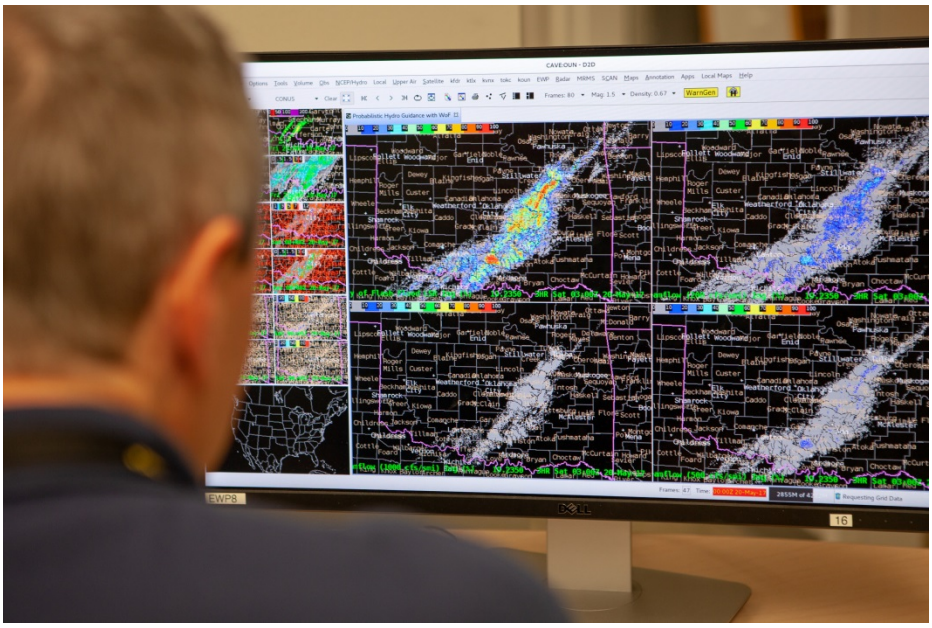
NOAA U.S. Weather Research Program (USWRP)

Hydrometeorology Testbed (HMT)
Multi-Radar Multi-Sensor (MRMS) Hydro Experiment

In Coordination with the HMT Flash Flood and Intense Rainfall (FFaIR) Experiment



-- 2018 HMT-Hydro Experiment Final Report --

June 25 – July 20, 2018

Hazardous Weather Testbed Facilities
National Weather Center
Norman, OK

Updated October 15, 2018
Version 1.1

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure 11. Objective assessment of the reliability of experimentally-issued flash flood warnings for major (blue) and minor (red) flash flood events.

Figure 12. Number of responses stating whether the participant deviated from the gridded probabilistic value in their assigned minor flash flood probability value for all experimental FFW.

Figure 13. Number of responses stating whether the participant deviated from the gridded probabilistic value in their assigned minor flash flood probability value for only verified experimental FFW.

Figure 14. Number of responses stating whether the participant deviated from the gridded probabilistic value in their assigned minor flash flood probability value for only unverified experimental FFW.

Figure 15. Number of responses stating whether the participant deviated from the gridded probabilistic value in their assigned minor flash flood probability value for experimental FFW that utilized the Probability of Receiving a Flash Flood LSR product as the primary product in the warning decision making process.

Figure 16. Number of responses stating whether the participant deviated from the gridded probabilistic value in their assigned minor flash flood probability value for experimental FFW that utilized the Probability of Exceeding Maximum Unit Streamflow Value (2 m^3 s^{-1} km^{-2}) product as the primary product in the warning decision making process.

Figure 17. Attention maps for 2100 UTC 19 May 2017 (top) and 0300 UTC 20 May 2017 (bottom). Each map represents a participating forecast and the locations that were mentioned in their data collection forms. The cool colors represent the counties where forecasters had their attention with the QPE-forced data (deterministic, probabilistic, or both). The orange color represents where forecasters had their attention with the probabilistic data driven by Warn-on-Forecast QPF. The dots represented on the map for 0300 UTC 20 May 2017 are flash flooding LSRs from the start of the archived case through 0400 UTC.

Figure 18. Location of Murray Carter Counties in south-central Oklahoma (yellow contour) highlighted on a map depicting the Probability of Receiving a Flash Flood LSR product valid at 0000 UTC 20 May 2017 showing a one-hour forecast using Warn-on-Forecast QPF for 0100 UTC 20 May 2017.

Figure 19. Timeline of all actions considered by the participating forecasters using Condition #1 for the area of Murray and Carter Counties in south-central Oklahoma from 2000 UTC 19 May 2017 to 0300 UTC 20 May 2017. Cells containing multiple colors describe multiple actions taken by the participating forecaster as specifically described by their entries in the data collection form. The column for 0200 UTC 20 May represents the period when the first flash flood LSR was reported (specifically at 0155 UTC).

Figure 20. Same as Figure 19 except for Condition #2.

Figure 21. Same as Figure 19 except for Condition #3.

I. Introduction

The National Oceanic and Atmospheric Administration (NOAA) Hydrometeorology Testbed Program (HMT) is administered by the Office of Water and Air Quality (OWAQ). The HMT promotes hydrometeorological research that will have quick and direct impact on operations within the National Weather Service (NWS), especially in regards to flash flood forecasting. The HMT provides a conceptual framework to foster collaboration between researchers and operational forecasters to test and evaluate emerging technologies and science for NWS operations. The project described herein is unique in that it addresses objectives of the HMT program while leveraging the physical facilities of the Hazardous Weather Testbed (HWT) at the National Weather Center (NWC) located in Norman, OK.

The fourth edition of the Multi-Radar Multi-Sensor (MRMS) Hydro Experiment (hereinafter denoted as "HMT-Hydro Experiment") focused on the issuance of experimental flash flood warnings for the hydrologic extreme of flash flooding during a select period of the warm season. The 2018 HMT-Hydro Experiment contained a blend of experiments with real-time data and archived case playback using prototype products and techniques. The experiment was conducted in close coordination with the sixth annual Flash Flood and Intense Rainfall (FFaIR) Experiment at the NOAA/NWS Weather Prediction Center (WPC) located in College Park, MD.

