consistency and stability issues.

Forecasters cited several strengths of the **tornado PHI**:

- Probability (~40% of positive tornado comments) – Many described probability trends as meaningful and helpful in assessing evolving threats.

  "There have been a couple really good jumps in probabilities across Glades, Okeechobee, Martin, and St. Lucie counties corresponding with confirmed tornadoes."

- Identification (~30%) – Forecasters noted accurate placement of objects relative to radar rotation.

  "Tornado PHI seems to do a great job with the initial spin-up and intensification, and then identifies an apparent shift to a new mesocyclone on the western flank of the storm."

- Continuity and Coverage (~20%) – Some mentioned smooth temporal behavior *when* detections were maintained and sufficient storm coverage in individual cases.

  "PHI Tor continued to perform well…the plumes seemed to have good continuity and contours did not disappear and reappear as in earlier events."

- Other (~10%) – Some comments noted additional strengths such as identifying the storm motion or handling complex or unique events correctly.

  "It held on to a rather persistent and stubborn bookend vortex that aided our warning decisions."

Limitations of the **tornado PHI** were more varied and consisted of:

- Continuity and Coverage (~35% of negative tornado comments) – Many forecasters described dropped or shifting detections and gaps in linear or complex systems that undermined trust.

  "Tornado PHI struggling a bit this morning with our line of storms. A few times it had contours of 40 and 20 percent but disappeared, then re-appeared roughly in the same area."

- Storm motion (~30%) – Unrealistic or inconsistent motion estimates were a common frustration.

  "The plumes' orientation were consistently more north than the actual circulation motion and polygons."

- Probability (~20%) – Concerns often centered on spurious detections, fluctuating values, or an apparent high bias.

  "I prefer it being a little hot…but not too hot where you discard it as being overdone."

- Identification (~15%) – Some pointed to false detections or missed signals.

  "False positives due to radar sampling issues in the terrain in NW Arkansas…but they were easy to resolve."

In response to this feedback for both severe and tornado PHI, developers made coordinated improvements to both severe and tornado PHI storm motion attributes. Storm motion estimates initially relied on RAP guidance, which often misplaced storms across boundaries due to its coarser grid spacing (13 km). These were replaced with HRRR data with a finer grid spacing of 3 km. The switch to using HRRR data provided improvement over the RAP, but issues were still noted for different storm modes and left splits. To help address these issues, a SCP threshold was recently introduced to help distinguish between environments as described in the Data and Methods section above. Additionally, a grid of current severe storm motions derived from current longer-lived storms, which tend to be more stable and well defined, was created. This allowed newly identified objects, particularly tornado objects (shorter-lived) and those formed through splits or mergers, to inherit more realistic storm motion.

Development has also addressed continuity and false detection challenges for tornado PHI. False detections were commonly attributed to side-lobe contamination (e.g., Boettcher and Bentley 2022) or ground clutter from features such as wind farms. Side-lobe contamination is a known limitation of the TORP algorithm (Sandmæl et al. 2023), producing detections with artificially high probabilities and poor spatial accuracy and these remain difficult to fully address. Early versions of the QC filters removed all TORP detections beyond 160 km from radar, but feedback prompted an allowance for plume creation outside this range if other QC filters are not flagging the detections and object probability >60%. Similarly, all detections below 30% probability were initially discarded, but if an object had a history of exceeding that threshold, it was retained at all time steps to preserve continuity. Additional refinements targeted clutter from wind farms and persistent noise; a grid of long-term averages of reflectivity was created to filter out these spurious detections from the GR2 feed.

Although limitations were a major focus of feedback, forecasters also highlighted strong performance cases, especially after QC improvements, illustrating the value of iterative development and refinement. Two events specifically mentioned by forecasters are examined in greater detail in the Case Study Performance and Analysis section below.

The next most common area of feedback concerned how forecasters applied PHI guidance in their decision making and in communications with partners for IDSS. It was found that PHI information was used both to guide warning operations and to support downstream messaging. Several forecasters (nearly 10%) described instances where PHI directly supported challenging warning decisions. For example, one noted that PHI "helped immensely with a cell merger…that made ascertaining storm motion and polygon drawing difficult, so we relied on the PHI plume to anticipate storm motion and drew a polygon that worked out really well." Another forecaster reported that PHI probabilities from TORP influenced escalation from a severe thunderstorm warning to a tornado warning: "we were evaluating whether to continue a SVR warning…[TORP]

started showing a signal…This caught our eye and we evaluated the circulation more…and saw the tightening and strengthening meso. This prompted us to issue a TOR."

Feedback also emphasized PHI's role in partner communications. PHI plumes were used to provide more confident "downstream" information to emergency managers and other partners who were not yet within an active warning. One forecaster explained: "We're pretty active in our NWSChat room…and partners outside of the warning are always asking if the warning will be 'extended' downstream into their county/town. The PHI plumes allowed us to give them a data-driven answer instead of relying 90% on forecaster intuition."

Finally, some forecasters described using PHI plumes to enhance messaging graphics, even in cases where a warning was not yet issued. As one participant noted: "Using the PHI plumes on a non-severe storm to help message potential risk. The values seemed reasonable, so we used them to create a quick meso-update graphic." Together, these examples demonstrate that beyond algorithm performance, PHI influenced both operational decision making and external communications, providing forecasters with additional confidence and data-driven justification in both warning and IDSS contexts.

Additional discussion included questions on how the domains were chosen, potential biases in the data, probability calculations, available training material, and the need for archive playback capability. Of these, archive playback was one of the most frequent requests, raised both during training seminars and within follow-up emails to the team. Forecasters emphasized that limited time during active warning operations prevented a thorough exploration of PHI, and that greater exposure to past cases would help build calibration and trust. In response, an archive system is currently under development with an initial rollout planned for testing. This system is described in more detail in the Recommendations and Ongoing Development section below.

Performance Metrics of Automated Severe and Tornado PHI

Evaluating performance metrics offer insights into the quality of PHI plumes, helping to determine if underlying algorithms are well-calibrated or require retraining, and informs best practices for operational forecasters. Performance is evaluated using methods similar to those described in Cintineo et al. (2024) and Sandmæl et al. (2023). The evaluation is performed for all dates between 1 June 2023 to present. Reports were filtered based on severe criteria. For severe PHI, this consisted of hail (≥1"), thunderstorm wind (≥58 mph), and tornado reports. For tornado PHI, only tornado reports were used. Official metrics are calculated using NCEI Storm Data (NOAA 1950a). However, due to a publication latency of up to 90 days for certified reports in the Storm Data

database, any verification performed within the 90-day period (i.e., after 1 May 2025) utilizes SPC Storm Reports (NOAA 1950b) and is considered preliminary. Storm reports, despite known issues with reporting (e.g., Trapp et al. 2006, Allen and Tippett 2015, and Potvin et al. 2016), are used but limitations should be acknowledged. Future research could investigate using MRMS as a proxy for reports as was done in Wendt and Jirak (2021).

Periods during which the system experienced slowdowns, errors, or downtimes were excluded to ensure a more accurate assessment of performance under optimal conditions. Slowdowns were recognized when processing time or probability thresholds were exceeded (e.g., probability threshold not equal to 5% for severe PHI processing, or processing time exceeding 60 seconds for tornado PHI). Errors were indicated by "error," "exception," or "fail" in the processing logs. Downtimes were identified when subsequent processing times surpassed 10 min. The evaluation includes only areas defined by the active domain(s) for that period (see example domain on 25 July 2025 in Fig. 2).

Verification was performed using a time window technique which converts continuous forecasts to binary forecasts (refer to Cintineo et al. 2024 for a more in-depth discussion). This method closely aligns with NWS warning and verification practices but is adapted for a probabilistic framework which assesses each threshold bin individually. Each storm report was associated with the nearest PHI plume within a given spatial buffer. For severe PHI, a 5 km buffer was applied for hail, thunderstorm wind, and tornado reports occurring within a 60-min window. Note this window differs from Cintineo et al. (2022; ProbSevere version 2) and Cintineo et al. (2024; PSv3) both of which used 45-min for ProbHail and ProbWind and 30-min for ProbTor. For tornado PHI, a 10+ km buffer was applied to tornado reports occurring within a 30-min window except that each report was interpolated from start to end every minute and 25 m was added to the buffer for each minute after the report's start. For example, if a tornado report occurred between 1800 and 1810 UTC, the buffer would increase from 10 km to 10.125 km from start to end. These supplemental data points more accurately represent the real-time performance of the underlying model (Sandmæl et al. 2023). However, our dataset consisted only of interpolated report points whereas Sandmæl et al. (2023) integrated multiple datasets, including storm-report objects and manually-identified objects.

