BAMS
Article

The Experimental Warning Program of NOAA's Hazardous Weather Testbed

Kristin M. Calhoun, Kodi L. Berry, Darrel M. Kingfield, Tiffany Meyer, Makenzie J. Krocak, Travis M. Smith, Greg Stumpf, and Alan Gerard

ABSTRACT: NOAA's Hazardous Weather Testbed (HWT) is a physical space and research framework to foster collaboration and evaluate emerging tools, technology, and products for NWS operations. The HWT's Experimental Warning Program (EWP) focuses on research, technology, and communication that may improve severe and hazardous weather warnings and societal response. The EWP was established with three fundamental hypotheses: 1) collaboration with operational meteorologists increases the speed of the transition process and rate of adoption of beneficial applications and technology, 2) the transition of knowledge between research and operations benefits both the research and operational communities, and 3) including end users in experiments generates outcomes that are more reliable and useful for society. The EWP is designed to mimic the operations of any NWS Forecast Office, providing the opportunity for experiments to leverage live and archived severe weather activity anywhere in the United States. During the first decade of activity in the EWP, 15 experiments covered topics including new radar and satellite applications, storm-scale numerical models and data assimilation, total lightning use in severe weather forecasting, and multiple social science and end-user topics. The experiments range from exploratory and conceptual research to more controlled experimental design to establish statistical patterns and causal relationships. The EWP brought more than 400 NWS forecasters, 60 emergency managers, and 30 broadcast meteorologists to the HWT to participate in live demonstrations, archive events, and data-denial experiments influencing today's operational warning environment and shaping the future of warning research, technology, and communication for years to come.

KEYWORDS: Forecasting techniques; Nowcasting; Operational forecasting; Communications/decision making; Experimental design; Societal impacts

https://doi.org/10.1175/BAMS-D-21-0017.1
Corresponding author: Kristin Calhoun, kristin.calhoun@noaa.gov
In final form 9 June 2021

AFFILIATIONS: Calhoun, Berry, and Gerard—NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma; **Kingfield**—NOAA/Global Systems Laboratory, Boulder, Colorado; **Meyer**—UCAR/Unidata, Boulder, Colorado; **Krocak**—Cooperative Institute for Mesoscale Meteorological Studies, and Center for Risk and Crisis Management, University of Oklahoma, Norman, Oklahoma; **Smith**—Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma; **Stumpf**—Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado, and NOAA/NWS/Meteorological Development Laboratory, Silver Spring, Maryland

Serafin et al. (2002) stated the need for a national testbed, closely linked to an operational center, where suggestions for model improvements would be subjected to rigorous systematic evaluation. Participants would be intimately involved in the testing, with access to the full operational data stream and knowledge of the operational staff. The interactions between the research, academic, and operational communities would provide a direct transfer of research into operational forecasting while also working as a mechanism for the needs and challenges of operational forecasting to influence atmospheric research. While their concept of a national testbed focused on numerical weather prediction, Serafin et al. (2002) noted that additional test centers could focus on other aspects of transitioning research to operations. Over the past 18 years, NOAA established 12 testbeds and proving grounds to facilitate the transition of research capabilities to operational implementation.

NOAA's Hazardous Weather Testbed (HWT) is both a physical space and research framework that fosters collaboration and the testing and evaluation of emerging tools, technology, and products for NWS operations. The HWT is a joint project between the NWS Storm Prediction Center (SPC), NSSL, and NWS Norman, Oklahoma, Weather Forecast Office (WFO) all located in the National Weather Center (NWC) building in Norman, Oklahoma. The goal of the HWT is to accelerate the transition of new meteorological insights and technologies into forecasting and severe weather warning operations. The HWT comprises two programs: the Experimental Forecast Program (EFP) and the Experimental Warning Program (EWP). The EFP focuses on the use of convection-allowing model ensembles to improve predictions of hazardous and convective weather events from a few hours to a week in advance across various spatial scales ranging from several counties to the continental United States (Kain et al. 2003, 2006; Clark et al. 2012, 2018, 2020; Gallo et al. 2017). The EWP focuses on severe weather research and technology to improve the WFOs' severe weather warnings for hail, wind, and tornadoes.

The EWP began through outreach to individual WFOs to test new warning ideas, software, and algorithms. Prior to its formal development, operational tests and evaluations of NSSL's single and polarimetric radar algorithms (e.g., Scharfenberg et al. 2003, 2005a), Warning Decision Support–Integrated Information (WDSS-II) radar displays (e.g., Scharfenberg et al. 2004), and initial Multi-Radar Multi-Sensor (MRMS) algorithms (including rotation tracks and hail diagnosis products; e.g., Stumpf et al. 2003a) were completed at local WFOs. However, during these operational tests and evaluations, it was difficult for forecasters and researchers to focus on product and algorithm evaluation during warning operations when life and property were at stake. Furthermore, rapid prototyping of new data and applications was not easy or even permitted by outside researchers within individual WFOs. The HWT provides an environment not only for rapid prototyping, but also for exploratory research and proof-of-concept testing prior to operational implementation, embedding forecasters and researchers alongside software developers and training entities.

Following a proposal in 2005 to create an operational environment similar to a WFO within the HWT, the EWP formally became part of the HWT activities and WFO Norman forecasters were invited to participate. Since 2006, the EWP has continually expanded to include more projects and forecasters. Forecasters outside of WFO Norman were invited by 2008. In 2010, a formal application process was created for forecasters across the NWS. In 2011, some experiments included military forecasters. By 2014, hydrologists and broadcast meteorologists were invited to participate in select experiments. End users, including emergency managers, were first incorporated in testbed activities in 2010 and began routinely participating in experiments starting in 2015. Between 2008 and 2020, the EWP hosted 437 forecasters, 61 emergency managers, and 38 broadcast meteorologists to evaluate products and tools for severe weather events.

While the experiment design and types of participants has evolved over the last 12 years, the goals of EWP experiments in the HWT originate from the same ideas: 1) assess the operational utility of new scientific concepts and technologies, 2) provide direct feedback from forecasters to developers on the strengths and limitations of their concepts, 3) offer insights to better meet the needs of operational forecasters, and 4) transfer knowledge of research concepts and ideas to the operational environment. This article details the evolution of the EWP, the data collection process, and highlights results and outcomes from experiments over the last decade.

Technical design of the HWT

The physical space of the HWT is between the SPC and WFO Norman in the National Weather Center. The left half of the room is configured for EWP operations and the right for EFP operations (Fig. 1). NWS forecasters use the Advanced Weather Interactive Processing System (AWIPS) to view meteorological datasets and issue products. To emulate the operations at a WFO, the EWP has used AWIPS since 2008 as the software foundation upon which all experiments are built (Kingfield and Magsig 2009). In 2012, the EWP transitioned to the second generation AWIPS (AWIPS2) environment 2 years before the operational deployment to the NWS. This early implementation of AWIPS2 served as a risk-reduction exercise for the larger forecasting community. The EWP was able to vet a number of potential problems with the AWIPS2 system, finding solutions prior to the national software rollout. Furthermore, it provided a sandbox for investigators to evaluate the implementation of their technologies, allowing for iterative development and boosting readiness levels for future operational implementation.

Currently, the HWT has 14 standalone AWIPS2 workstations, two 27-in. monitors per workstation, nine large televisions in the corners of the room for situational awareness, two tables for group discussions, and a telephone to simulate communication between offices. For additional space, experiments can expand into the Development Laboratory (Fig. 1, bottom right), which is equipped with nine additional standalone AWIPS2 workstations, two monitors per workstation, three large televisions on two walls, and a table for group discussions. Displays can be shared between the two rooms.

The investigators developing EWP experiments consult with the HWT technical leads to ensure their hypotheses can be properly evaluated with the AWIPS2 system. This includes coordination on the types of products ingested, whether new AWIPS2 code is needed, and whether external systems are required to support the experiment (e.g., web-based displays or data collection systems). The AWIPS2 workstations are configured for either live or archived meteorological data. Live weather experiments allow for "real world" evaluation of new algorithms and technologies as if these components were operational. Experiments using archived data allow for controlled conditions and with repeatable experimental processes while maintaining an operational environment. The systems can move between live and archive modes seamlessly so multiple experiments can run simultaneously.

During live operations, the EWP servers ingest all data from the Satellite Broadcast Network (SBN) that are available to any NWS WFO. While many numerical models are clipped to an NWS office's area of responsibility, this ability was disabled to allow for participants to work in multiple NWS offices simultaneously during EWP operations. Higher-resolution WSR-88D

Fig. 1. (top) Activities in the HWT during both EWP and EFP simultaneous experiments. (bottom left) Schematic of the room set up for the majority of EWP experiments; forecasters are typically working in pairs at adjacent workstations acting as a single NWS weather forecast office. (bottom right) EWP activities in the Development Laboratory within the National Weather Center. This extra space facilitates multiple experiments simultaneously.

data are critical for NWS operations, but are not available on the SBN due to the high volume of information provided by each radar site. To incorporate this information into EWP operations, up to 10 Open Radar Product Generator (ORPG; Crum et al. 1998) processes can be activated via a web interface to ingest live Level-II data and provide a full suite of WSR-88D base and derived fields with minimal latency. Experimental datasets are ingested through the Local Data Acquisition and Dissemination (LDAD) system, similar to how NWS offices currently receive non-SBN data. All workstations can be localized to any NWS office and have the ability to issue simulated short-fused warning products (e.g., severe thunderstorm and tornado warnings) through the WarnGen application.