The 2018 HMT-Hydro Experiment ran for three weeks during a period from 25 June to 20 July 2018 with a one-week break during the Fourth of July holiday. Forecasters from the NWS Weather Forecast Offices (WFOs) and River Forecast Centers (RFCs) along with participation from the National Water Center in Tuscaloosa, AL worked with research scientists to assess emerging hydrometeorological technologies and products to improve the prediction, detection, and warning of flash flooding. There were two primary areas of interest with the 2018 HMT-Hydro Experiment: Use of probabilistic information to convey uncertainty of the flash flood threat and use of Warn-on-Forecast quantitative precipitation forecasts (QPFs) for short-term prediction of potential flash flooding. Various objectives of the experiment were conducted through real-time experimental warning operations and archived case studies. Each week finished up with a group discussion focusing on the probabilistic products and use of QPFs in the warning decision making process.

Researchers from the National Severe Storms Laboratory (NSSL) and the University of Oklahoma (OU) Cooperative Institute for Mesoscale Meteorological Studies (CIMMS) administered the project and the HWT provided physical space and computing resources. This report discusses the activities of the 2018 HMT-Hydro Experiment and presents findings from it with a specific emphasis on operational impacts and recommendations for future investigations.

II. Objectives

Past HMT-Hydro Experiments focused on deterministic products, including high-resolution distributed hydrologic model forecasts that operated on the flash flood time scale (Martinaitis et al. 2017). The hydrologic models examined in the past experiments were forced by MRMS radar-only quantitative precipitation estimates (QPEs). Different deterministic quantitative precipitation forecasts (QPFs) in combination with the QPE forcing for one of the hydrologic models were also evaluated.

The next evolution of hydrologic modeling and flash flood prediction will integrate probabilistic information and uncertainty into the warning decision making process. Activities within the 2018 HMT-Hydro Experiment were split between real-time operations and archived case playback. For real-time operations, forecasters focused on the decision to issue experimental flash flood warnings using probabilistic gridded information. Archived case playback for a variety of flash flood events analyzed flash flood prediction tools from a hydrologic model that were 1) forced by QPE only and 2) forced by QPE combined with ensemble QPFs.

The HMT-Hydro Experiment was conducted in collaboration with the FFaIR Experiment (Barthold et al. 2015) to simulate the real-time workflow from forecast and guidance products in the 6–24 h timeframe from WPC to experimental flash flood warnings issued in the 0–6 h timeframe. For the days utilizing real-time experimental warning operations, the HMT-Hydro Experiment team acted as a "virtual, floating forecast office" to shift its area or responsibility to where heavy precipitation events and subsequent flash flooding was anticipated to occur. The participating forecasters had the ability to issue products for any county warning area (CWA) in the CONUS.

The primary scientific goals of the 2018 HMT-Hydro Experiment were as follows:

- Evaluate the relative skill of experimental probabilistic flash flood monitoring and short-term predictive tools.
- Determine the potential benefits/limitations of utilizing precipitation forecasts for flash flood prediction and warning decision making. This will be conducted using the NSSL Warn-on-Forecast ensemble QPFs in archived case studies and OU/CAPS ensemble QPFs and HRRR QPFs in real-time operations.
- Assess the utility and perceived skill of experimental flash flood warnings that communicate the uncertainty and magnitude of the flash flood threat.
- Enhance cross-testbed collaboration and coordination as well as the collaboration between the operational forecasting, research, and academic communities on the forecast challenges associated with short-term flash flood forecasting.
- Identify how the use of probabilistic information can advance the science and societal impacts of conveying the threat of flash flooding within the Forecasting a Continuum of Environmental Threats (FACETs) paradigm.

III. Experiment Design and Activities

The HMT-Hydro Experiment ran Monday through Friday for three weeks from 25 June to 20 July 2018 with a break taken during the week of 2 July. The physical location of the experiment was in the Hazardous Weather Testbed (HWT) on the second floor of the National Weather Center (NWC) in Norman, OK. A total of ten participants from a variety of NWS offices and the National Water Center contributed to the operational and evaluation activities of the HMT-Hydro Experiment. See Appendix A for the list of participants and the officers from OU/CIMMS and NSSL that conducted the experiment.

The mix of real-time experiment warning operations and archived case studies varied the day-to-day running of the HMT-Hydro Experiment. The period from Monday through Thursday contained two days of real-time operations and two days of case studies. The weather and daily briefings from the FFaIR Experiment dictated what activities will be conducted on each day, and more importantly, which days would have the focus on real-time experimental warning operations. The daily schedules for each week are provided in Appendix B.

Experiment Introduction and Training

Participants underwent an application and selection process under the aegis of the HWT in the months prior to the commencement of the experiment. NWS service hydrologists and forecasters expressing interest in storm-scale hydrology and in related scientific research received preference. Prior to their arrival in Norman, participants were given general information about the principal scientific goals of the experiment, including a copy of the operations plan and links to associated training material. Participants were not officially exposed to any experimental products or tools until the Monday afternoon session.

The introduction session on Monday morning featured a series of presentations given by the HMT-Hydro Experiments principal investigators (PIs) and officers. The presentations focused on the following topics:

- A reiteration of the scientific goals along with a history and past-results from the HMT-Hydro Experiment
- Detailed descriptions and usage examples of all products available for use in both the real-time operations and archived case studies
- Description of the experimental design and expectations of the use of Warn-on-Forecast with the archived case studies
- Use of AWIPS and Hazard Services in the HMT-Hydro Experiment

The introduction session was modified after the first week to move the Warn-on-Forecast archived case study presentation to just prior to the first case being used instead of it being given only on Monday morning. This was a result of the potential of the archived cases not being used until Wednesday; thus, this allowed the training on the cases to be fresh for the forecasters prior to starting the first case.

FFaIR Daily Briefings

The daily weather briefing given at 12:30PM CDT was directed by the FFaIR Experiment in College Park, MD. HMT-Hydro Experiment participants and officers joined the briefing in the HWT using screen-sharing software, which were projected on to the large situational displays in the HWT. The primary goals of these briefings were to:

- Conduct a post-mortem on experimental products issued the prior day,
- Provide present synopsis of rainfall and flooding for situational awareness, and
- Summarize model-based forecasts of heavy rainfall and guidance for probabilistic flash flooding for the day.

The briefing provided by the FFaIR Experiment along with the follow-up question-and-answer session usually lasted approximately 30 minutes but sometimes lasted 45–60 minutes. Each briefing stepped through various experimental products focused on the 6-24 hour forecasting of flash flooding. The final products derived by the FFaIR Experiment were the Excessive Rainfall Outlook (ERO) valid through 1200 UTC the following day and the six-hour probabilistic flash flood forecast (PFFF; Figure 1). The details and products from these briefings, along with a beginning-of-the-week forecast, helped determine what days featured real-time experimental warning operations and what days focused on archived case studies.