Fig. 5 summarizes the overall performance for the binary classification of the plumes for each probability bin. For severe PHI, the highest Critical Success Index (CSI) is observed for probabilities between 30 and 60%. There is an under forecast for probabilities below 20% and an increasing over forecast at higher probabilities. In contrast, tornado PHI shows the highest CSI for probabilities 50% and above, with an over forecast across all probability bins. Over forecasting is likely inflated in comparison to Sandmæl et al. (2023) due to two factors: interpolating TORP to PHI resolution, which can exacerbate issues with noise detections, and differences in verification

datasets. There is still skill added despite over forecasting for both severe and tornado PHI, especially when considering no guidance versus with guidance. In fact, forecaster feedback overwhelmingly indicates a preference for over forecasting rather than under forecasting, as it provides complete situational awareness and allows forecasters to determine which storms to monitor.

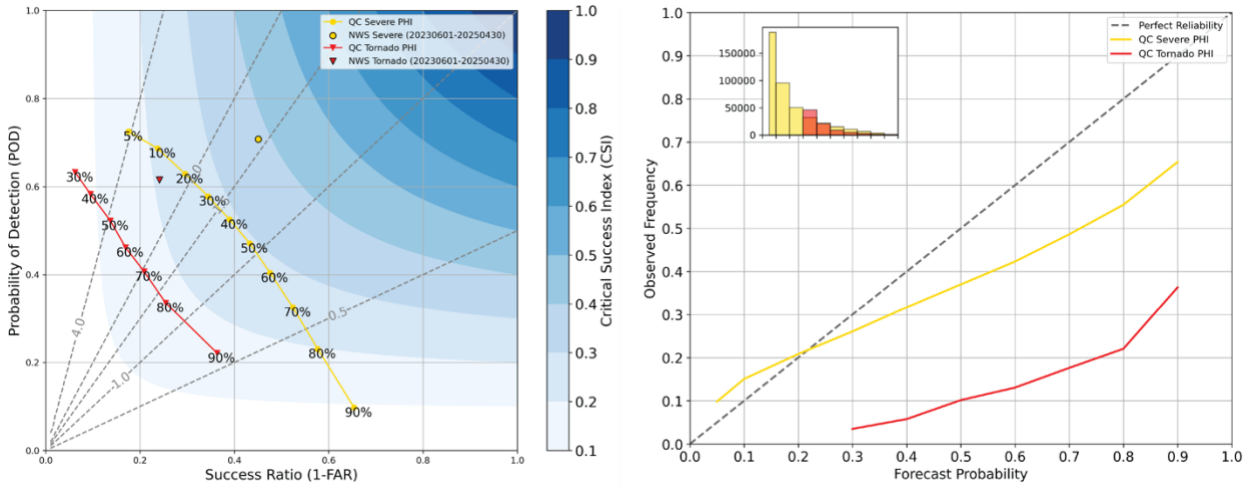

Figure 5: *(a) Performance diagram showing success ratio (x-axis) and probability of detection (y-axis) of severe (yellow) and tornado (red) PHI probabilities from 1 June 2023 through 30 April 2025. NWS severe thunderstorm and tornado warning performance over the same period of time is shown with an outlined yellow circle and red triangle, respectively. (b) Reliability diagram showing the PHI forecast probability (x-axis) relative to the observed frequency (y-axis) over the same time period as panel (a). This includes 2,125,033 (n) 2-min severe PHI predictions for 163,092 different identified storms, and 282,436 (n) 2-min tornado PHI predictions for 48,960 detected circulations.*

Case Study Performance and Analysis

8 July 2024: Hurricane Beryl Landfall

Hurricane Beryl made landfall as a Category 1 Hurricane on 8 July at 0850 UTC near Matagorda, Texas with the eye continuing to move northeastward across Texas near the west Houston metro area. Following landfall, Beryl was downgraded to a tropical depression by 0000 UTC on 9 July and continued to move northeast across Texas and into Louisiana, becoming an extratropical low by 1200 UTC on 9 July (Beven et al. 2025). At least 65 tornadoes were associated with Beryl, the majority occurring as Beryl moved across Texas, Louisiana, and Arkansas. Additional tornadoes were reported further north into Indiana and New York and Canada. This evaluation will focus on the period beginning just before landfall at 0600 UTC on 8 July through 0700 UTC on 9 July 2025 across Texas, Louisiana, and Arkansas when most of the tornadoes occurred.

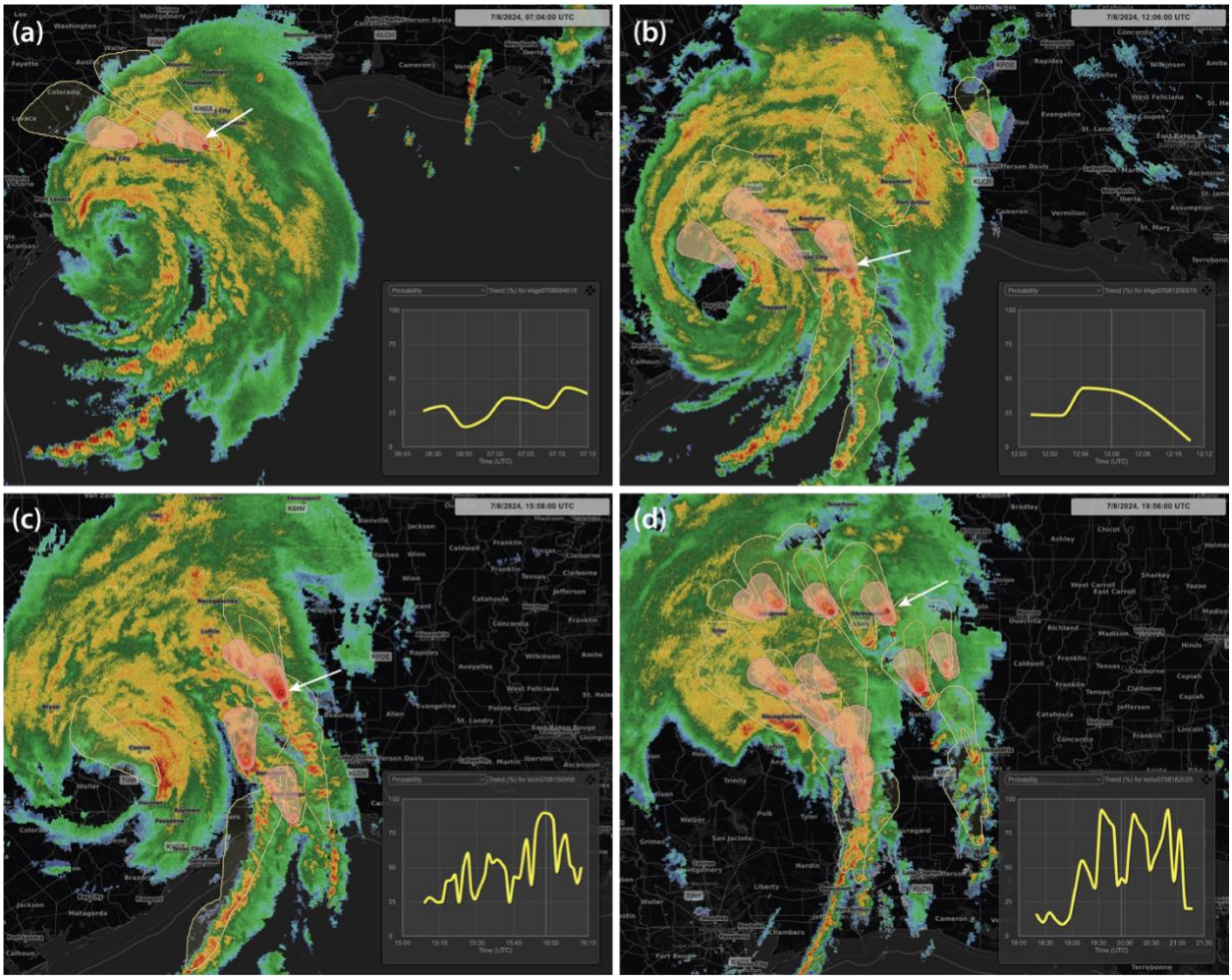

Figure 6: *Evolution of Hurricane Beryl following landfall in Texas on 8 July 2024. PHI plumes (Severe - yellow contours; Tornado - red shaded contours) and reflectivity mosaic are shown at (a) 0704, (b) 1206, (c) 1558, and (d) 1956 UTC. Inset time series is for tornado probability associated with plume highlighted by white arrow, note the longer storm lifetimes in (c) and (d).*