To support the ingesting of archived meteorological datasets for archive operations, software was developed to feed information into AWIPS2 at a controlled pace. The ingest time could be either the archived event time or a false date and time to help reduce recognition of past events. Screen capture ability and large digital clocks were added to the display environments to document the forecasters' actions and streamline subsequent data analysis by the experiment investigators.

EWP experiment design

Though the EWP began solely as a spring experiment focused on current weather, archive data allows experiments to occur year-round. Experiments in the EWP now consist of a mixture of live demonstrations, archive cases, and data-denial experiments. Typical experiments include three to six NWS forecasters traveling to Norman, Oklahoma, each week with experiments lasting between 3 and 5 weeks. For most experiments, Monday is an orientation and training day, in which forecasters familiarize themselves with the experimental tools, software, and/or datasets. Tuesday through Thursday are operational days. On these days, forecasters test experimental data and tools using live data and/or archived weather cases. Many experiments begin the day with an archive weather case and follow up with a live weather evaluation later that day. Each day concludes with surveys and a semi-structured discussion focused on the utility of the tools and data. Friday is typically reserved as a debrief day, in which participants reflect on their experiences with the data and tools throughout the entire week.

The design of each EWP project is tailored to specific research goals. As a result, EWP projects are, in general, either exploratory or experimental (also commonly referred to as confirmatory or a priori hypothesis testing). Exploratory projects typically involve rapid prototyping and queries of new algorithms, ideas, and concepts that have not fully been defined. These types of experiments aim to provide insight to researchers and developers on how a forecaster might interact with new technologies or data. Exploratory research may also focus more on the communication and collaboration of new concepts or data with end users. These exploratory experiments are typically flexible and adapt to feedback as the experiment progresses.

Experimental projects follow a specific outline of repeatable hypothesis testing across multiple participants from different backgrounds for a variety of weather and locations. This can provide a more thorough examination of how a new product/algorithm/tool would impact the operational environment, corresponding output, and information dissemination and use. In forecaster-focused research, these experiments focus primarily on product or instrument evaluations and typically use data-denial or descriptive research tools, such as ranked surveys. Often, the goal is to provide statistical details on the use of products and algorithms, including the changes in opinion and behavior of forecasters over time. In end-user-focused research, these types of experiments can explore, for example, connections between individual predispositions toward uncertainty information and preferences for product design.

As seen in the following examples, many experiments move from exploratory to more formal experimental research over subsequent years as part of the EWP. This design allows for incremental development of larger concepts and ideas, while still making progress toward the operational transition of research. This repeated, incremental development ultimately provides better outcomes while also increasing the speed of the transition process and rate of adoption of beneficial applications and technology (Clark et al. 2012; Gallo et al. 2017).

Researchers use several methods to collect feedback from participants, including focus groups, surveys, and researcher notes (e.g., Calhoun et al. 2014). Some of the most valuable participant feedback collected during exploratory research is through focus groups and individual discussions. These conversations are crucial to gathering participant opinions about changes in workload, utility of the new tool/product/process, and interpretation of the information. During these discussions, researchers use a focus group discussion guide where specific topics are designated but emergent issues can be investigated in more depth by the focus group moderator. This flexibility is important because it allows participants to offer qualitatively rich insights on important topics that are not known a priori. Oftentimes, participants will highlight a challenge or benefit of the experimental product or system that researchers have not considered previously. The integration of users in a naturalistic decision environment is essential for the research-to-operations process.

In addition to focus groups, researchers also use surveys to collect data before, during, and after the participants are introduced to the experimental items. This method allows for the collection of anonymous data and unfiltered opinions. Over numerous experiments, surveys can also be collated for meaningful statistical analysis of more hypothesis-driven work.

Researchers often employ a pretest and posttest method to elicit feedback about information gaps and then assess whether or not the new product/tool/system helped improve those gaps. While most surveys are designed with close-ended questions for ease of analysis, open-ended questions are also employed to gather more specific details about the benefits and challenges of the experimental items.

During many experiments, participants are paired with a researcher that observes what data are being used and how they are used for warning decisions. Researchers take notes of actions, decisions, and challenges the participants are experiencing, particularly related to interface usage and workflow. These researchers also ask questions and clarify why a participant is using a new tool in certain ways to elicit feedback about potential modifications to experimental products. The notes taken during the simulations are useful for researchers to reference during data analysis because they add context to survey responses and focus group notes.

Finally, if the primary goal of the experiment is to determine whether the experimental datasets help forecaster warning decisions, blog posts can be a useful way for forecasters to express their thoughts as they go through the warning process. Participants are encouraged to take screen captures of what products they are looking at and write a short blog post explaining how the experimental products played a role in their warning decision. These posts are particularly helpful because they are written from the participants' perspective. Having participants explain the benefits and challenges of an experimental product in their own words highlights details that researchers may not have considered previously.

Past and current EWP experiments

Experiment names/subjects, acronym definitions, years of activity, and associated publications are included in Table 1. A brief summary including any unpublished results or unique methods is included for many of the experiments below.

Radar experiments.

MRMS-Severe. Prior to becoming operational (Smith et al. 2016), the MRMS-Severe algorithms involved several years of iterative testing and development. Initial development was coordinated with individual WFOs. The associated limitations of testing MRMS products in WFOs during severe weather operations was part of the impetus for the creation of the EWP. Once in the HWT, investigators employed flexible testing methods including open-ended questions and discussions during a combination of live weather and archived cases that they could not have completed during actual warning operations. These exploratory experiments provided guidance for additional development on the most beneficial algorithms (e.g., rotation tracks and the maximum expected size of hail) and allowed for incremental modifications to visualizations to appropriately fit within the typical forecaster workflow. Later testing used archived cases in a controlled environment to examine whether MRMS products could reduce false alarm area, improve polygon alignment, speed up storm diagnosis, and enhance lead time. These tests helped determine best practices for use and were coordinated with the training community to provide recommendations for integration with other operational products (Bates et al. 2015).

Phased Array Radar Innovative Sensing Experiment (PARISE). Using multiple cases of tornadic and severe storms in simulated operational environments with archive data, the PARISE experiments investigated the hypothesis that higher temporal resolution radar data would positively impact the warning decision process (Heinselman et al. 2012, 2015). These experiments used a controlled environment providing different temporal resolution data (e.g., 4.5-min versus sub-1-min updates) to different forecasters. This research established that

Table 1. The experiment names, year(s), principal investigator(s), and relevant references for the experiments conducted in the EWP.

Experiment	Years	Principal investigators	References
Multi-Radar Multi-Sensor (MRMS) Severe	2005, 2008–10, 2013–14	T. Smith, G. Stumpf, K. Scharfenberg, K. Manross, J. LaDue and K. Ortega	Smith et al. (2003); Stumpf et al. (2003a,b); Scharfenberg et al. (2005b); LaDue et al. (2013)
Phased Array Radar Innovative Sensing Experiment (PARISE)	2008–10, 2014–15	P. Heinselman, K. Wilson, D. Kingfield	LaDue et al. (2010); Heinselman et al. (2012, 2015); Bowden et al. (2015); Bowden and Heinselman (2016); Wilson et al. (2016, 2017a,b, 2018)
Collaborative Adaptive Sensing of the Atmosphere (CASA)	2007–10	B. Philips, J. Brotzge, T. Smith, G. Stumpf	Philips et al. (2008, 2010); Brotzge et al. (2010); Bass et al. (2011); Rude et al. (2012)
Dual-Polarization Hail Size Discrimination	2013	K. Ortega	Ortega et al. (2016)
Radar Convective Applications	2018–21	B. Smith, T. Sandmael	Sandmael et al. (2020)
Conditional Probability of Tornado Intensity	2018	B. Smith, M. Mahalik	
GOES–JPSS	2009–21	S. Goodman, D. Lindsay, C. Siewart, B. Line, M. Bowlan, K. Calhoun	Bedka et al. (2010); Sieglaff et al. (2011); Bikos et al. (2012); Walker et al. (2012); Goodman et al. (2013); Cintineo et al. (2014); Schmit et al. (2014); Line et al. (2016); Calhoun (2018, 2019); Bruning et al. (2019); Esmaili et al. (2020)
Lightning Jump Algorithm	2013–16	K. Calhoun, L. Carey, D. Kingfield, E. Schultz,	Chronis et al. (2014); Calhoun et al. (2015)
Earth Networks Total Lightning Network	2015–16	K. Calhoun, D. Kingfield, T. Meyer	Calhoun et al. (2016)
3DVAR	2011–12	K. Calhoun, T. Smith, J. Gao, D. Stensrud	Calhoun et al. (2014)
OUN-WRF	2011–14	G. Garfield, A. Anderson	
Variational Local Analysis and Prediction System	2013–14	H. Jiang, Y. Xie, S. Albers, I. Jankov, L. Wharton, Z. Toth	Jiang et al. (2015)
MRMS Hydrometeorological Testbed–Hydrology	2014–16, 2018–19	J. Gourley, S. Martinaitis, P.-E. Kirstetter, H. Vergara, K. Wilson, N. Yussouf	Argyle et al. (2017); Gourley et al. (2017); Martinaitis et al. (2017, 2020); Yussouf et al. (2020)
Probabilistic Hazard Information (PHI)	2008, 2014–21	G. Stumpf, T. Smith, K. Manross, K. Ortega, K. Calhoun, C. Karstens, J. Correia, C. Ling, J. James, L. Rothsfusz	Kuhlman et al. (2009); Karstens et al. (2015, 2018); Ling et al. (2015, 2017); Stumpf et al. (2015, 2018); Bates et al. (2019); James et al. (2020); Manross et al. (2021)
PHI End Users	2015–20	K. Berry, H. Obermeier, K. Klockow-McClain, D. LaDue, M. Krocak, K. Wilson	Nemunaitis-Berry and Obermeier (2017); Klockow-McClain et al. (2020); Obermeier et al. (2018, 2019, 2020)
Severe Weather and Society Dashboard	2020	K. Klockow-McClain, J. Ripberger, M. Krocak	Ripberger et al. (2019, 2020)