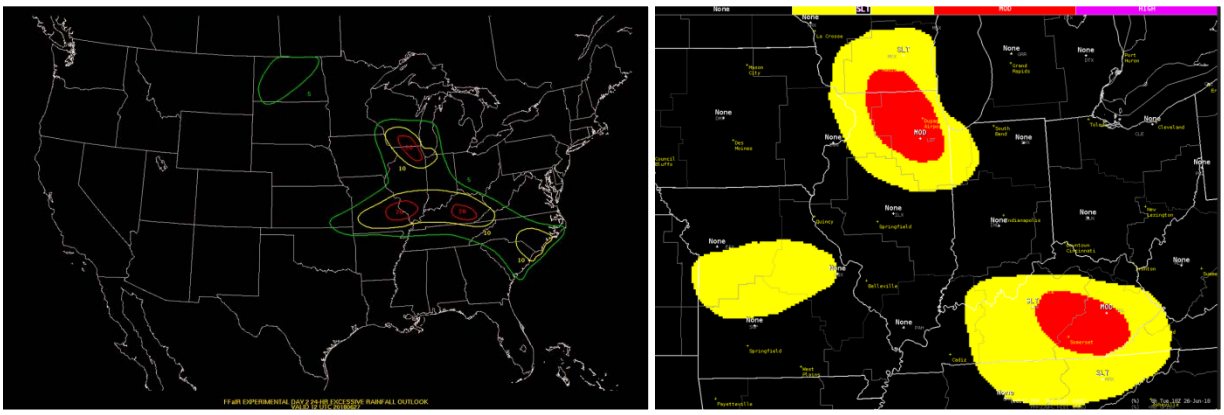


Figure 1. FFaIR ERO (left) and PFFF (right) for daily briefing given on 26 June 2018. The ERO was valid from 1500 UTC 26 June to 1200 UTC 27 June. The PFFF was valid from 1800 UTC 26 June to 0000 UTC 27 June.

FFaIR Forecast Collaboration

The HMT-Hydro Experiment collaborated with FFaIR Experiment participants in the creation of a six-hour PFFF. These occurred on the days of real-time experimental warning operations except for Mondays due to the introduction session. The collaborative process was primarily directed by the FFaIR Experiment and occurred at 11:30AM CDT. The FFaIR Experiment shared two different screens that were displayed throughout the HWT to help visualize their forecast process (Figure 1). The details from the collaboration helped set the foundation for the FFaIR daily briefing that followed and the forecast area(s) for real-time experimental warning operations.

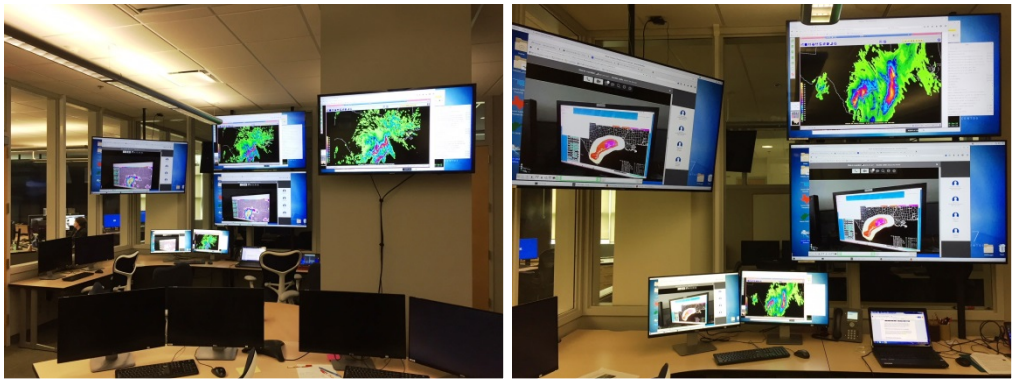


Figure 2. Demonstration of display capabilities of FFaIR forecast collaboration using the various screens implemented throughout the HWT.

Experimental Warning Operations

The start of experimental warning operations depended on the daily schedule. The focus region(s) for product issuance initially corresponded to the FFaIR Experiment guidance and current observations. Participating forecasters primarily used products with the Flooded Locations and Simulated Hydrographs (FLASH) system with radar reflectivity and QPEs provided by MRMS. Some hydrologic model outputs were also forced by different model QPFs.

The experimental warning operations design was intended to mimic the responsibilities of a local forecast office, but with the ability to change to any county warning area in the testbed. In the event of multiple flash flooding events occurring in separate regions of the CONUS, the HMT-Hydro Experiment officers prioritized the operations to the domain(s) with the anticipated biggest impacts and perhaps population density (in order to obtain dense reports). Participating forecasters were allowed to work the separate regions or collaborate in a specific region.

The experimental warnings differed from those issued in operations in that they included two probability of occurrence corresponding to flash flooding magnitudes. The Hazard Services software was used to issue experimental products. The Hazard Services GUI was tailored to solicit information from the forecaster regarding the decision-making process and the products used for the issuance of each experimental warning. The participant responses were then analyzed to see what products and probabilities influenced the warning decision. More information regarding the use of the Hazard Services software in the HMT-Hydro Experiment are detailed in the Experiment Datasets section.

Subjective Evaluation Sessions

Experimental product and warning evaluation sessions followed up real-time experimental warning operations. This usually occurred on the following day; however, deviations in the operational schedule due to current weather conditions meant that evaluations were rescheduled to more appropriate periods. One of the HMT-Hydro Experiment officers guided the participants through a series of questions related to the experimental products

and the associated experimental FFWs issued for verified flash flood events. Each participant had an equal vote in the evaluation process through the use of TurningPoint™ software and individual clickers used to collect, display, and archive forecaster responses for each question. A discussion followed each question with comments captured by HMT-Hydro Experiment officers.