For the six hours following landfall, lower probability severe PHI plumes covered much of the precipitation area east and northeast of the inner core. Intermittent tornado PHI plumes also appeared in response to episodes of rotation in both the rainbands as well as the inner-core region. These tornado PHI plumes were typically short-lived (<15 min) and contained lower probabilities (<45%) during this period (Fig. 6). After 1500 UTC on 8 July 2025, the context of these plumes changed within the rainbands between Houston, TX and Lake Charles, LA. These plumes were typically associated with long-track, low-topped supercell storms (often lasting more than 1 hr) with higher probabilities than seen earlier in the event. An EF2 tornado reported at 1557 UTC near Jasper, LA kicked off a period of multiple tornado reports (Fig. 6c). From 1800 to 2200 UTC,

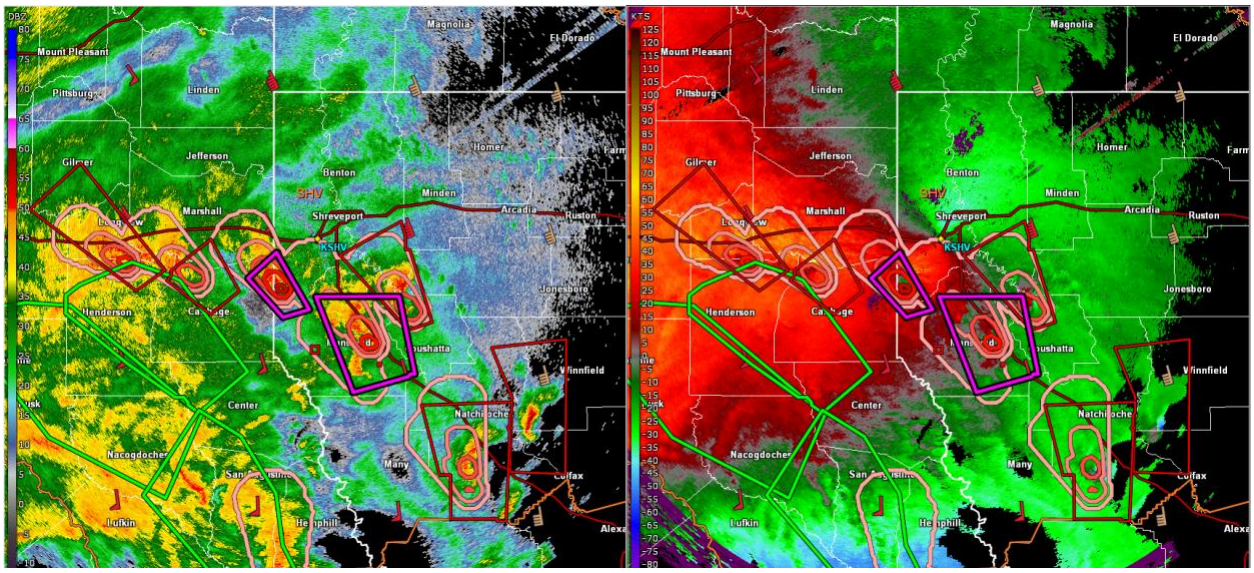

Figure 7: *Screenshot from NWS forecaster of tornado PHI plumes in GR2 (pink and pale red contoured plume) with NWS tornado warnings (red and magenta polygons) and flash flood warnings (green polygons) over KSHV 0.3° reflectivity (left) and velocity (right) at 1916 UTC 8 July 2024.*

there were more than 7-10 different supercell storms in the right front quadrant of Beryl, many of them tornadic during this period (Fig. 6d and 7).

Both the severe and tornado PHI plumes maintained accurate storm motions as the supercells moved north then northwest around the hurricane center. NWS warning polygons consistently overlapped well with the tornado PHI plumes, with similar coverage areas and estimated storm motions (Fig. 7). Due to the nature of the event, severe PHI plumes often covered much of the eastern side of the hurricane, though the overall severe probabilities remained low with the vast majority of plumes having no higher than 10% likelihood. This results in much lower performance metrics for this case than overall severe performance during the evaluation. The highest CSI (0.18-0.19) were for 20-30% probabilities for NWS Southern Region from 0600 UTC on 8 July to 0659 UTC on 9 July (Fig. 8a). Probabilities above this had both much lower Probability of Detection (POD) while also having decreased Success Ratios (SRs; 1-FAR). These skill scores may have been at least partially a result of underreporting of severe reports due to competing hazards and overwhelmed systems during a hurricane landfall. This can lead to under-reporting across the board, including severe storm impacts (e.g., Trapp et al. 2006). However, it may be that the severe PHI is simply not tuned well for identifying hazards adequately for this type of event. Tornado PHI plumes demonstrated higher skill scores for the event with the highest CSI for the tornado PHI between 0.33-0.35 for 60-70% likelihood. Higher probabilities had lower POD, but unlike the severe PHI plumes, they also had an increased SR. Both severe and tornado probabilities suffered from an over forecasting bias at most thresholds with consistent higher forecast probabilities than observed occurrences (Fig. 8b).

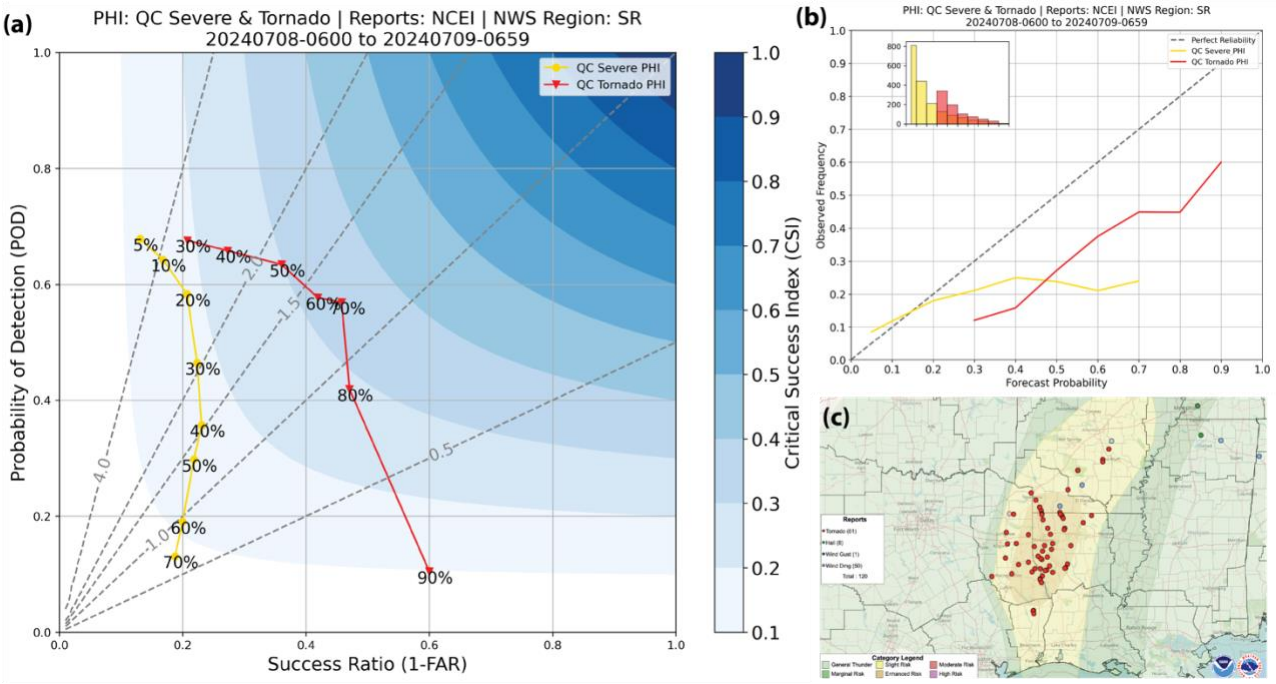

Figure 8: *(a) Performance diagram for severe (yellow) and tornado (red) PHI by probability threshold for 0600 UTC 8 July through 0659 UTC 9 July 2024. (b) Reliability diagram of severe (yellow) and tornado (red) forecast PHI probabilities (x-axis) relative to observed frequency (y axis) with inset bar chart showing the frequency count in each bin for the same period as shown in (a). (c) SPC storm reports for 1200 UTC 8 July through 0700 UTC 9 July overlaid on Day 1 SPC Severe Weather Outlook (shaded).*