additional lead time and increased probability of detection (POD) was directly related to access to the higher temporal resolution data. The PARISE experiments (Heinselman et al. 2015; Bowden et al. 2015) emphasized discussion with forecasters postevent to establish how phased-array radar data were blended into the forecaster workflow to account for the additional lead time. Later experiments combined this retrospective discussion with eye-tracking software to better understand the specific attention and cognitive process of the warning forecaster (Wilson et al. 2016).

Collaborative Adaptive Sensing of the Atmosphere (CASA). The CASA experiment evaluated the use of gap-filling radars during warning decisions and the associated impacts to end-user communication (Bass et al. 2011; Rude et al. 2012). Similar to the MRMS experiments, initial years of the CASA experiment were exploratory. These experiments focused on the strengths and limitations of the new gap-filling technology and assessed other potential benefits of the data. Later years of the CASA experiments wished to expressly determine: 1) which fine-scale rotations warranted a warning, 2) how 1-min updates resolved forecaster uncertainty, and 3) a model for decision-making and communication interactions among spotters, emergency managers, and NWS forecasters for severe weather.

Radar convective applications. The ongoing NWS Radar Operations Center (ROC) experiments examine new and updated single-radar products and algorithms for the WSR-88D network including 1) velocity-derived azimuthal shear (AzShear), 2) a new mesocyclone detection algorithm (MDA), and 3) a new tornado detection algorithm (TDA). The primary goal of these evaluations is to determine if the new algorithms are justified replacements for the legacy versions. Results from 2019 emphasized the impact single-radar AzShear could have on NWS warning operations as it highlights key features in velocity data that are precursors to tornadic circulations, especially those associated with challenging quasi-linear convective systems. Crucially, investigators also learned that any updates to the MDA and TDA would be ineffectual unless the visualizations were improved from the original design. In 2019, the initial implementation of the new algorithms mimicked the same table listing and readout as the original MDA/TDA (Fig. 2, left). Forecasters informed researchers that it was the clunky visualizations combined with a high false alarm rate that made the legacy detection algorithms inadequate. Thus, subsequent development following the initial experiment focused on updating the visualization from the undesirable table design to an object-based interactive design with temporary readouts that did not interfere with ongoing analysis (Fig. 2, right). Virtual experiments in 2021 examine whether the new foundation for the MDA and TDA increases forecaster use and satisfaction with the algorithms.

Satellite and lightning experiments.

GOES–JPSS. Prior to the launch of the GOES-R series of satellites, there was keen interest in demonstrating new capabilities and algorithms as well as providing operational forecasters a chance to increase knowledge of the upcoming upgrades. Since the new GOES series had temporal and spatial resolutions that allow for use in the prediction and diagnosis of hazardous weather, Satellite Proving Ground (initially referred to as "GOES-R Proving Ground"; Goodman et al. 2012) evaluations were included in the EWP. Today, the

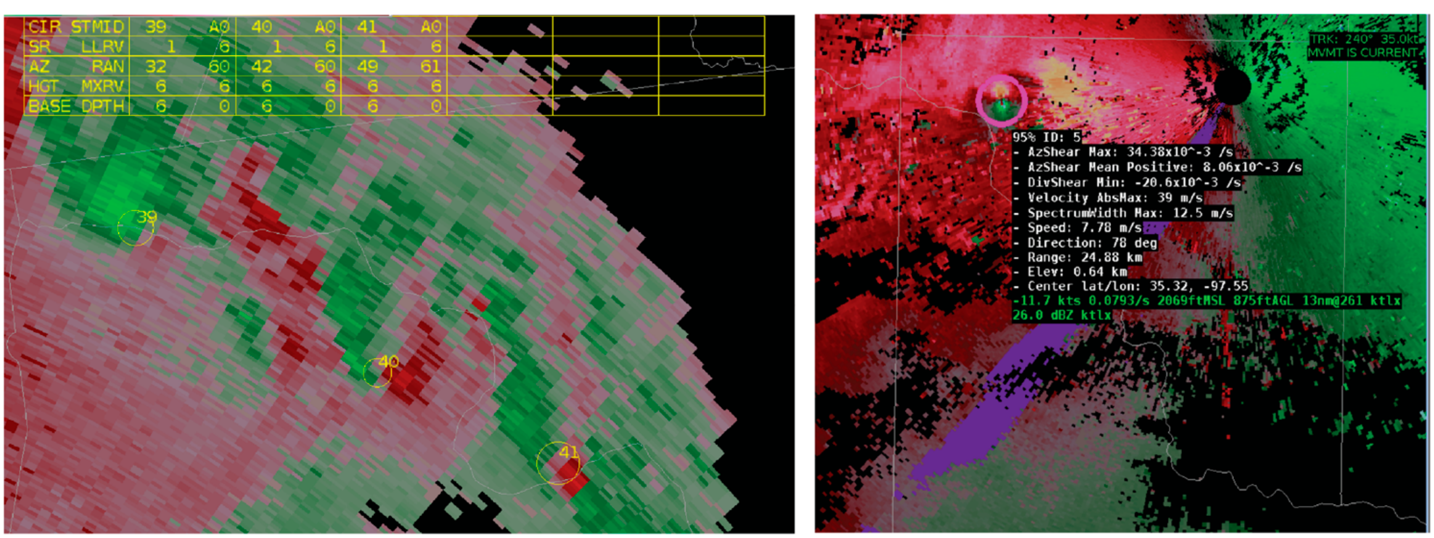

Fig. 2. (left) Original implementation of the tornado detection algorithm and (right) updated version based on forecaster feedback. New visualization follows the design of the ProbSevere algorithm seen by forecasters in the satellite experiments.

EWP continues to provide the GOES and Joint Polar Satellite System (JPSS) programs an opportunity each spring to conduct demonstrations of baseline, future capabilities, and experimental products.

The initial satellite demonstrations focused on preparation for the launch of the GOES-R satellite. In 2009, this first demonstration was limited to weather briefings and overviews of product performance to participants. By 2010, the importance of including the satellite-based products directly into forecaster software and workflow was keenly noted by the principal investigators and the data and algorithms were examined within AWIPS2 for the first time. This change provided the opportunity to more directly understand the impact on the storm diagnosis and severe weather warnings. These initial experiments consisted of expected baseline and level-2 products and algorithms including convective initiation nowcasting and probabilities (Sieglaff et al. 2011; Walker et al. 2012), overshooting top and thermal couplet detection (Bedka et al. 2010), total lightning detection and a pseudo-Geostationary Lightning Mapper Product (Goodman et al. 2013), simulated satellite imagery (Bikos et al. 2012), and a 0–3 h severe hail probability (a preliminary version of the ProbSevere Statistical Model; Cintineo et al. 2014).

In 2013, to better prepare a wider audience of NWS operational forecasters for the new GOES-R series of satellites, participants were asked to participate in a webinar hosted by the NWS Warning Decision Training Division (WDTD) called "Tales from the Testbed." Forecasters presented the experimental products and how they performed during different events and locations throughout the week. In 2014, *GOES-14* Super Rapid Scan Operations were introduced to see how the higher temporal resolution expected from the GOES-R series impacted use of satellite data during warning operations (Schmit et al. 2014; Line et al. 2016). Feedback from the forecasters in the EWP experiment during these first five years helped to determine which products would be integrated into the GOES-R baseline and level-2 product dissemination after launch.