The subjective evaluation sessions were broken into two sections: Experimental products and experimental FFWs. The first session on experimental products focused on a single flash flood event that occurred during real-time warning operations. If no flash flood event occurred or if a more significant flash flood event occurred outside of the experimental operational hours but within 24-hours of the evaluation period, it could be utilized during this session. The experimental product evaluation focused on the following:

- FLASH CREST Maximum Unit Streamflow
- Probability of Receiving a Flash Flood LSR
- Probability of Exceeding Maximum Unit Streamflow Value (2 $m^3 s^{-1} km^{-2}$)
- Probability of Exceeding Maximum Unit Streamflow Value (5 $m^3 s^{-1} km^{-2}$)
- Probability of Exceeding Maximum Unit Streamflow Value (10 $m^3 s^{-1} km^{-2}$)

For each of the five evaluated products, the participants were given the following two statements to subjectively rank the product performance: 1) Using all available flash flood observations, supply an approximate weight quantifying how the [product] depicted the spatial extent of flash flooding, and 2) Using all available flash flood observations, rate how the [product] depicted the magnitude of flash flooding. The subjective analysis of the spatial coverage of each product compared to verified flash flooding was conducted on a scale from 0–100 in ten-point intervals. The subjective analysis of each product depicted the magnitude of the verified flash flooding was on a seven-point scale from "Much Too Low" to "Much Too High." The deterministic FLASH CREST Maximum Unit Streamflow product is operational in the NWS and was used as the control product for probabilistic product comparisons.

Evaluation questions were prepared for the Probability of Receiving a Flash Flood LSR products forced by the HRRR and OU/CAPS QPFs; however, challenges with properly displaying the various forecast hours for each run of each model prevented proper analysis of the forecast models.

The second session focused on the experimental FFWs issued by the participants. Only experimental FFWs where flash flooding occurred within the polygon and within the valid time were used in this analysis. The participants were given the following three statements to subjectively evaluate the experimental FFW performance: 1) Using all available flash flood observations, rate the spatial accuracy of the experimental FFW vs. FFW(s) issued operationally, 2) Using all available flash flood observations and tools, rate the probability of minor flash flooding that was given in the experimental flash flood warning, and 3) Using all available flash flood observations and tools, rate the probability of major flash flooding that was given in the experimental flash flood warning. Up to four experimental FFWs could be evaluated in a single evaluation session.

Archived Case Studies

Two days during the testbed week were dedicated to the evaluation of past flash flood events and the potential impacts of including Warn-on-Forecast QPFs into the flash flood prediction process. Three cases were utilized throughout the three weeks of the experiment:

- Falls Creek, OK Event (2000 UTC 19 May 2017 to 0300 UTC 20 May 2017)
- Hurricane Harvey (0100 UTC 26 August 2017 to 0700 UTC 26 August 2017)
- El Reno/OKC Event (2300 UTC 31 May 2013 to 0100 UTC 1 June 2013)

The Falls Creek, OK and Hurricane Harvey events were utilized all three weeks of the HMT-Hydro Experiment. The El Reno/OKC event was utilized in the second and third weeks of the experiment. The Hurricane Harvey case ran until 1500 UTC 26 August 2017 during the first week of the experiment, which was reduced to 0700 UTC for the second and third weeks. More information regarding the change in length of this case is detailed in the Analysis and Recommendations section. Another case was developed for the experiment but was not utilized due to data quality issues.

Three data conditions were developed for each case and were made available to the forecasters using built-in procedures in the playback of the case through the Weather Event Simulator (WES) in AWIPS-II. These three conditions are the following:

- Condition #1: Deterministic Products (QPE-Only Forcing)
- Condition #2: Probabilistic Products (QPE-Only Forcing)
- Condition #3: Probabilistic Products (QPE + Warn-on-Forecast QPF Forcing)

Forecasters evaluated each condition in the order presented above. The data is paused at the top of each hour, allowing the participants to analyze the data without it playing in a displayed real-time mode. A data collection form was provided for each case for the participants to fill out. For Conditions #1 and #2, the following prompts were provided to the participants:

- What is your current understanding of the flash flood threat? What information did you extract from these products? Use county names and/or cities to identify the different threat areas.
- Would you have taken any action (by issuing an advisory, warning, follow-up statement, etc.) at this time based on these products? Use county names and/or cities to identify the different threat areas.

When moving to Condition #3, the participants were able to move through a three-hour forecast of probabilistic information based on the Warn-on-Forecast QPF input. The data collection form included a time line for that three-hour forecast period in ten-minute intervals. The participant was then given the following prompt to fill out the three-hour timeline:

- Going through the probabilistic guidance with the Warn-on-Forecast forcing, add information to the timeline that answers the following for any and all areas in the study domain. Please focus on the forecast times that are of most interest to you (i.e., you do not have to fill out information for every 10-minute segment). Use county names and/or cities to identify the different potential threat areas.
    - What is your understanding of the flash flood threat using the probabilistic products?
    - What information did you extract from these products?
    - Would you have taken action (by issuing an advisory, warning, follow-up statement, etc.) at this time and/or do you anticipate taking action at a future forecast time based on these products?

Once the forecaster completed the questions for that hour, he/she would advance the time to the top of the next hour and repeat the process.

Group Discussion

A group discussion was held on Friday morning to garner feedback on the activities that occurred during the week, with an emphasis on the archived cases and the Warn-on-Forecast use. HMT-Hydro Experiment officers and PIs presented focused questions to guide the discussion to gain insight and feedback on the probabilistic grids and the use of short-term QPFs in the flash flood prediction and warning process.

Feedback Survey

At the end of the week, participants filled out an online feedback survey. The feedback received within this survey was useful in determining what was working during the experiment and what changes could be made either between weeks or for future HMT-Hydro experiments.

IV. Experiment Datasets

This section details the various gridded fields and observations that were utilized throughout each week of the 2018 HMT-Hydro Experiment and the associated products issued by the forecasters during each week.

Forecast Tools

HMT-Hydro Experiment participants had access to a range of NWS operational products and experimental products during real-time experimental warning operations. The suite of NWS operational products is similar to that available via satellite broadcast network (SBN) from the National Centers for Environmental Prediction (NCEP). This included surface observations and precipitation gauges, observed soundings from the NWS upper-air network of rawinsondes, full Multi-Radar Multi-Sensor (MRMS) product suite, forecast guidance from various numerical and convective-allowing models, and satellite data. The satellite data included base and derived products from the recently implemented GOES-16 (Geostationary Operational Environmental Satellite) satellite. Up to ten individual WSR-88D radars were made available during the operational period. The radars were chosen in regions defined by current or anticipated heavy rainfall and flash flooding.

In contrast, because of the limited available WSR-88Ds that could be turned on and the ability to work anywhere in the CONUS, the Flash Flood Monitoring and Prediction (FFMP) and Four-dimensional Stormcell Investigator (FSI) software were not available. Local numerical and convective-allowing models were also not available. Operational products, including FFWs were not viewable; moreover, the forecasters were asked not to look at NWS web pages and other applications that would display operational NWS products.