27-28 June 2025: Tornadic Supercells in North Dakota

This event was characterized by the development of numerous supercell storms and larger storm clusters along a weak surface front extending across western and central North Dakota as an upper-level shortwave trough moved across the area. A tornado watch was issued for the region at 2220 UTC with the first storms initiating within a broad area of instability shortly before 2330 UTC in central and northern North Dakota. The 0000 UTC sounding from Bismarck, ND, indicated CAPE exceeding 2000 J kg^{-1}, steep low-level lapse rates, and effective shear greater than 40 kts. Additional tornadic supercells subsequently formed and progressed across the region over the following five hours, supported by a strengthening low-level jet. By 0330 UTC, mesoscale analysis from SPC identified an area over central North Dakota where the effective-layer Significant Tornado Parameter (STP) exceeded a value of 5 (values > 3 are typically associated with tornadoes of EF3 intensity or greater; Thompson et al. 2012). After the event, it was reported that the PHI plumes were particularly helpful for warning decision-making and storm management. However,

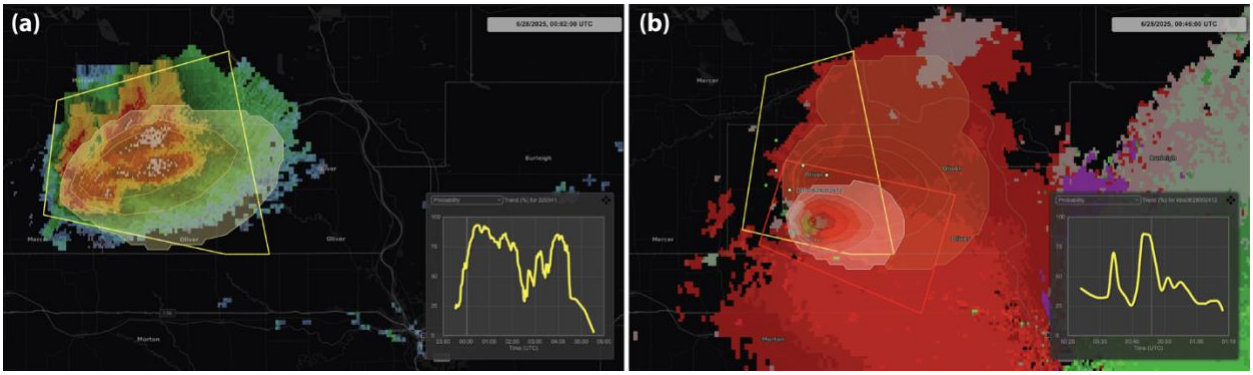

Figure 9: *(a) Severe PHI plume (yellow shading/contour), NWS severe thunderstorm warning (yellow polygon) and reflectivity from KBIS at 0002 UTC 28 June 2025. (b) Tornado PHI plume (red shading/contours), NWS severe thunderstorm (yellow polygon) and tornado (red polygon) warnings and velocity from KBIS at 0046 UTC 28 June 2025. Inset trend graphs on each plot show the PHI probability trends for (a) severe and (b) tornado; vertical line on each graph represents the current time of the plot.*

there were instances later in the event in which side-lobe contamination adversely affected the algorithm's performance.

The first tornadic supercell of the event initiated at 2330 UTC (defined using >30 dBZ reflectivity) and intensified rapidly over Mercer and Oliver counties. A severe PHI plume first appeared at 2336 UTC with a 25% probability, increasing steadily to 92% by 0022 UTC. The initial severe thunderstorm warning was issued at 0000 UTC, when the probability had reached 65% (Fig. 9, left). Tornado PHI guidance followed at 0022 UTC initially from the KMBX radar with a 50% probability; another automated plume from KBIS followed shortly after at 0024 UTC with an initial probability of 40%. Tornado probabilities remained between 50–70% over the next 20 min. The plume using KBIS data had more fluctuation, likely due to more inconsistency of the rotation lower in the storm where the KBIS radar was scanning. The first tornado warning was issued at 0038 UTC, just after a 70% probability peak within the KBIS tornado PHI plume. The NWS tornado warning polygon and tornado PHI plume were closely aligned with the first spotter-reported tornado at 0041 UTC (Fig. 9, right). Tornado PHI plumes were maintained with the storm with probabilities between 25–40% through 0210 UTC, while the NWS maintained tornado warnings until 0145 UTC. Another tornado was reported between 0130 and 0138 UTC.

A second tornadic supercell initiated in Oliver County around 0130 UTC, following a similar track behind the first. An initial severe PHI plume with probabilities of 20% appeared at 0142 UTC and had a southeastward storm motion. A severe thunderstorm warning was issued slightly later at 0156 UTC, with estimated eastward movement that better matched the storm at this time (Fig. 10a). By 0224 UTC, the storm developed a mesocyclone, and its motion shifted more southeasterly, better aligning with the automated plume. A tornado PHI plume emerged at 0227

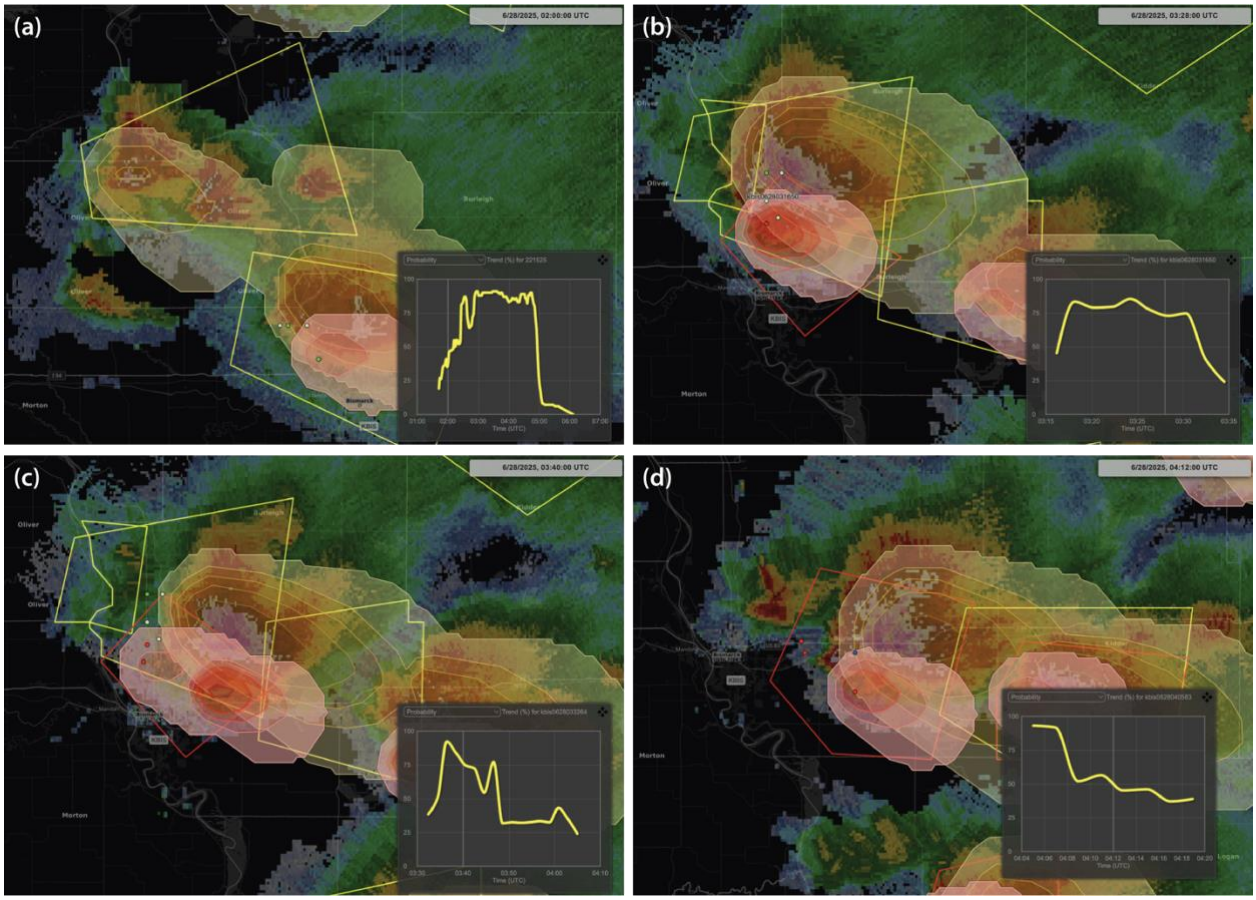

Figure 10: *Reflectivity is shown in all plots. Plot (a) has the severe probability inset for the more than four hours the storm was tracked (0142-0600 UTC). Plots (b-d) show the tornado probability trends for the three different tracked TORP circulations*.