Beginning in 2015, JPSS algorithms from the NOAA Unique Combined Atmospheric Processing System (NUCAPS) were included to assess the value of soundings from polar orbiting satellites in filling the temporal and spatial gaps between the standard NWS daily sounding radiosonde measurements. As Esmaili et al. (2020) discuss, this EWP evaluation both helped NUCAPS developers gain a better appreciation of operational requirements and limitations. This led to improvements of NUCAPS functionality within sounding toolkits (such as within the SHARPpy software by Blumberg et al. 2017) and focused investigators toward issues that impact usability in the severe weather environment (such as more accurately representing the boundary layer).

Following the launch of the GOES-R satellite, activity during 2017–19 spring experiments shifted to focus on both validation of algorithms and testing new data [e.g., the Geostationary Lightning Mapper (GLM)]. Forecasters evaluated new convectively applicable baseline products, such as total precipitable water, derived stability indices, and derived motion winds, as well as multispectral (red–green–blue) composites and channel differences. Not only did the 1-min imagery aid in the initiation and updraft monitoring phase of convection, but it also aided forecasters in null cases and in identifying areas that were more stable and might not result in storm development.

The baseline Legacy Atmospheric Profile retrieval algorithms and layer predictable water products for moisture and stability coverage were typically used prior to convective initiation to analyze gradients. However, use of these products were minimal in the warning and mesoscale analysis environment as products were only available every 30 min with forecasters commonly noting that blending with a higher-resolution convective allowing model over CONUS for 5- to 15-min products would be much preferred.

Additionally, the initial presentation of GLM data within AWIPS2 in 2017 was quickly found to be problematic. EWP scientists strongly suggested that the data and associated visualization should not be provided to NWS operational offices until both geolocation errors within the ground-system and the initial visualization were fixed. This feedback drove a focused effort between academic and federal partners to develop gridded imagery that retains the quantitative physical measurements and better illustrates how lightning discharges illuminate thunderstorms (Bruning et al. 2019). The new gridded GLM products were part of the 2018 and 2019 EWP demonstrations where feedback from forecasters helped shape the individual product visualizations and determined new products, such as minimum flash size, for operational implementation and use (Calhoun 2018, 2019). As a new instrument with unique visualizations, forecasters reported having a subject matter expert available to answer questions on GLM data made the greatest impact on product understanding and use throughout the week.

Overall, the feedback from 10 years of GOES and JPSS experiments has greatly shaped not only the products and visualizations available to the NWS, but has also produced training and best practices for the use of satellite data in hazardous weather forecasting for all forecasters across the NWS.

Lightning jump algorithm. In severe storms, rapid increases in lightning flash rate, or "lightning jumps," are coincident with pulses in the storm updraft and typically precede severe weather at the surface by tens of minutes. In support of future GLM capabilities, different implementations of a total lightning jump algorithm (LJA) were tested during the 2013–16 EWP Spring Experiments in coordination with the GOES–JPSS proving ground demonstrations. The goals were to determine if the LJA could be used by NWS forecasters to enhance situational awareness, diagnose convective trends, and potentially improve the short-term prediction of severe weather following the results of Schultz et al. (2009, 2011). Initially, LJA testing was limited to regions with Lightning Mapping Arrays. Later, the LJA was expanded to include data from Earth Networks Total Lightning Network (ENTLN) to test CONUS-wide.

The LJA greatly benefitted from repeated testing and incremental development. It initially focused solely on the 2-sigma value (or twice the standard deviation of the 1-min flash rate) and initial experiments quickly informed researchers that this one-size-fits-all approach would not work. Between annual demonstrations, significant changes were made to the algorithm and visualization to show not only the degree of the jump (e.g., the standard deviation), but also the associated trends (Fig. 3). While one jump was important, forecasters were also interested in noting cases where multiple jumps happened for the same storm. Similarly, while forecasters liked the rapid 1- or 2-min updates, they found sometimes they missed the peak sigma value, so a 5-min max sigma product (updating every minute) was created. By the 2016 evaluation, primary feedback focused on the visualization; suggestions included either moving the visualization to an opaque display or to an outline with details in a "mouseover" for the current flash rate to facilitate quick storm comparisons, similar to the ProbSevere product (Fig. 3). This demonstration also highlighted how reviewing products in an operational setting next to other products is beneficial, not only because it provides context relative to operational products, but this can also spur advancements in other developmental products. In the case of the LJA and ProbSevere, the corresponding evaluations led to discussions across principal investigators from both projects, ultimately leading to the addition of lightning data to ProbSevere which reduced the false alarm rate (Cintineo et al. 2018). The 1-min update and 5-min max LJA products were equally regarded as individual forecasters generally chose depending on the operational focus of the day. As such, it was suggested that both move forward as operational products. Forecaster feedback was incorporated into MRMS training through the WDTD and both the 5-min and 1-min products became part of operational MRMS version 12 in 2020.

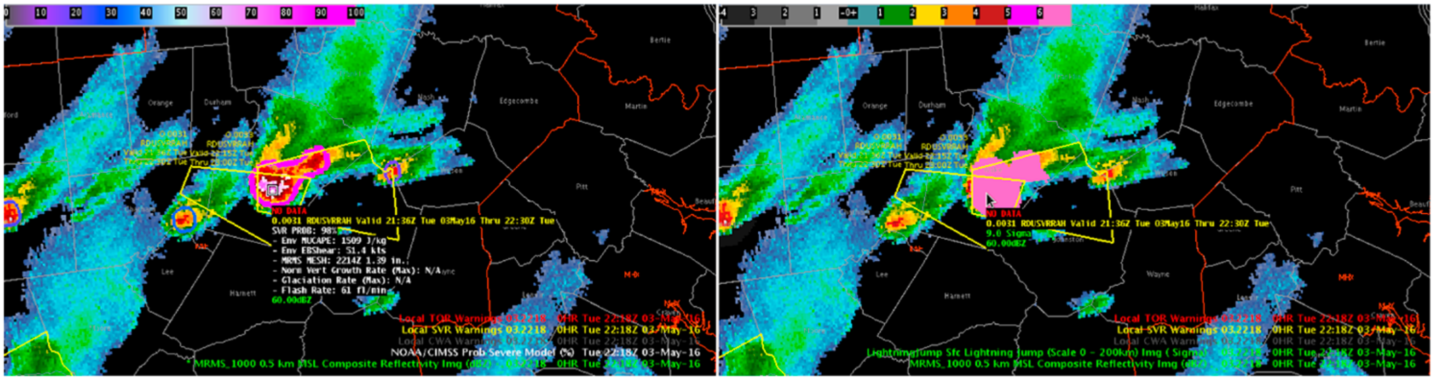

Fig. 3. Forecaster display from AWIPS2 during HWT evaluation on 3 May 2016 in Wake County, NC, showing experimental warnings with MRMS 0.5 km MSL composite reflectivity, (left) ProbSevere or (right) lightning jump algorithm. As noted in a blog post by the forecaster, the forecaster had high confidence in extending the warning due to a 9-sigma lightning jump (as shown in the mouseover).

Earth networks' dangerous thunderstorm alerts. Earth Networks, Incorporated (ENI), indicated the potential for their total lightning data and automated Dangerous Thunderstorm Alerts (DTA; Liu and Heckman 2012) system to help increase lead times over current NWS severe weather and tornado warnings, while maintaining a similar POD.

To test the value and impact of ENI total lightning data and algorithms, both the ENTLN data and products from the DTA system were evaluated during two experiments.

The first experiment used data denial and rotating control group testing to evaluate the hypothesis that different data access would affect forecaster warning decisions. This experiment included 18 NWS forecasters over 6 weeks in 2014. Each forecaster was isolated and rotated through a series of six 2-h-long archived weather-warning simulations across a variety of convective regimes ranging from marginally severe to high-impact tornadic events. Using a repeated measures design, the forecaster was randomly assigned one of three tiers of data during the simulation: 1) the full suite of WSR-88D radar products, 2) ENTLN total lightning point data and products available in Tier 1, or 3) ENI total lightning cell tracking, flash rate products, and associated alert polygons in addition to all products available in Tier 2.

All tiers performed similarly with the Tier 2 group (total lightning + radar) performing slightly better in terms of overall false alarm ratio and POD. Based on these skill scores, forecasters that were not already experts at radar interrogation and severe storm forecasting saw the most benefit from the inclusion of total lightning data and associated products during warning operations. Overall, the indicated that while the forecasters found the total lightning and derived products useful in warning operations, the DTA polygons themselves had limited value.