A separate menu was dedicated to the HMT-Hydro Experiment that included select operational MRMS (Zhang et al. 2016) and Flooded Locations and Simulated Hydrographs (FLASH; Gourley et al. 2017) products for use during operations. A list of available products used in both the real-time operations and archived case studies is provided in Appendix C. The deterministic products focused on the MRMS precipitation products, more specifically the radar-only QPE, the seamless hybrid scan reflectivity, Radar Quality Index (RQI; Zhang et al. 2012), and surface precipitation type. This allowed for participants to have high spatio-temporal resolution radar reflectivities and QPEs across the CONUS to compensate for the limited number of WSR-88D data feeds into the HWT.

There were two sections of the deterministic FLASH product suite provided to participants: Hydrologic model output and QPE comparison products. The featured hydrologic model in FLASH is the Coupled Routing and Excess Storage (CREST) model. It is a fully distributed hydrologic model that provides maximum streamflow, maximum unit streamflow, and soil moisture products to forecasters. Similar data were also provided for the Sacramento Soil Moisture Accounting (SAC-SMA) model and a hydrophobic model. The QPE comparison products has a suite of grids creating a ratio between various MRMS radar-only QPE accumulations and flash flood guidance produced at NWS RFCs (Clark et al. 2014) via a mosaic created at WPC. Another suite of products in this section compares the MRMS

radar-only QPEs to precipitation average recurrence intervals from NOAA Atlas 14 (Perica et al. 2013). NOAA Atlas 14 analyses are not yet available for states in the Pacific Northwest or Texas. Rainfall frequencies were modeled in these regions using a multivariate regression approach, thus enabling the computation of rainfall ARI products in these states.

There were a total of four probabilistic gridded products generated from the FLASH system for the HMT-Hydro Experiment:

- Probability of Receiving a Flash Flood LSR
- Probability of Exceeding Maximum Unit Streamflow Value (2 $m^3 s^{-1} km^{-2}$)
- Probability of Exceeding Maximum Unit Streamflow Value (5 $m^3 s^{-1} km^{-2}$)
- Probability of Exceeding Maximum Unit Streamflow Value (10 $m^3 s^{-1} km^{-2}$)

The Probability of Receiving a Flash Flood LSR product corresponds to a range of unit streamflow values that have been associated with flash flood LSRs. The probability values depicted in this product indicates how frequent the corresponding values of simulated unit streamflow have been associated with flash flood LSRs. A value of 100% means that the corresponding simulated unit streamflow value has always been associated with flash flood LSRs, while a value of 0% means that the corresponding simulated unit streamflow value has never been associated with flash flood LSRs. These probabilities are generated in a post-processing algorithm using probabilistic models that have been trained on historical data. This means that the hydrologic model is integrated in a deterministic way and the resulting output (unit streamflow) is used as input to the probabilistic model. For this product, the flash flood LSR probabilistic model is based on reports from 2005–2011, and uses a logistic regression algorithm.

The Probability of Exceeding Maximum Unit Streamflow Value products use three different thresholds to represent a different potential magnitude of the flash flood hazard (i.e., minor, moderate, and major flash flooding). The determination of these values for the different magnitudes was based on past studies and observations from previous events and HMT-Hydro Experiments. The probabilities are generated in a post-processing algorithm using probabilistic models that have been trained on historical data. This means that the hydrologic model is integrated in a deterministic way and the resulting output (unit streamflow) is used as input to the probabilistic model. The probabilistic model is based on probabilities of observed USGS unit streamflow values conditional to values of simulated unit streamflow with a bias correction based on the USGS observations. This probabilistic model was trained on archived data from 2002–2011.

All work was conducted within the AWIPS-II (Advanced Weather Interactive Processing System) software in the Display 2-Dimension (D2D) perspective. No other perspectives were utilized in the testbed environment.

Observations

Three separate sources of flash flood and flood observations were available to participants and officers during the experiment:

- Local Storms Reports (LSRs) gathered by local NWS Weather Forecast Offices (WFOs)
- Automated stream gauge measurements collected by the United States Geological Survey (USGS)
- Unsolicited public geolocated smartphone or mobile devices reports from the mPING (Meteorological Phenomena Identification Near the Ground) product run by NSSL and OU (Elmore et al. 2014)

NWS LSRs are issued during or immediately after a given hazardous weather event (Horvitz 2012). These reports include the date and time of the event, the city and county of the event, the type of event, the source of the report, and the location in decimal degrees. Flood and flash flood LSRs typically include a short description of the exact impact of the reported event in plain English.

Closely related to NWS LSRs are reports in the NWS publication *Storm Data* (MacAloney 2007). In contrast to LSRs, they can contain a range of times and a spatial range (a series of latitude/longitude points versus a single point in a LSR). *Storm Data* reports are generally correlated with LSRs, but there are situations when a flash flood event only comes to light days after an event, and thus, is absent from the LSR database but present in the *Storm Data* database.

USGS stream gauges are located on catchments of various sizes across the United States. In order to quantify for inclusion in this observation database, a flash flood event recorded at a stream gauge must exceed the NWS-defined minor flood stage for the gauged location or the USGS-defined two-year return period for the gauged location and satisfy a requirement for a quick time-of-rise (0.90 m h^{-1}) of the stage (Cosgrove 2014, personal communication). Only stream gauges with contributing drainage areas of less than 2000 km^2 were considered.

The mPING project uses the recent proliferation of GPS-enabled smart phones and other mobile devices to crowd-source surface weather conditions (Elmore et al. 2014). Users can identify the relative severity of an observed flood or flash flood using a 1-4 integer scale defined by the following:

1) River/creek overflowing or cropland/yard/basement flooding
2) Street/road flooding or closure; Vehicles stranded
3) Homes/buildings with water in them
4) Homes/buildings/vehicles swept away

NWS local storm reports were rated in a similar fashion by HMT-Hydro Experiment officers. Any report will be used to validate a minor flood, while a report rated as a 3 or 4 is

required for a major flood. The major flood category also includes personal impacts such as rescues, evacuations, injuries, and fatalities. If a flood is captured by a USGS stream gauge, then the reported flood stage can be used to validate the magnitude associated with the warning. The experimental coordinators will also examine social media and local news stations for reports that are informative to the validation process.

Issued Products

In common NWS parlance, "product" refers to a text message disseminated by an operational unit of the agency. Common products include watches, warnings, and advisories. In this report, two types of products are considered: operational flash flood warnings and experimental flash flood warnings.