UTC with a peak probability of 41% and remained with the storm through 0301 UTC. After a brief lull, a new tornado PHI plume formed at 0316 UTC with 45% probability, rapidly increasing to 83% by 0318 UTC. A tornado warning was issued at 0325 UTC, coinciding with the first tornado report from this storm (Fig. 10b). Multiple tornadoes were reported through 0334 UTC, after which both radar velocity and tornado probability quickly dropped below 25% as a new updraft formed farther east and cut-off inflow to this region of the storm. A new tornado PHI plume based on KBIS appeared at 0332 UTC at this location and quickly increased from 38% to 92% likelihood over four min (Fig. 10c). A new tornado warning was issued once again coinciding with a tornado

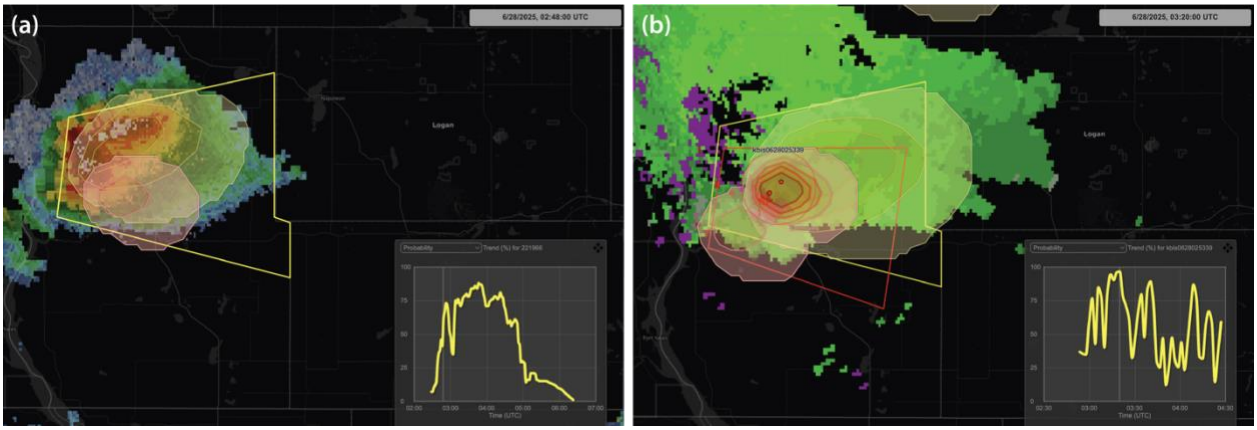

Figure 11: *As in Fig. 9 but for new convection to the south in Emmons County at (a) 0248 and (b) 0320 UTC. The red circles provide the locations of SPC tornado reports valid at the current time.*

report at 0344 UTC. The storm continued to cycle with additional plumes developing at 0405 UTC with 90% probability and another tornado report at 0412 UTC (Fig. 10d). The final tornado PHI plume disappeared at 0430 UTC as the storm merged with the dissipating storm just ahead of it. Severe probabilities rapidly dropped from 85% to <10% between 0448 and 0506 UTC as the storm merged with larger clusters to the south.

Another isolated supercell storm developed in Emmons County at 0224 UTC, situated south of the previous supercells and north of a bowing storm cluster. An automated severe PHI plume began at 0228 UTC with a low 7% probability but intensified quickly. The first severe thunderstorm warning followed at 0245 UTC, when the severe probability had risen to 62% (Fig. 11). Tornado PHI guidance based on KBIS began at 0242 UTC with a 30% probability. A second plume developed at 0253 UTC due to a new rotational feature farther back in the storm. Tornado probabilities quickly increased with this area, reaching 76% by 0301 UTC and 85% by 0305 UTC. A tornado warning was issued at 0304 UTC, remaining in effect until 0330 UTC. Multiple tornadoes were reported beginning at 0314 UTC, well-aligned with both the peak plume probabilities and warning polygon (Fig. 10). Later it was reported that the tornado probabilities suffered from side-lobe contamination.

While the examples above focused on the tornadic supercell storms, there were a number of merging and splitting storms as convection grew upscale overnight with a bowing cluster moving southeast across central and southern South Dakota. Based on overall performance of both PSv3 and TORP for these storm modes and region as well as the implementation of multiple quality control methods already in place, the general expectation is that performance metrics for automated PHI plumes should be better than the two-year average, but not as good as an event with fully isolated storms.

For severe probabilities of <70%, POD remained somewhat steady from 0.8 to 0.85 while the SR increased steadily from 0.24 to 0.58 leading to the highest CSI of 0.51 at a probability of 70% (Fig. 12a). This CSI is a bit higher than the CSI in this region from PSv3 of 0.35-0.45 (Cintineo et al. 2024) or national CSI for the two-year period of automated PHI plumes. Overall reliability was similar to what we see in the national metrics and PSv3, with a relatively reliable forecast that had a slight under forecast below 30% probability and a slight over forecast that increases at probabilities above 50% (Fig. 12b).

For the tornado PHI plumes, the highest CSI in this event was for probabilities between 80-100% with CSI between 0.4-0.44 (Fig. 12a). However, while there is only a modest difference in CSI between the probability thresholds of 80 and 90%, the trade-off between POD and SR for the same CSI is quite apparent. At 80% probability, the POD is 0.85 while the SR is 0.45, but at the 90% threshold, POD falls to 0.53 while the SR increases to 0.67. Forecasters should keep this in mind when thinking about thresholds to use for both situational awareness and warning decisions. Due to the nature of both the environment and the strength of the rotation, when present, almost every tornado PHI plume during this event had probabilities >30% and every plume associated with a tornado report had probabilities >60% (Fig. 12b).

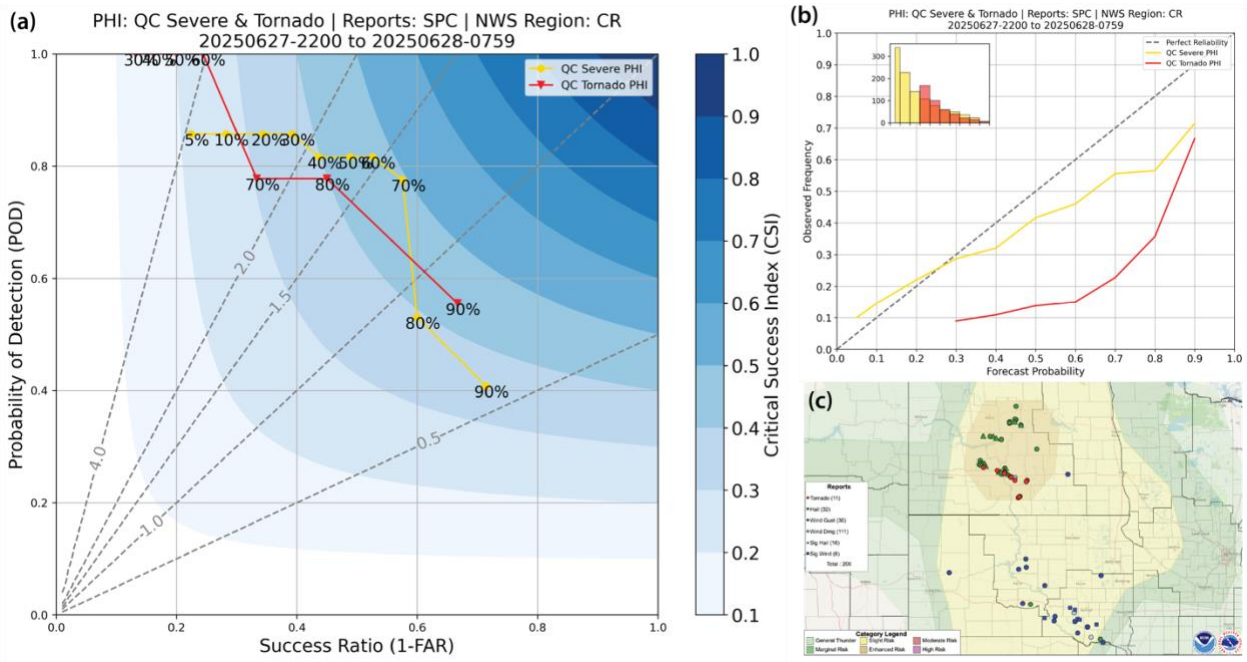

Figure 12: *As in Fig. 8 but for 2200 UTC 27 June 2025 to 0759 UTC 28 June 2025.*