A successive live weather evaluation was completed during the 2015 spring experiment. Like other live experiments, the operational domains were decided daily based on likelihood of severe weather. During this experiment, we operated in 42 different NWS county warning areas with 31 forecasters evaluating ENTLN total lightning and associated products. This second experiment provided insight on how the forecasters would use the data for warning operations alongside all currently available products and provided a stress test for the timeliness and usability of the ENTLN data and tools within operations. Similar to year one, forecasters gravitated to the total lightning data and storm-based derived flash rate trend information as well as the time series display. Again, multiple forecasters noted the three layers of alerts (including the DTAs) cluttered the screen and had too high of a false alarm rate to add value to the warning process. However, a majority of forecasters did find value in the total lightning tools, such as the storm tracking and time series information. Furthermore, some forecasters required these additional tools and derived products to feel comfortable utilizing the total lightning data in operations due to limited training currently available on lightning data.

Modeling/data assimilation experiments.

3DVAR. An adaptive storm-scale three-dimensional variational data assimilation (3DVAR) analysis system was introduced to forecasters as part of the Spring Experiment in 2011 and 2012 (Calhoun et al. 2014). This exploratory research investigated a possible implementation of Warn-on-Forecast (Stensrud et al. 2009) concepts with operational forecasters through the visualization of dynamically consistent gridded analyses driven by data assimilation. The storm-scale analyses provided updraft, storm-top divergence, and vorticity in addition to multilevel wind vectors. While forecasters found that the analyses added confidence in warning decisions, often with additional lead time when storms were close to a WSR-88D radar, data latency as little as 5 min prevented consistent use within warning operations as new radar volumes were completed within this period. As such, forecasters placed more weight on the new radar information that could highlight new and/or different information from the 3DVAR analyses. Even with the positive feedback by forecasters, transition to operations was not continued due to the high impact of data latency.

OUN-WRF. In 2011, an exploratory investigation of the Norman Weather Forecast Office Weather Research Forecast Model (OUN-WRF) was created to better understand the potential impacts of higher-resolution and locally relevant numerical weather data on convective-scale analysis and warnings to better inform the development of concepts such as Warn-on-Forecast (Stensrud et al. 2009). Unlike high-resolution models run at national centers, the configuration of the local OUN-WRF Model for these experiments was flexible, allowing for parameterization sets to be optimized daily for expected weather over the Southern Plains. Through discussions and use, participating forecasters acquired expertise in identifying the impact of parameterizations, enabling them to account for the impact on a forecast. Investigators used the experiment to better understand not only which particular products were of use for situational awareness and warnings, but also how they were used. Iterative years of the experiment tuned the type of products available (as products such as updraft helicity were found to be especially relevant) and also tested how model cycling and time-lagged ensembles could be visualized within the forecaster workflow. The end result of this experiment was a transfer of knowledge not only to the operational community, but also within the high-resolution modeling community regarding product development and visualizations for operations.

Flash flood experiments. The Hydrometeorology Testbed (HMT) MRMS–Hydrology (HMT-Hydro) experiments used a structured framework to better understand how experimental NWP forecasts and hydrologic model guidance affected warnings for flash flood events (Martinaitis et al. 2017, 2020). Forecasters and hydrologists from NWS WFOs and River Forecast Centers evaluated the Flooded Locations and Simulated Hydrographs (FLASH; Gourley et al. 2017) system, use of the Hazard Services software (Argyle et al. 2017) with initial assessments of flash flood recommenders (Martinaitis et al. 2017), and coupled probabilistic hydrologic modeling with probabilistic ensemble QPFs (Yussouf et al. 2020) across multiple years of testing. These experiments uniquely created collaborations between the Weather Prediction Center's annual Flash Flood and Intense Rainfall (FFaIR) experiment (Barthold et al. 2015) and HMT-Hydro experiments. This joint effort successfully simulated the workflow between two testbed environments across multiple national centers in different time zones (e.g., Martinaitis et al. 2020). The most recent experiments compared the output of forecasters using the experimental products to the operational products to provide an objective measure of increased lead time and POD, while using discussions and surveys to better understand the various strengths and challenges and participant perceptions on how new products influenced flash flood operations. Similar to other EWP experiments such as the GLM demonstration, discussion with participants underscored the impact of having subject-matter experts present with forecasters and other end users (Martinaitis et al. 2020). Future work will examine the implementation of the probabilistic flash flood data into the Hazard Services software platform.

Warning-paradigm experiments. In 2008, scientists began testing the concept of a rapidly updating high-resolution gridded Probabilistic Hazard Information (PHI) system with forecasters as part of the EWP. During these initial exploratory experiments, forecasters were asked to create PHI in lieu of the current deterministic warnings available to the public today. However, researchers found forecasters needed a baseline probability for calibration across events (Kuhlman et al. 2008). PHI has greatly evolved since this first implementation, based on both forecaster and end-user feedback as well as through the incorporation of new algorithms and tools that better address the ideas of Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018). The major emphasis of more recent PHI experiments has been on initial testing of concepts related to human–computer interaction (e.g., Karstens et al. 2015, 2018) while generating short-fused high-impact PHI for severe weather. Human factors (Ling et al. 2015; James et al. 2020) and end-user experiments with emergency managers and broadcast meteorologists (described in more detail below) have been coordinated with many of the NWS forecaster experiments providing an end-to-end experiment to better understand the impact of PHI from creation to use. Simultaneously, we continue to test these refined concepts and methodologies from earlier PHI prototype experiments and transition them into an experimental version of Hazard Services (the next generation warning tool for the NWS), including initial steps toward this paradigm through threats-in-motion (Stumpf and Gerard 2021), for further testing and evaluation prior to deployment.

End user experiments. The inclusion of core partners of the NWS within experiments adds an extra layer to discussions and product use and creation. In 2014, the GOES–JPSS experiment included roughly one broadcast meteorologist each week, assuming the role of warning forecaster. The benefits of incorporating broadcasters in the GOES–JPSS experiment included familiarizing them with NWS operations, the warning decision process, AWIPS2, and the challenges forecasters face. Additionally, broadcasters contributed unique feedback and provided NWS forecasters with the broadcast perspective on severe weather warning coverage. As a result of the engagement and shared edification of all partners, the EWP made a concerted effort to incorporate NWS core partners into more experiments going forward.

Emergency managers and broadcast meteorologists were incorporated into PHI experiments as both standalone participant groups and collective integrated warning team participants dating back to 2015. Research with these core partners focuses on how the continuous flow of probabilistic information from days before the event through the warning time scale may be received, understood, and used to make important decisions (Klockow-McClain et al. 2020). Emergency management participants simulated decisions for towns or areas that matched the scale of their jurisdiction (e.g., university, city, county, state). Researchers investigated how the experimental products changed the decisions they made or the timeframes during which they made decisions during severe weather events. Broadcast participants performed typical job functions in a mock television studio environment (Fig. 4) as they received experimental probabilistic information from forecasters. Research protocols were used to systematically study how broadcast meteorologists interpreted, used, and communicated

Fig. 4. Broadcast meteorologists participating in an EWP experiment in a mock television studio. Participants took turns broadcasting on the wall and attending to social media responsibilities.

probabilistic information and rapidly updating warnings to their hypothetical viewing audience (Obermeier et al. 2018, 2019, 2020).

Social data experiments. The 2020 Severe Weather and Society Dashboard experiment was conducted to assess the utility of a new type of product with forecasters—one that provides social and behavioral data about the communities that NWS forecasters serve. Containing data from the annual Severe Weather and Society Survey (Silva et al. 2017, 2018, 2019; Ripberger et al. 2019, 2020), the dashboard was the first primarily social science tool to be tested in the HWT. A wide variety of data were presented to participants, from composite indices that measure tornado warning reception, understanding, and response, to geographic risk perceptions of different weather hazards and demographic variables. The overall utility of the dashboard tool was evaluated along with the usefulness of the data within the dashboard and the time scales on which the data would be most useful.

Researchers are pursuing paths for operationalization of the dashboard and future community-level data. Surveys focusing on other weather hazards including tropical cyclones, winter storms, and fire weather are currently under development, with future plans to create dashboards similar to the severe weather prototype.

The future of the EWP

During the period of the COVID-19 pandemic, the HWT faced new challenges that are shaping the future of EWP operations. The inability to host activities in person resulted in the transition to virtual activities. Initial virtual efforts were limited, but 2021 EWP evaluations use AWIPS2 in the cloud for five experiments. Moving forward, the HWT EWP hopes to incorporate a combination of in-person and virtual activities.

Another shift in HWT experiment design is to expand across the EFP and EWP bridging across time scales and reference classes. Motivated by FACETs (Rothfusz et al. 2018), severe weather research is now focusing on providing a continuous flow of information across the entire space–time continuum of a severe weather event. This approach allows researchers to consider all reference classes and reference frames from days before to within minutes of an event (Klockow-McClain 2019; Klockow-McClain et al. 2020) to ensure they tell a consistent and cohesive story. Expanding experiments beyond only forecast or warning time scales will also allow for the inclusion of interactions that occur between national centers and WFOs.