Operational flash flood warnings (FFWs) are issued for "storm-term events which require immediate action to protect life and property" (Clark 2011). Warnings are polygons that can be drawn independent of county or other political boundaries. They can be issued for multiple causative factors; however, in the context of the HMT-Hydro Experiment, flash flood events caused by heavy rainfall are of chief interest. FFWs are issued by local WFOs and therefore cannot cross County Warning Area (CWA) boundaries. They are created in these WFOs by an add-on application to AWIPS-II called WarnGen. In the HMT-Hydro Experiment, Hazard Services software was used for issuing experimental products (Figure 3). The forecaster draws a polygon with as many vertices as needed to accurately encompass the threat. Based on this polygon, the Hazard Services software determines which counties and locations should be in the warning text, produces the appropriate text, and then disseminates the warning.

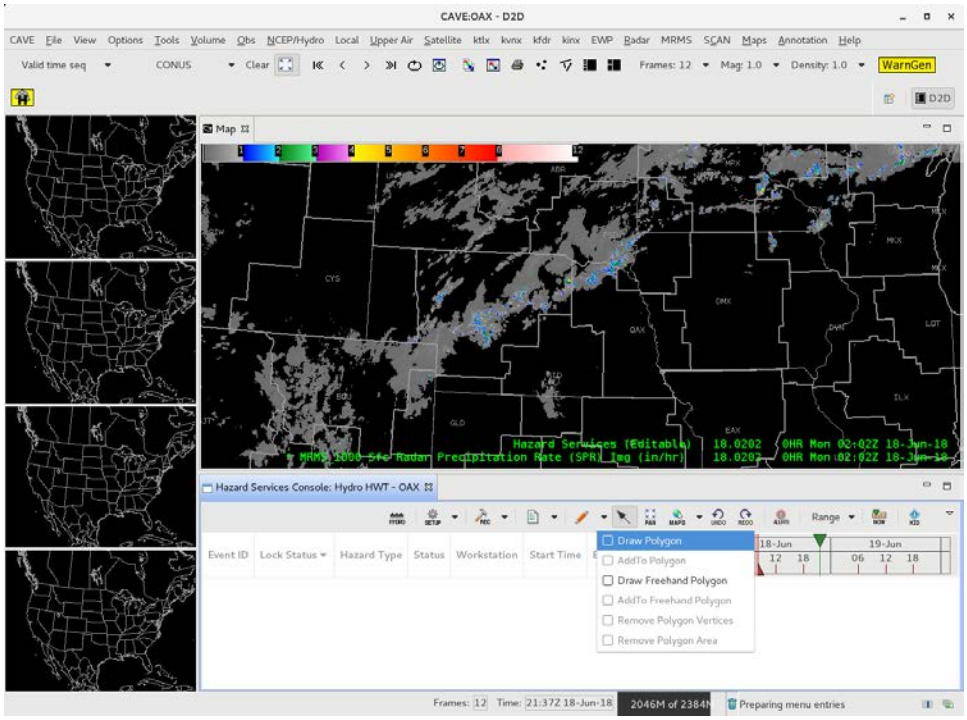


Figure 3. The Hazard Services user interface shown in the AWIPS-II system.

Experimental FFWs work similarly to their operational counterpart but with some important differences. In the testbed environment, participants were told to act as a national forecast office (i.e., the participants were responsible for monitoring conditions for and predicting potential flash flooding across the entire CONUS). In previous years of the HMT-Hydro Experiment, the Hazard Services software was modified to allow FFWs to be drawn across CWA boundaries. The version of Hazard Services in the 2018 HMT-Hydro Experiment would not allow for these capabilities. The participants were directed to use the service backup mode in the software, and the participant had to select the CWA that he/she would want to draw an experimental FFW.

The investigators used a modified version of Hazard Services that required forecasters to quantify their uncertainty about the magnitude of flash flooding expected within the warning polygon. The probability of minor flash flooding (corresponding to mPING impact classes "1" and "2") ranged from 10–100%, and the probability of major flash flooding (corresponding to mPING impact classes "3" and "4") ranged from 0–100%. The flash flood probability values were available in ten-point increments for both minor and major probabilities. Although forecasters could identify a variety of valid lengths for their experimental warnings (ranging from one to six hours), the default warning length of three hours was set in Hazard Services. A total of 57 experimental FFWs were issued during the three weeks of the 2018 HMT-Hydro Experiment.

The Hazard Services software was further modified to survey the participant about their warning decision making process (Figure 4). The following questions were presented to the participants while they are issuing an experimental FFW:

- What was the primary probability product type that prompted your decision to issue the FFW?
  - Probability of Receiving a Flash Flood LSR
  - Probability of Exceeding Maximum Unit Streamflow Value (2 $m^3 s^{-1} km^{-2}$)
  - Probability of Exceeding Maximum Unit Streamflow Value (5 $m^3 s^{-1} km^{-2}$)
  - Probability of Exceeding Maximum Unit Streamflow Value (10 $m^3 s^{-1} km^{-2}$)
- What gridded probability values prompted your decision to issue the FFW?
  - 10–100% in ten-point increments
- Is your FFW minor flash flood probability a deviation from the gridded values?
  - Yes (Higher)
  - Yes (Lower)
  - No (Same)
- Is your FFW major flash flood probability a deviation from the gridded values?
  - Yes (Higher)
  - Yes (Lower)
  - No (Same)
- If your flash flood probability value(s) deviate from the gridded product(s), please provide some reasoning why:
  - Text box for written response
- Other notes about FFW issuance:
  - Text box for written response

The participants had the option to issue follow-up statements (FFS) to their experimental FFWs. The Hazard Services software had the same questions and prompts for FFSs as those presented when issuing the initial FFW. Only four FFS were issued during the three week period of the 2018 HMT-Hydro Experiment. All FFWs and follow-up statements were archived as text products for evaluation by HMT-Hydro Experiment officers.