4. Recommendations for Use and Ongoing Development

Best Practices for Using PHI in Operations

Based on feedback and efforts to adequately handle the QC aspects described in Data and Methods, fully automated PHI may be used by forecasters in several complementary ways: 1) for calibration, by providing a statistical "anchor" for storm-to-storm comparisons during an event; 2) for storm triage, by accelerating cognition when multiple areas compete for attention; and 3) for communication, by offering a defensible numerical reference for expressing risk prior to warning issuance. Similar patterns are seen in other high-stakes fields: for example, clinical decision support systems give physicians an initial ranked diagnosis that improves calibration while leaving final judgment to the doctor (e.g., Sutton et al. 2020), and cockpit automation provides guidance cues that help pilots prioritize attention during complex flight situations (e.g., Causse et al. 2025). Though automation has shown improved performance and reduced mental workload in a wide range of fields, over-reliance on automation without knowledge of its limitations can lead to misuse or over-dependency, making it essential for forecasters to remain aware of its strengths and limitations.

Building on previous HWT results, operational forecaster feedback, and performance metrics, forecasters may need to consider different thresholds for communicating hazard potential and making warning decisions. Thresholds may also need to vary by storm mode and environment. The strongest performance is observed for supercell storm environments, such as the 27-29 June 2025 event described in the second case study. This is where PHI can be used for triage as well as understanding storm trends. These are also cases with the highest CSI scores for both severe and tornado hazards, demonstrating the feasibility of use of PHI as both a calibration tool and a decision aid. Even in more complex convective scenarios, however, PHI can still offer useful initial calibration for storm comparison, though forecasters must supplement with deeper storm interrogation when signals are less clear.

Understanding storm-by-storm PHI trends can also provide forecasters with situational awareness and highlight which storms require deeper interrogation. While parallels exist with other domains where automated cues support professionals in synthesizing competing data streams, the primary value here lies in enabling forecasters to focus cognitive resources where they matter most, without relinquishing expertise to automation. Continued evaluation should also consider how to mitigate risks of automation bias, ensuring that forecasters use PHI as a guide rather than a determinant, and that flexibility remains to override automated probabilities when local expertise or evolving storm structures suggest otherwise.

Beyond storm interrogation, PHI has also shown value in communication with external partners. For example, forecast offices have used *NWSChat* to relay PHI-derived trends in storm probability to high-end user groups such as emergency managers. These early messages provided valuable context and a "heads up" about potential warning issuance, allowing partners to anticipate protective action even before official warnings were issued. In this way, PHI supports IDSS by offering a transparent, quantitative reference that helps users understand both the timing and likelihood of escalating threats.

Computing Resources Necessary for PHI in Operations

The iterative and developmental nature of the PHI provided thus far has been well-suited to a computing framework that is relatively light on resources. Most PHI and TORP processing has taken place on a powerful on-premesis server machine with output disseminated via a minimal cloud architecture. Resources have been marginally increased and improved over the course of development, however some compromises in the quantity of output have been necessary to match the resources available. Techniques such as domain limitation or the variable probability threshold outlined in the Domain Selection and Filter Method section have been the primary methods of preventing overload on research-level computing resources.

Computationally, the goals of PHI in operations would be to expand domain coverage, reduce probability cutoff thresholds, and maintain a very robust system availability, in addition to enabling continuous improvements in the quality of PHI output. Optimization of PHI processing code can assist with each of these goals, however additional computing resources will be required to effectively support CONUS-wide operations. Increased server compute power will be required for continuous PHI coverage of CONUS, with higher server specifications or greater numbers of servers translating directly into increased domain coverage and lower probability thresholds. TORP processing currently has an effective limit of 10 radars per server (described in the Domain Selection and Filter Method section). As such, increased domain coverage for TORP output will require an additional processing server for every 10 radars desired. Increased computing resources may also assist with how quickly PHI becomes available following radar scans.

A transition to a cloud-based architecture would also offer some advantages to an operational system. Increases in compute power can be easily scaled up in a cloud environment, load balancing schemes can be utilized to automatically increase or decrease compute power on demand, and the runtime availability of cloud-based systems is very high relative to on-premesis systems. During this demonstration, some PHI and TORP processing has already been run in temporary cloud-based environments.

Archive Case Review and On-Demand Verification

A case-by-case evaluation approach, similar to the case studies previously discussed, can help forecasters identify PHI's strengths and weaknesses across various events, especially those impacting their own County Warning Areas (CWAs). This can ultimately foster forecaster trust and facilitate more in-depth discussions. The Severe Weather Research Map (SWRM) serves as a platform for this type of evaluation and is currently undergoing a phased release with NWS forecasters.

The SWRM is an experimental, interactive web-based mapping tool that allows forecasters, researchers, and developers to readily examine cutting-edge research products (Steeves et al. 2020; NOAA 2021). Research products, such as PHI Recommenders (see the Warning Recommender within Hazard Services section below), and software capabilities can be rapidly incorporated into the web tool, making it an ideal platform to test out novel concepts and ideas related to PHI. Using the SWRM, forecasters (as well as researchers and developers) can review past events in detail, as the SWRM is set up to streamline the review of both real-time and archived PHI data. In order to allow users to efficiently review PHI data, the web tool displays queryable PHI data and provides enhanced visualization features, including data animation, display customization, and hazard object trend analysis charts. Visualization features of the SWRM are exemplified in Fig. 6 and Figs. 9-11; all of the images in these figures were taken from the SWRM.

An on-demand verification tool, accessible via the PHI webpage (phi.nssl.noaa.gov), is currently under development. This tool will allow forecasters to assess PHI performance for specific date ranges and regions of interest within the evaluation period (1 June 2023 to present) and provide a baseline for performance comparison across different events.

Integrating Predicted Probabilities

The present implementation of PHI plumes uses the current probability taken from either PSv3 or TORP, applies a Gaussian or linear filter, respectively, and reduces that probability to zero at the end of the forecast period (either 1 hour for severe PHI or 30 min for tornado PHI). However, future iterations of PHI could incorporate probabilistic guidance from the Warn-on-Forecast System (e.g., WoFS-PHI; Loken et al. 2025) or other machine learning approaches, such as extended forecast TORP probabilities (Fig. 13). By integrating numerical modeling with current observations, these methods could maintain the accuracy of instantaneous probabilities while also providing improved forecasts of subsequent storm evolution. This approach would allow PHI to represent evolving storm trends beyond simply tapering to zero at the end of the forecast period. In addition, these methods could extend the forecast horizon beyond the current 30-60 min window, giving forecasters earlier indications of storm initiation and evolution and supporting

probabilistic IDSS initiatives within the NWS. Overall, this capability would enable forecasters to better anticipate changes in storm intensity (increase, decrease, or persistence), enhance the consistency and timeliness of warnings, and help communicate critical information to partners with greater lead time.