We expect funding and interest in the research-to-operations process to continue to bring a diverse group of projects and ideas through the EWP. Though the scope of individual projects may change, the EWP will continue to provide a means of transitioning knowledge, new technologies, and applications to operations while also exploring innovative concepts and ideas that will shape the future of warning applications and research for years to come.

The EWP in a Virtual Environment

While impossible to entirely replicate the operational and collaborative capabilities of the HWT, remote collaboration and evaluation provides an opportunity for continued testing and evaluation instead of shutting experiments completely down. The COVID-19 era has had impacts on multiple sectors, and the EWP experiments are no different. The initial onset provided little time to fully prepare to replicate the EWP testing and operational capabilities in a virtual environment in spring 2020. However, through much development and dependence on virtual and cloud-based platforms, multiple experiments are testing the waters of remote evaluation in 2021. From previous lessons learned, we knew the ability to use the operational software would be crucial for testing items near transition to operations. Leveraging AWIPS2 in the cloud as the backbone for testing, we are using a combination of WES-2 Bridge software from the NWS Warning Decision Training Division and locally developed software to continue experiments using archive data. Live weather evaluations use local data managers to provide data to the cloud platform. As expected though, a completely new environment brings complications; we are now dependent on participant home hardware, bandwidth and capacity, and consideration of time zones in planning experiments. This requires multiple hours for rehearsal of logistics, dependencies, and platform configurations. While remote collaboration and evaluation can provide to an increased number and wider array of participants without additional travel costs, we are finding a need to simplify archive cases and evaluations to meet the new limitations. Even with an increased number of collaborative tools, such as Google Meet and Slack, we are limited in how much of the operational environment we can recreate. We currently find ourselves isolating forecasters into multiple Meets to maintain simultaneous discussions with principal investigators, but this loses the collaborative aspect of being in the same location—handing off storms and discussions of who is covering what threat becomes a juggling activity bouncing across multiple Google Meets and crossing messages as opposed to casually asking a question across a room. Additionally, troubleshooting technical problems on remote desktops leads to full break in data collection and discussion as the forecaster is forced to participate in the troubleshooting process. Yet, even with these limitations, we have still been able to continue testing products, ideas, and algorithms for operational implementation. Forecasters that have participated in virtual EWP experiments in 2021 have said they would definitely participate again virtually, but overall prefer the in-person experience and collaborative aspect of being in the same location with other forecasters, scientists, and end users. Looking years in the future, we hope to use this virtual experience to find a balance between the benefits of in-person and remote experiments.

Acknowledgments. The HWT EWP was made possible by the many dedicated participants, investigators, and internal support staff over the last decade. Special thanks to Dave Andra, Adrian Campbell, Vicki Farmer, Mike Foster, Gabe Garfield, Kim Klockow-McClain, William Line, Jonathan Madden, Kevin Manross, Justin Monroe, James Murnan, Kiel Ortega, Lans Rothfusz, Thea Sandmael, Kevin Scharfenburg, Russ Schneider, Brandon Smith, and Alex Zwink. HWT Infrastructure Support was provided by NOAA/Office of Oceanic and Atmospheric Research/Weather Program Office (WPO). Travel support for forecasters and participants was provided by the GOES-R program office, WPO, NWS, and individual grants from Private Investigators. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA16OAR4320115, U.S. Department of Commerce.

Data availability statement. No datasets were generated or analyzed during the current study.

References

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

Barthold, F. E., T. E. Workoff, B. A. Cosgrove, J. J. Gourley, D. R. Novak, and K. M. Mahoney, 2015: Improving flash flood forecasts: The HMT-WPC flash flood and intense rainfall experiment. *Bull. Amer. Meteor. Soc.*, **96**, 1859–1866, https://doi.org/10.1175/BAMS-D-14-00201.1.

Bass, E. J., and Coauthors, 2011: A method for investigating real-time distributed weather forecaster-emergency manager interaction. *2011 IEEE Int. Conf. on Systems, Man, and Cybernetics*, Anchorage, AK, IEEE, 2809–2815, https://doi.org/10.1109/ICSMC.2011.6084098.

Bates, A. V., and Coauthors, 2019: Preparing for FACETs: The need for inclusive National Weather Service forecaster training and innovative collaboration tools. *9th Conf. on Transition of Research to Operations*, Phoenix, AZ, Amer. Meteor. Soc., 13A.5, https://ams.confex.com/ams/2019Annual/webprogram/Paper349804.html.

Bedka, K., J. Brunner, R. Dworak, W. Feltz, J. Otkin, and T. Greenwald, 2010: Objective satellite-based detection of overshooting tops using infrared window channel brightness temperature gradients. *J. Appl. Meteor. Climatol.*, **49**, 181–202, https://doi.org/10.1175/2009JAMC2286.1.

Bikos, D., and Coauthors, 2012: Synthetic satellite imagery for real-time high-resolution model evaluation. *Wea. Forecasting*, **27**, 784–795, https://doi.org/10.1175/WAF-D-11-00130.1.

Blumberg, W. G., K. T. Halbert, T. A. Supinie, P. M. Marsh, R. L. Thompson, and J. A. Hart, 2017: SHARPpy: An open-source sounding analysis toolkit for the atmospheric sciences. *Bull. Amer. Meteor. Soc.*, **98**, 1625–1636, https://doi.org/10.1175/BAMS-D-15-00309.1.

Bowden, K. A., and P. L. Heinselman, 2016: A qualitative analysis of NWS forecasters' use of phased-array radar data during severe hail and wind events. *Wea. Forecasting*, **31**, 43–55, https://doi.org/10.1175/WAF-D-15-0089.1.

——, P. Heinselman, D. M. Kingfield, and R. Thomas, 2015: Impacts of phased array radar data on forecaster performance during severe hail and wind events. *Wea. Forecasting*, **30**, 389–404, https://doi.org/10.1175/WAF-D-14-00101.1.

Brotzge, J., K. Hondl, B. Philips, L. Lemon, E. J. Bass, D. Rude, and D. L. Andra Jr., 2010: Evaluation of distributed collaborative adaptive sensing for detection of low-level circulations and implications for severe weather warning operations. *Wea. Forecasting*, **25**, 173–189, https://doi.org/10.1175/2009WAF2222233.1.

Bruning, E. C., and Coauthors, 2019: Meteorological imagery for the geostationary lightning mapper. *J. Geophys. Res. Atmos.*, **124**, 14 285–14 309, https://doi.org/10.1029/2019JD030874.

Calhoun, K. M., 2015: Forecaster use of total lightning data for short-term forecasts and warnings in the Hazardous Weather Testbed. Seventh Conf. on Meteorological Applications of Lightning Data, Phoenix, AZ, Amer. Meteor. Soc., 3.5, https://ams.confex.com/ams/95Annual/webprogram/Paper268336.html].

——, 2018: Feedback and recommendations for the Geostationary Lightning Mapper (GLM) in severe and hazardous weather forecasting and warning operations. NOAA Rep., 15 pp., https://hwt.nssl.noaa.gov/ewp/projects/GLM-HWT-report_2018.pdf.

——, 2019: Feedback and recommendations for the Geostationary Lightning Mapper (GLM) in severe and hazardous weather forecasting and warning operations. NOAA Rep., 16 pp., https://hwt.nssl.noaa.gov/ewp/projects/GLM-HWT-report-2019.pdf.

——, 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, https://doi.org/10.1175/WAF-D-13-00107.1.

——, D. M. Kingfield, T. Meyer, W. Roberts, J. Ramer, B. Motta, and L. Rothfusz, 2016: Earth networks total lightning data and dangerous thunderstorm alerts evaluation in the NOAA hazardous weather testbed. NOAA Doc., 33 pp., http://hwt.nssl.noaa.gov/ewp/ENI_HWT_Finalreport_Mar2016.pdf.

Chronis, T., and Coauthors, 2014: National demonstration and evaluation of a real time lightning jump algorithm for operational use. *26th Conf. on Weather Analysis and Forecasting/22nd Conf. on Numerical Weather Prediction*, Atlanta, GA, Amer. Meteor. Soc., 4B.1, https://ams.confex.com/ams/94Annual/webprogram/Paper232183.html.

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

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

Clark, A. J., and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed experimental forecast program spring experiment. *Bull. Amer. Meteor. Soc.*, **139**, 55–74, https://doi.org/10.1175/BAMS-D-11-00040.1.

——, and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed spring forecasting experiment. *Bull. Amer. Meteor. Soc.*, **99**, 1433–1448, https://doi.org/10.1175/BAMS-D-16-0309.1.

——, and Coauthors, 2020: A real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bull. Amer. Meteor. Soc.*, **101**, E2022–E2024, https://doi.org/10.1175/BAMS-D-19-0298.1.

Crum, T. D., R. E. Saffle, and J. W. Wilson, 1998: An update on the NEXRAD program and future WSR-88D support to operations. *Wea. Forecasting*, **13**, 253–262, https://doi.org/10.1175/1520-0434(1998)013<0253:AUOTNP>2.0.CO;2.