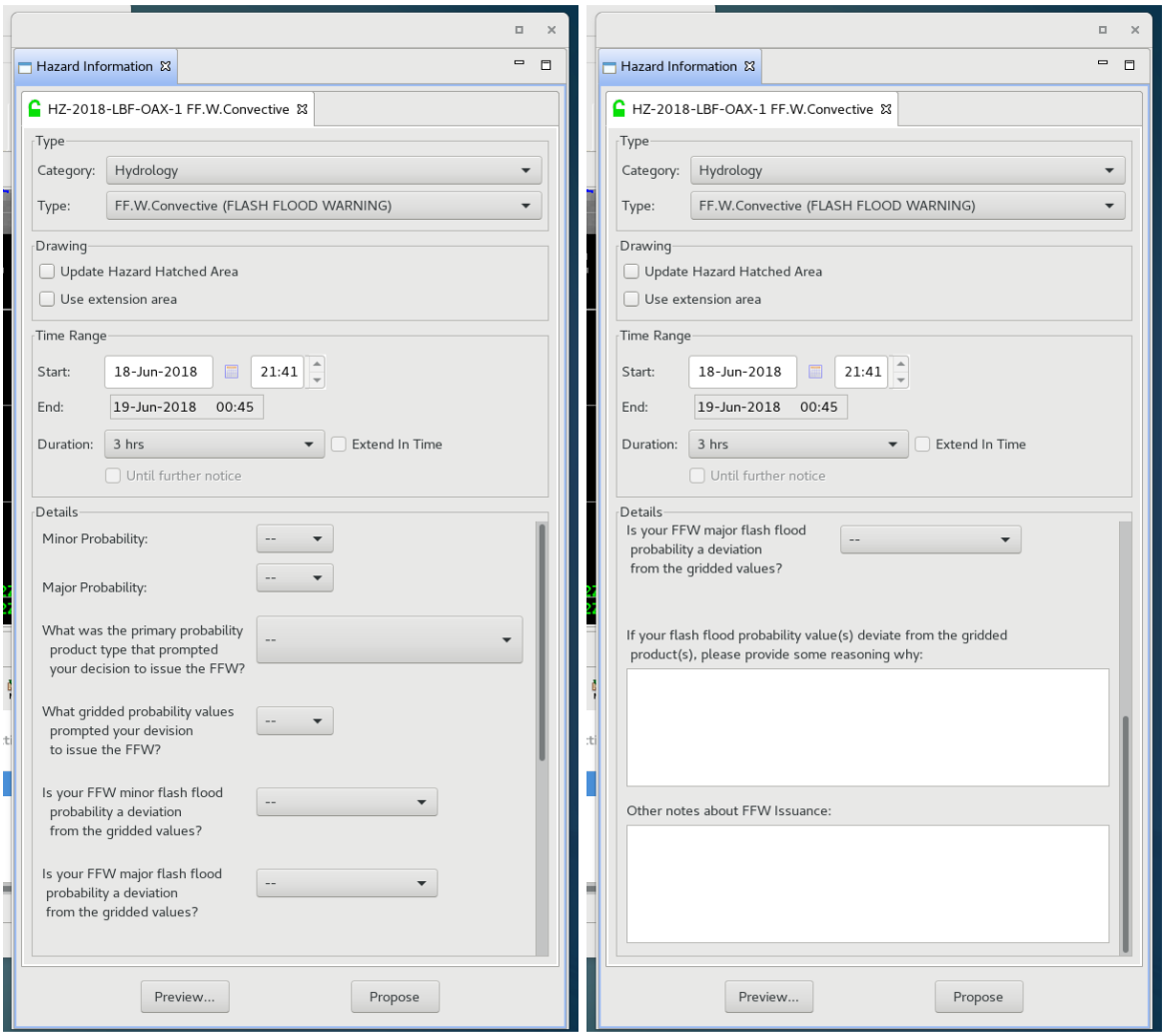


Figure 4. The Hazard Information GUI in the Hazard Services software that participants use to create experimental FFWs. This GUI was modified to survey participants about their warning decision making process using the experimental probabilistic flash flood products.

V. Results: Real-Time Operations

As described in Section III, the subjective evaluations of the real-time experimental warning operations were broken down into two sections: 1) the experimental products and 2) the subsequently issued FFWs were broken into two section. This section of the final report will look at the subjective evaluation results with some objective results as well.

Note that neither specific warning performance metrics nor statistical evaluations of the experimental probabilistic products are detailed here. These will be provided in future HMT-Hydro Experiment reports and related future publications.

Product Evaluations

A verified flash flood event was chosen after each real-time experimental warning operations. Each event was chosen based on its significance and the availability of LSRs. The event did not have to occur during the real-time experimental warning operations, but events that did occur during operations were favored. A total of eight events were analyzed during the three week period (there were additional evaluations that occurred after two real-time operation session given the evolution of the weather event, which allowed the time for the extra evaluations). Five of the eight chosen events were during real-time experimental warning operations (e.g., Figure 5). The three evaluated events that were not during the HMT-Hydro Experiment operational period were chosen for a number of reasons, including having no verified flash flooding during the real-time experimental warning operations period.

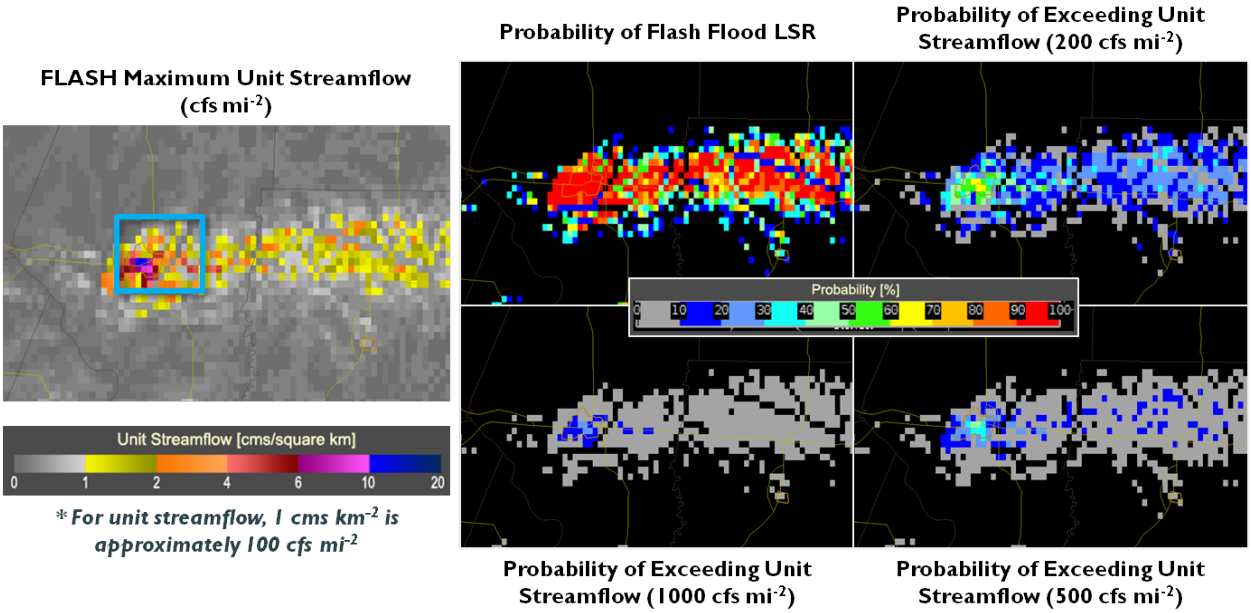


Figure 5. The Columbia, Missouri flash flood event as seen from the deterministic (left) and probabilistic (right) data at 1930 UTC 26 June 2018. The images were taken from the flash.ou.edu web page, which was used during the evaluations. Note that the units for the FLASH CREST Maximum Unit Streamflow product is in metric units ($m^3\ s^{-1}\ km^{-2}$).