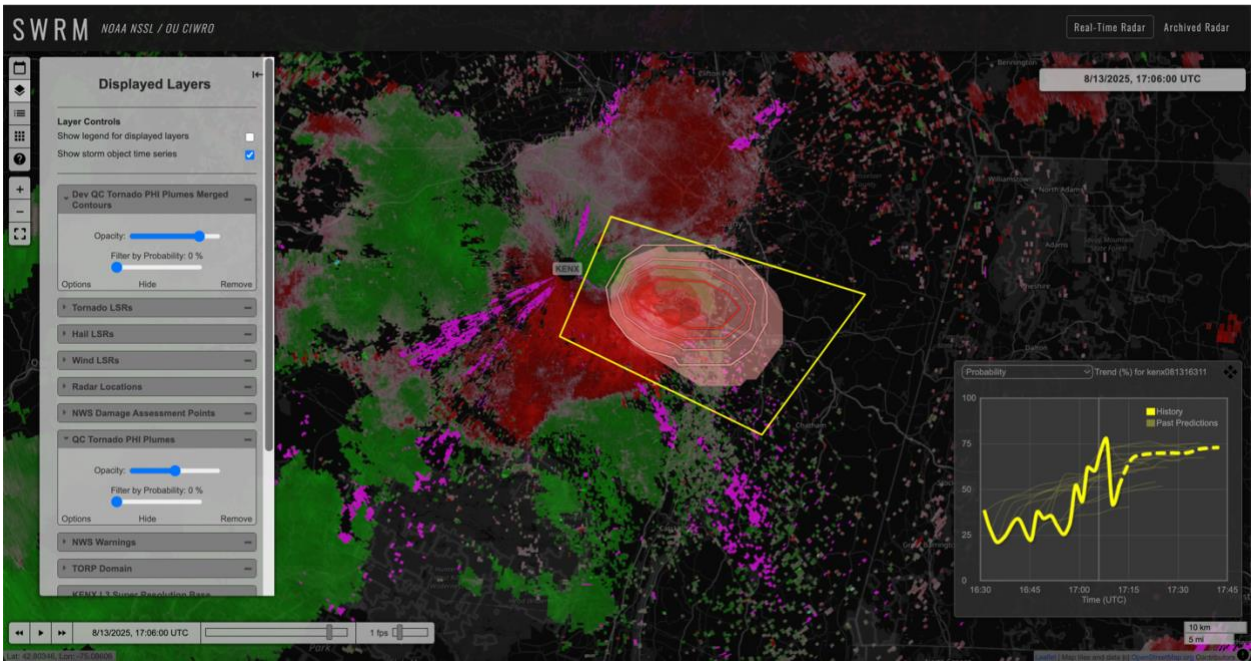

Figure 13: *Tornado PHI plume (red shaded contours) and NWS severe thunderstorm warning (yellow polygon) overlaid on 0.5° velocity from KENX at 1706 UTC 13 August 2025. Inset trend graph illustrates the probability trends for tornado PHI. The actual probabilities are represented by the bold line, while the darker thin lines show predictive probabilities from each current time. A vertical line on the graph indicates the current time of the plot.*

Warning Recommender within Hazard Services Severe

As options are considered for the operational implementation of PHI, one possible avenue is its use as a warning recommender. Hazard Services version 4 (HSv4), developed by the Global Systems Laboratory (GSL), is scheduled for nationwide installation in Q3 FY26 and is being tested at 14 WFOs. HSv4 may serve as the primary platform for issuing severe thunderstorm and tornado warnings. Building on forecaster feedback from this GR2 evaluation and verification statistics from the HWT, a "PHI Recommender" could provide as a first guess for severe thunderstorm warning areas, enhancing the accuracy, efficiency, and consistency of warning operations. A prototype of PHI as a recommender in HSv4 is shown in Fig. 14. This prototype ingests geoJSON files generated in the same manner as the GR2 placefiles used in this demonstration. By default,

the recommender could display thresholds informed by overall CSI from this PHI evaluation (40-60%) or allow customization by office using PSv3 performance results (Cintineo et al. 2024), with an adjustable slider for individual forecasters. Initial implementation would likely focus on severe thunderstorm warnings, with expansion to a tornado recommender at later stages. Importantly, the recommender is not intended to replace the forecaster, but rather to increase efficiency by providing a baseline for lower-priority storms, enabling forecasters to devote more attention to complex or high-impact situations. Establishing such a baseline would also promote greater consistency in both warning polygons and issuance practices across forecasters.

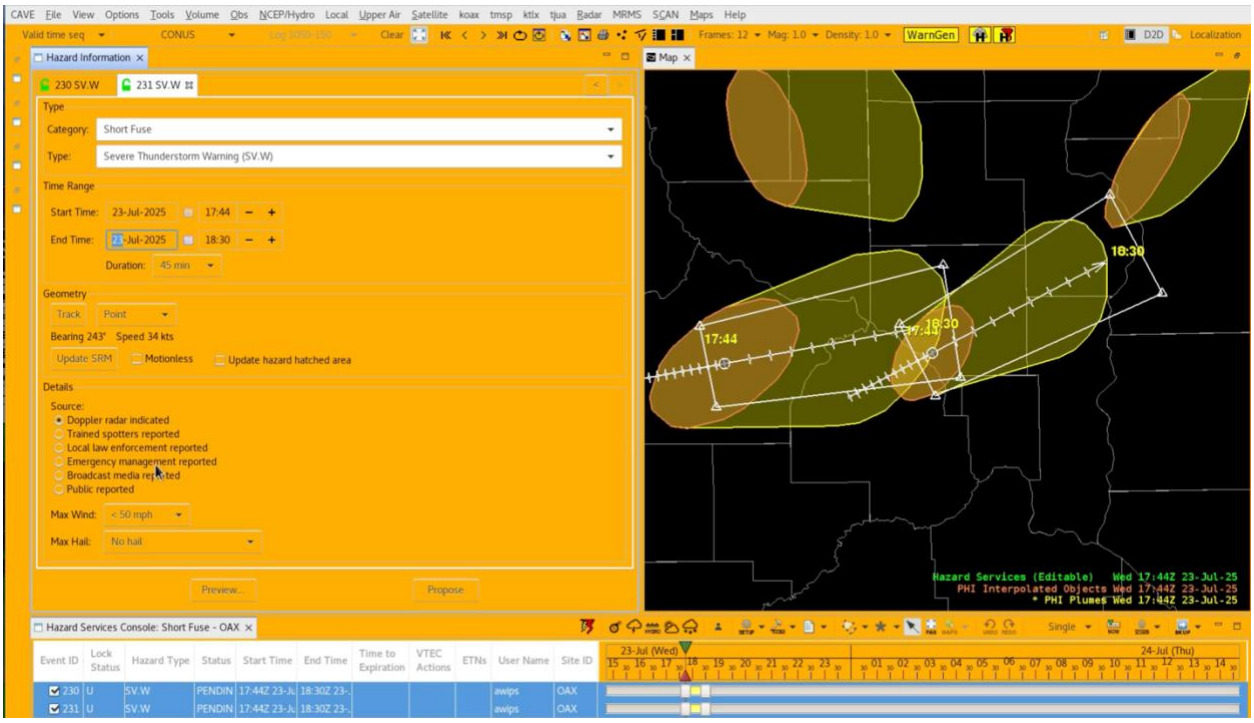

Figure 14: *A screen capture of a prototype HSv4 illustrates severe PHI plumes (yellow shaded) acting as a first guess or recommender for NWS severe thunderstorm warnings (white polygons). Warnings originate at the centroid of the PHI object (darker shaded ellipse), which indicates the current hazard location, and extend downstream to the outer edges of the PHI plume using the motion extracted from the PHI plume for the specified duration (45 min in the example shown here).*

5. Acknowledgements

We want to acknowledge all of the National Weather Service forecasters that have participated in reviewing the PHI plumes and the performance, as seen above, this feedback is incredibly valuable in shaping the development of PHI and the associated products. In particular, we want to thank Brian Carcione for helping to organize the initial implementation in NWS Southern Region and for continued discussion throughout the evaluation as well as Jeff Walstreicher for coordinating the extension into NWS Eastern Region offices. Credit goes to Kevin Manross at Global Systems Laboratory and the Cooperative Institute for Research in the Atmosphere for the PHI Hazard Services development and prototyping. Additionally, we want to thank John Cintineo for providing comments and suggestions that improved the report.

All graphics created using the SWRM use Basemap capabilities provided by OpenStreetMap under the Open Database License (https://www. openstreetmap.org/copyright) with base radar data from Iowa Environmental Mesonet or the National Weather Service. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA21OAR4320204, US. Department of Commerce. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect those of OAR or the Department of Commerce.

6. References

Allen, J. T., and M. K. Tippett, 2015: The Characteristics of United States Hail Reports: 1955–2014. Electron. J. Severe Storms Meteor., 10 (3), https://doi.org/10.55599/ejssm.v10i3.60.

Berry, K. and coauthors, 2024: The Experimental Warning Program, 2022–2023 Experiment Summary. NOAA/NSSL Technical Report. https://hwt.nssl.noaa.gov/ewp/archive/EWP2022-2023-Summary.pdf

Beven II, J. L., C. Fritz, and L. Alaka, 2025: *National Hurricane Center Tropical Cyclone Report: Hurricane Beryl (AL022024)*. NOAA/National Weather Service. 1 May 2025. https://www.nhc.noaa.gov/data/tcr/AL022024_Beryl.pdf

Boettcher, J. B., and E. S. Bentley, 2022: WSR-88D Sidelobe Contamination: From a Conceptual Model to Diagnostic Strategies for Improving NWS Warning Performance. *Wea. Forecasting*, 37, 853–869, https://doi.org/10.1175/WAF-D-21-0155.1.