Esmaili, R. B., and Coauthors, 2020: Adapting satellite soundings for operational forecasting within the hazardous weather testbed. *Remote Sens.*, **12**, 886, https://doi.org/10.3390/rs12050886.

Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/hazardous weather testbed spring forecasting experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1.

Goodman, S. J., and Coauthors, 2012: The GOES-R proving ground: Accelerating user readiness for the next-generation geostationary environmental satellite system. *Bull. Amer. Meteor. Soc.*, **93**, 1029–1040, https://doi.org/10.1175/BAMS-D-11-00175.1.

——, and Coauthors, 2013: The GOES-R Geostationary Lightning Mapper (GLM). *Atmos. Res.*, **126**, 34–49, https://doi.org/10.1016/j.atmosres.2013.01.006.

Gourley, J. J., and Coauthors, 2017: The Flooded Locations and Simulated Hydrographs (FLASH) project: Improving the tools for flash flood monitoring and prediction across the United States. *Bull. Amer. Meteor. Soc.*, **98**, 361–372, https://doi.org/10.1175/BAMS-D-15-00247.1.

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

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

James, J. J., C. Ling, C. Karstens, J. Correia, K. Calhoun, T. Meyer, and D. Ladue, 2020: Forecasters' cognitive task analysis and mental workload analysis of issuing probabilistic hazard information (PHI) during FACETs PHI prototype experiment. *Wea. Forecasting*, **35**, 1505–1521, https://doi.org/10.1175/WAF-D-19-0194.1.

Jiang, H., and Coauthors, 2015: Real-time applications of the variational version of the Local Analysis and Prediction System (vLAPS). *Bull. Amer. Meteor. Soc.*, **96**, 2045–2057, https://doi.org/10.1175/BAMS-D-13-00185.1.

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

Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF Model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181, https://doi.org/10.1175/WAF906.1.

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

——, and Coauthors, 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.

Kingfield, D. M., and M. A. Magsig, 2009: Leveraging National Weather Service technology for collaboration and training. *25th Conf. on International Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Phoenix, AZ, Amer. Meteor. Soc., P1.9, https://ams.confex.com/ams/pdfpapers/150678.pdf.

Klockow-McClain, K. E., 2019: Probability of… what, exactly? Establishing a simple ground rule for the communication of forecast uncertainty. *5th Conf. on Weather, Warnings, and Communications.* San Diego, CA, Amer. Meteor. Soc., J8.1, https://ams.confex.com/ams/47BC5WxComm/webprogram/Paper358771.html.

——, and Coauthors, 2020: Putting multiple probabilistic products before end users: The 2019 HWT Emergency Manager experiments. *15th Symp. On Societal Applications: Policy, Research and Practice*, Boston, MA, Amer. Meteor. Soc., 17.6, https://ams.confex.com/ams/2020Annual/webprogram/Paper369752.html.

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. *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., 8A.2, https://ams.confex.com/ams/24SLS/techprogram/paper_142027.htm.

——, E. Gruntfest, K. A. Scharfenberg, and G. J. Stumpf, 2009: Beyond storm-based warnings: An advanced WAS*IS workshop to study communication of probabilistic hazardous weather information. 4th Symp. on Policy and Socio-Economic Research, Phoenix, AZ, Amer. Meteor. Soc., 3.5, https://ams.confex.com/ams/89annual/techprogram/paper_150887.htm.

LaDue, J. G., and Coauthors, 2013: The multiple-radar/multiple-sensor (MRMS) severe weather algorithm experiment in the hazardous weather testbed in 2013. *37th Annual Meeting,* Madison, WI, National Weather Association, https://nwas.org/annual-meeting-events/past-meetings/2013-agenda/.

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

Line, W. E., T. J. Schmit, D. T. Lindsey, and S. J. Goodman, 2016: Use of geostationary super rapid scan satellite imagery by the storm prediction center. *Wea. Forecasting*, **31**, 483–494, https://doi.org/10.1175/WAF-D-15-0135.1.

Ling, C., L. Hua, C. D. Karstens, G. J. Stumpf, T. M. Smith, K. M. Kuhlman, and L. Rothfusz, 2015: A comparison between WarnGen system and probabilistic hazard information system for severe weather forecasting. *2015 International Annual Meeting*, Los Angeles, CA, Human Factors and Ergonomics Society, 1791–1795, https://doi.org/10.1177/1541931215591387.

——, and Coauthors, 2017: Forecasters' mental workload while issuing Probabilistic Hazard Information (PHI) during 2016 FACETs PHI Hazardous Weather Testbeds. *33rd Conf. on Environmental Info. Processing Tech.*, Seattle, WA, Amer. Meteor. Soc., J9.3, https://ams.confex.com/ams/97Annual/webprogram/Paper314172.html.

Liu, C., and S. Heckman, 2012: Total lightning data and real-time severe storm prediction. *Conf. on Meteorological and Environmental Instruments and Methods of Observation*, Brussels, Belgium, World Meteorological Organization, P5 (10), https://library.wmo.int/pmb_ged/iom_109_en/Session5/P5_10_Liu_Total_Lightning_Data_and_Real-Time_Severe_Storm_Prediction.pdf.

Manross, K. L., and Coauthors, 2021: Updates on hazard service—Probabilistic hazard information (HS-PHI). *11th Conf. on Transition of Research to Operations*, Online, Amer. Meteor. Soc., 532, https://ams.confex.com/ams/101ANNUAL/meetingapp.cgi/Paper/384508.

Martinaitis, S. M., and Coauthors, 2017: The HMT Multi-Radar Multi-Sensor Hydro Experiment. *Bull. Amer. Meteor. Soc.*, **98**, 347–359, https://doi.org/10.1175/BAMS-D-15-00283.1.

——, and Coauthors, 2020: The 23 June 2016 West Virginia flash flood event as observed through two Hydrometeorology Testbed experiments. *Wea. Forecasting*, **35**, 2099–2126, https://doi.org/10.1175/WAF-D-20-0016.1.

Nemunaitis-Berry, K. L., and H. M. Obermeier, 2017: Broadcast meteorologist decision making in the 2016 Hazardous Weather Test Probabilistic Hazard Information Project. *Bull. Amer. Meteor. Soc.*, **98**, 2045–2047, https://doi.org/10.1175/BAMS_9810_2039-2052_Nowcast.

Obermeier, H., K. L. Nemunaitis-Berry, K. E. Klockow, C. D. Karstens, A. Gerard, and L. P. Rothfusz, 2018: Broadcast meteorologist decision-making in the 2017 Hazardous Weather Testbed Probabilistic Hazard Information Project. *13th Symp. on Societal Applications: Policy, Research & Practice*, Austin, TX. Amer. Meteor. Soc., 8.4, https://ams.confex.com/ams/98Annual/webprogram/Paper329820.html.

——, ——, ——, T. C. Meyer, P. A. Campbell, A. Gerard, and C. Kolakoski, 2019: Communicating probabilistic hazard information: Broadcast meteorologists in the 2018 Hazardous Weather Testbed. *47th Conf. on Broadcast Meteorology*, San Diego, CA, Amer. Meteor. Soc., J7.6, https://ams.confex.com/ams/47BC5WxComm/webprogram/Paper358620.html.

——, ——, ——, ——, ——, ——, J. E. Trujillo, and C. Carithers, 2020: Communicating probabilistic hazard information: Broadcast meteorologists in the 2018-19 Hazardous Weather Testbed. 15th Symp. on Societal Applications: Policy, Research and Practice, Boston, MA. Amer. Meteor. Soc., 12A.4, https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/368267.

Ortega, K. L., J. M. Krause, and A. V. Ryzhkov, 2016: Polarimetric radar characteristics of melting hail. Part III: Validation of the algorithm for hail size discrimination. *J. Appl. Meteor. Climatol.*, **55**, 829–848, https://doi.org/10.1175/JAMC-D-15-0203.1.

Philips, B., D. Westbrook, D. Pepyne, J. Brotzge, E. J. Bass, and D. Rude, 2008: User evaluations of adaptive scanning patterns in the CASA spring experiment 2007. *2008 IEEE Int. Geoscience and Remote Sensing Symp.*, Boston, MA, IEEE, 156–159, https://doi.org/10.1109/IGARSS.2008.4780051.

——, V. Chandrasekar, J. Brotzge, M. Zink, H. Rodriguez, C. League, and W. Diaz, 2010: Performance of the CASA radar network during the May 13, 2009 Anadarko tornado. *15th Symp. on Meteorological Observation and Instrumentation*, Atlanta, GA. Amer. Meteor. Soc., 9.3, https://ams.confex.com/ams/90annual/techprogram/paper_165864.htm.

Ripberger, J. T., M. J. Krocak, W. W. Wehde, J. N. Allan, C. Silva, and H. Jenkins-Smith, 2019: Measuring tornado warning reception, comprehension, and response in the United States. *Wea. Climate Soc.*, **11**, 863–880, https://doi.org/10.1175/WCAS-D-19-0015.1.

——, C. Silva, H. Jenkins-Smith, J. Allan, M. J. Krocak, W. Wehde, and S. Ernst, 2020: Exploring community differences in tornado warning reception, comprehension, and response across the United States. *Bull. Amer. Meteor. Soc.*, **101**, 936–948, https://doi.org/10.1175/BAMS-D-19-0064.1.

Rothfusz, L. P., R. Schneider, D. Novak, K. Klockow-McClain, A. E. Gerard, C. Karstens, G. J. Stumpf, and T. M. Smith, 2018: FACETs: A proposed next-generation paradigm for high-impact weather forecasting. *Bull. Amer. Meteor. Soc.*, **99**, 2025–2043, https://doi.org/10.1175/BAMS-D-16-0100.1.

Rude, D. J., E. J. Bass, and B. Philips, 2012: Quantifying the impact of adding gap filling radar data on forecaster wind assessments, warnings, and confidence. *Meteor. Appl.*, **19**, 355–370, https://doi.org/10.1002/met.269.

Sandmael, T., K. L. Elmore, and B. R. Smith, 2020: A new machine learning-based tornado detection algorithm for the WSR-88D Network. *19th Conf. on Artificial Intelligence for Environmental Science*, Boston, MA, Amer. Meteor. Soc., P364, https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/363586.

Scharfenberg, K. A., D. L. Andra Jr., and M. P. Foster, 2003: Operational uses of

polarimetric radar in severe local storm prediction. *31st International Conf. on Radar Meteorology*, Seattle, WA, Amer. Meteor. Soc., P4B.3, https://ams.confex.com/ams/pdfpapers/64415.pdf.
——, D. J. Miller, D. L. Andra Jr., and M. P. Foster, 2004: Overview of spring 2004 WDSS-II demonstration at WFO Norman. *22nd Conf. on Severe Local Storms*, Hyannis, MA, Amer. Meteor. Soc., 8B.7, https://ams.confex.com/ams/11aram22sls/techprogram/paper_81808.htm.
——, and Coauthors, 2005a: The joint polarization experiment: Polarimetric radar in forecasting and warning decision-making. *Wea. Forecasting*, **20**, 775–788, https://doi.org/10.1175/WAF881.1.
——, T. M. Smith, and G. J. Stumpf, 2005b: The testing of NSSL multi-sensor applications and data from prototype platforms in NWS forecast operations. Preprints, *21st Conf. on Weather Analysis and Forecasting*, Washington, DC, Amer. Meteor. Soc., 6A.2, https://ams.confex.com/ams/WAFNWP34BC/techprogram/paper_95026.htm.
Schmit, T. J., and Coauthors, 2014: Rapid refresh information of significant events: Preparing users for the next generation of geostationary operational satellites. *Bull. Amer. Meteor. Soc.*, **96**, 561–576, https://doi.org/10.1175/BAMS-D-13-00210.1.
Schultz, C. J., W. A. Petersen, and L. D. Carey, 2009: Preliminary development and evaluation of lightning jump algorithms for the real-time detection of severe weather. *J. Appl. Meteor. Climatol.*, **48**, 2543–2563, https://doi.org/10.1175/2009JAMC2237.1.
——, ——, and ——, 2011: Lightning and severe weather: A comparison between total and cloud-to-ground lightning trends. *Wea. Forecasting*, **26**, 744–755, https://doi.org/10.1175/WAF-D-10-05026.1.
Serafin, R. J., A. E. MacDonald, and R. L. Gall, 2002: Transition of weather research to operations: Opportunities and challenges. *Bull. Amer. Meteor. Soc.*, **83**, 377–392, https://doi.org/10.1175/1520-0477(2002)083<0377:TOWRTO>2.3.CO;2.
Sieglaff, J. M., L. M. Cronce, W. F. Feltz, K. M. Bedka, M. J. Pavolonis, and A. K. Heidinger, 2011: Nowcasting convective storm initiation using satellite-based box-averaged cloud-top cooling and cloud-type trends. *J. Appl. Meteor. Climatol.*, **50**, 110–126, https://doi.org/10.1175/2010JAMC2496.1.
Silva, C., J. Ripberger, H. Jenkins-Smith, and M. Krocak, 2017: Establishing the baseline: Public reception, understanding, and responses to severe weather forecasts and warnings in the contiguous United States. Reference Rep., Center for Risk and Crisis Management, 31 pp., http://risk.ou.edu/downloads/news/WX17-Reference-Report.pdf.
——, ——, ——, ——, and W. Wehde, 2018: Refining the baseline: Public reception, understanding, and responses to severe weather forecasts and warnings in the contiguous United States. Reference Rep., Center for Risk and Crisis Management, 29 pp, http://risk.ou.edu/downloads/news/WX18-Reference-Report.pdf.
——, ——, ——, ——, S. Ernst, and A. Bell, 2019: Continuing the baseline: Public reception, understanding, and responses to severe weather forecasts and warnings in the contiguous United States. Reference Rep., Center for Risk and Crisis Management, 33 pp, http://risk.ou.edu/downloads/news/WX19-Reference-Report.pdf.
Smith, T. M., G. J. Stumpf, K. L. Manross, and C. Thomas, 2003: Warning Decision Support System B Integrated Information (WDSS-II). Part I: Multiple-sensor severe weather applications development at NSSL during 2002. *19th Intl. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, Long Beach, CA, Amer. Meteor. Soc., 14.8, https://ams.confex.com/ams/annual2003/techprogram/paper_56700.htm.
——, and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630, https://doi.org/10.1175/BAMS-D-14-00173.1.
Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499, https://doi.org/10.1175/2009BAMS2795.1.
Stumpf, G. J., and A. Gerard, 2021: National Weather Service severe weather warnings as threats-in-motion. *Wea. Forecasting*, **36**, 627–643, https://doi.org/10.1175/WAF-D-20-0159.1.
——, T. M. Smith, and C. Thomas, 2003a: The National Severe Storms Laboratory's contribution to severe weather warning improvement: Multiple-sensor severe weather applications. *Atmos. Res.*, **66**, 657–669, https://doi.org/10.1016/S0169-8095(03)00079-6.
——, ——, K. L. Manross, and A. E. Gerard, 2003b: Warning Decision Support System B Integrated Information (WDSS-II). Part II: Real-time test at Jackson Mississippi NWSFO. *19th Int. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, Long Beach, CA, Amer. Meteor. Soc., P1.36, https://ams.confex.com/ams/pdfpapers/56695.pdf.
——, C. D. Karstens, and L. P. Rothfusz, 2015: Probabilistic Hazard Information (PHI): Highlighting the benefits via new verification techniques for FACETs. *3rd Conf. on Weather Warnings and Communication*, Raleigh, NC, Amer. Meteor. Soc., 5.7, https://ams.confex.com/ams/43BC3WxWarn/webprogram/Paper272745.html.
——, and Coauthors, 2018: Three years of hazard services—Probabilistic hazard information (HS-PHI) experiments at the NOAA hazardous weather testbed. *29th Conf. on Severe Local Storms*, Stowe, VT, Amer. Meteor. Soc., 112, https://ams.confex.com/ams/29SLS/webprogram/Paper348359.html.
Walker, J. R., W. M. MacKenzie Jr., J. R. Mecikalski, and C. P. Jewett, 2012: An enhanced geostationary satellite-based convective initiation algorithm for 0-2 hour nowcasting with object tracking. *J. Appl. Meteor. Climatol.*, **51**, 1931–1949, https://doi.org/10.1175/JAMC-D-11-0246.1.
Wilson, K. A., P. L. Heinselman, and Z. Kang, 2016: Exploring applications of eye-tracking in operational meteorology research. *Bull. Amer. Meteor. Soc.*, **97**, 2019–2025, https://doi.org/10.1175/BAMS-D-15-00148.1.
——, ——, and C. M. Kuster, 2017a: Considerations for phased-array radar data use within the National Weather Service. *Wea. Forecasting*, **32**, 1959–1965, https://doi.org/10.1175/WAF-D-17-0084.1.
——, ——, ——, D. M. Kingfield, and Z. Kang, 2017b: Forecaster performance and workload: Does radar update time matter? *Wea. Forecasting*, **32**, 253–274, https://doi.org/10.1175/WAF-D-16-0157.1.
——, ——, and Z. Kang, 2018: Comparing forecaster eye movements during the warning decision process. *Wea. Forecasting*, **33**, 501–521, https://doi.org/10.1175/WAF-D-17-0119.1.
Yussouf, N., K. A. Wilson, S. M. Martinaitis, H. Vergara, P. L. Heinselman, and J. J. Gourley, 2020: The coupling of NSSL warn-on-forecast and FLASH systems for probabilistic flash flood prediction. *J. Hydrometeor.*, **21**, 123–141, https://doi.org/10.1175/JHM-D-19-0131.1.