Boettcher, J., S. Torres, F. Nai, C. Curtis, and D. Schvartzman, 2022: A Multidisciplinary Method to Support the Evolution of NWS Weather Radar Technology. *Wea. Forecasting*, 37, 429–444, https://doi.org/10.1175/WAF-D-21-0159.1

Bunkers, M. J., B. A. Klimowski, J. W. Zeitler, R. L. Thompson, and M. L. Weisman, 2000: Predicting supercell motion using a new hodograph technique. Wea. Forecasting, 15, 61–79, https://doi.org/10.1175/1520-0434(2000)015<0061:PSMUAN>2.0.CO;2.

Calhoun, K.M., P.A. Campbell, R.B. Steeves, T. Sandmael, C.N. Satrio, P.T. Hyland, J. G. Madden, and J.W. Monroe, 2024: Probabilistic Hazard Information and Threats-in-Motion: Testing the Future of Warnings and Storm-Based Hazard Creation and Communication. *104th AMS Annual Meeting, 2nd Symp. on the Future of Weather, Forecasting, and Practice.* Baltimore, MD. https://ams.confex.com/ams/104ANNUAL/meetingapp.cgi/Paper/430084

Causse, M., M. Mercier, O. Lefrançois, and N. Matton, 2025: Impact of Automation Level on Airline Pilots' Flying Performance and Visual Scanning Strategies: A Full Flight Simulator Study. *Applied Ergonomics*, 125, https://doi.org/10.1016/j.apergo.2024.104456.

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.

Cintineo, J. L., 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.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, A. Wimmers, J. Brunner, and W. Bellon, 2020: A Deep-Learning Model for Automated Detection of Intense Midlatitude Convection Using Geostationary Satellite Images. *Wea. Forecasting,* 35, 2567–2588, https://doi.org/10.1175/WAF-D-20-0028.1

Cintineo, J. L., M. J. Pavolonis, and J. M. Sieglaff, 2024: ProbSevere Version 3: Improved Exploitation of Data Fusion and Machine Learning for Nowcasting Severe Weather. *Wea. Forecasting,* 39, 1937–1958, https://doi.org/10.1175/WAF-D-24-0076.1.

Dhami, M. K. and D. R. Mandel, 2022: Communicating Uncertainty using Words and Numbers. *Trends in Cognitive Sciences*, 26 (6), 514-526. https://doi.org/10.1016/j.tics.2022.03.002.

Gesell, Ian, 2020: Verification of the Tornado and Lightning Plumes and Evaluation of a New Kernel for the Tornado and Lightning Plumes. *MS Thesis*, Univ. of Oklahoma, School of Meteorology. https://hdl.handle.net/11244/325322.

Joslyn, S. and S. Savelli, 2010: Communicating forecast uncertainty: public perception of weather forecast uncertainty. *Meteorol. Appl.*, 17, 180–195. https://doi.org/10.1002/met.190

Joslyn, S. L., and J. E. LeClerc 2012: Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. Journal of Experimental Psychology: Applied, 18 (1), 126–140. https://doi.org/10.1037/a0025185

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.

Karstens, C. D., and Coauthors, 2018: Development of a Human–Machine Mix for Forecasting Severe Convective Events. *Wea. Forecasting,* 33, 715–737, https://doi.org/10.1175/WAF-D-17-0188.1.

Kox, T., L. Gerhold, and U. Ulbrich, 2015: Perception and use of uncertainty in severe weather warnings by emergency services in Germany, *Atmospheric Research*, 159, 292-301, https://doi.org/10.1016/j.atmosres.2014.02.024 .

Krippendorff, K., 2013: Content Analysis: An Introduction to its Methodology (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Loken, E. D., and Coauthors, 2025: Combining Model and Observational Data Using Machine Learning for Short-Term Severe Weather Hazard Prediction. *Artif. Intell. Earth Syst.*, https://doi.org/10.1175/AIES-D-24-0102.1.

National Weather Service Learning Office, 2025: *Probabilistic Hazard Information for the NWS Field* [Video] WDTD ROOTs Seminar Series. YouTube. Recorded 27 Mar 2025. Accessed 15 August 2025. https://youtu.be/ETUFHDhbc1Y?si=rJpUK9DuTds3MnIP.

NOAA, 1950a: Storm Events Database. NOAA/National Centers for Environmental Information, accessed July 2025, https://www.ncdc.noaa.gov/stormevents.

NOAA, 1950b: SPC Local Storm Reports. NOAA/Storm Prediction Center, accessed July 2025, http://www.spc.noaa.gov/climo/online/.

NOAA, 2021: Research Tools: Decision Support. NOAA/National Severe Storms Laboratory, accessed August 2025, https://www.nssl.noaa.gov/tools/decision/.

Potvin, C. K., C. Broyles, P. S. Skinner, H. E. Brooks, and E. Rasmussen, 2019: A Bayesian Hierarchical Modeling Framework for Correcting Reporting Bias in the U.S. Tornado Database. *Wea. Forecasting*, 34, 15–30, https://doi.org/10.1175/WAF-D-18-0137.1.

Ripberger, J., A. Bell, A. Fox, A. Forney, W. Livingston, C. Gaddie, C. Silva, and H. Jenkins-Smith, 2022: Communicating Probability Information in Weather Forecasts: Findings and Recommendations from a Living Systematic Review of the Research Literature. *Wea. Climate Soc.*, 14, 481–498, https://doi.org/10.1175/WCAS-D-21-0034.1.

Rudlosky, S. D., and K. S. Virts, 2021: Dual Geostationary Lightning Mapper Observations. *Mon. Wea. Rev.*, 149, 979–998, https://doi.org/10.1175/MWR-D-20-0242.1.

Sandmæl, T. N., and Coauthors, 2023: The Tornado Probability Algorithm: A Probabilistic Machine Learning Tornadic Circulation Detection Algorithm. *Wea. Forecasting,* 38, 445–466, https://doi.org/10.1175/WAF-D-22-0123.1.

Smith, T. M., V. Lakshmanan, G. J. Stumpf, K. L. Ortega, K. Hondl, K. Cooper, K. M. Calhoun, D. M. Kingfield, K. L. Manross, R. Toomey, and J. Brogden, 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.

Steeves, R. B., P. A. Campbell, and T. M. Smith, 2020: A Web-Based Visualization Tool for FACETs. *100th AMS Annual Meeting, 36th Conf. on Environ. Inf. Processing Tech.,* 15 Jan 2020, Boston, MA. https://ams.confex.com/ams/2020Annual/webprogram/Paper368627.html

Sutton, R. T., D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, K. I. Kroeker, 2020: An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med.* Feb 6. 3:17. doi: 10.1038/s41746-020-0221-y.

Thompson, R. L., R. Edwards, and C. M. Mead, 2004: An update to the supercell composite and significant tornado parameters. 22nd Conf. on Severe Local Storms, Hyannis, MA, Amer. Meteor. Soc., P8.1, https://ams.confex.com/ams/11aram22sls/techprogram/paper_82100.htm.

Thompson, R. L., B. T. Smith, J. S. Grams, A. R. Dean, and C. Broyles, 2012: Convective Modes for Significant Severe Thunderstorms in the Contiguous United States. Part II: Supercell and QLCS Tornado Environments. *Wea. Forecasting*, 27, 1136–1154, https://doi.org/10.1175/WAF-D-11-00116.1.

Trapp, R. J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and R. Wolf, 2006: Buyer Beware: Some Words of Caution on the Use of Severe Wind Reports in Postevent Assessment and Research. *Wea. Forecasting*, 21, 408–415, https://doi.org/10.1175/WAF925.1.

Wendt, N. A., and I. L. Jirak, 2021: An Hourly Climatology of Operational MRMS MESH-Diagnosed Severe and Significant Hail with Comparisons to Storm Data Hail Reports. *Wea. Forecasting*, 36, 645–659, https://doi.org/10.1175/WAF-D-20-0158.1.

Zhu, Y., M. Stock, J. Lapierre, and E. DiGangi, 2022: Upgrades of the Earth Networks Total Lightning Network in 2021. *Remote Sensing*, 14, 9, 2209. https://doi.org/10.3390/rs14092209