Using the available observations, it was determined by HMT-Hydro Experiment officers if the flash flood event was a minor or major flash flood event. This was based on the criteria outlined in Section IV. All of the events analyzed in this section were only considered as minor flash flood events.

Participants first focused on the spatial coverage of the FLASH CREST Maximum Unit Streamflow product and the four experimental probabilistic products (Figure 6). Using the FLASH CREST Maximum Unit Streamflow product as the control data, all of the probabilistic product scores were similar to the deterministic FLASH data. The Probability of Receiving a Flash Flood LSR product was ranked the lowest amongst the four probabilistic products. This was based on the higher probability values (generally >90%) were considered as being overdone spatially. The three Probability of Exceeding Maximum Unit Streamflow Value products were ranking higher in comparison, with the product focusing on the 2 m^3 s^{-1} km^{-2} (200 ft^3 s^{-1} mi^{-2}) threshold having the lowest standard deviation value. There were greater ranges of values for the higher probability thresholds, with the product focusing on the 10 m^3 s^{-1} km^{-2} (1000 ft^3 s^{-1} mi^{-2}) threshold having the highest average value.

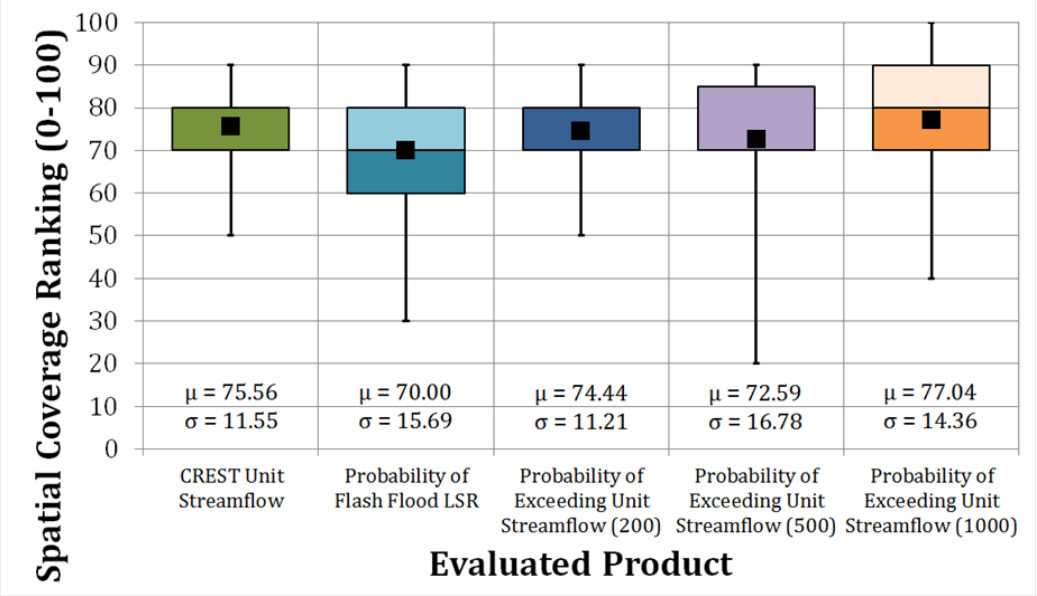


Figure 6. Subjective ranking of the spatial coverage of the FLASH CREST Maximum Unit Streamflow product and the four experimental probabilistic products when compared to verified flash flood events using a box-and-whisker plot. The top (bottom) of each box represents the 75th (25th) percentile with the line in the middle of each box representing the median subjective ranking value. The top (bottom) whisker represents the maximum (minimum) ranking. The black dot represents the mean subjective ranking. The mean (μ) and standard deviation (σ) values for each product are shown below each box-and-whisker plot.

The subjective evaluations of the probability magnitudes for the four probabilistic flash flood products had varying results. The FLASH CREST Maximum Unit Streamflow product was evaluated again as the control dataset. Most evaluations ranked this product as "About Right" with an equal distribution on either side on that response (Figure 7).

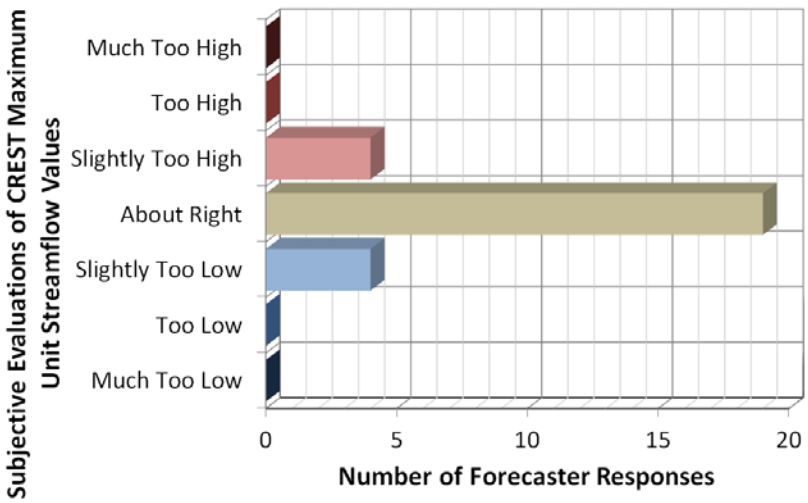


Figure 7. Subjective evaluation of the magnitude of values for the FLASH CREST Maximum Unit Streamflow product.

The evaluation the four probabilistic products yielded different results for each product (Figure 8). The magnitude of the values of the Probability of Receiving a Flash Flood LSR product was generally perceived as higher than expected. When moving to the Probability of Exceeding Maximum Unit Streamflow Value for the 2 $m^3\ s^{-1}\ km^{-2}$ (200 $ft^3\ s^{-1}\ mi^{-2}$) threshold, it was perceived as being too low. The perception became more centric with the 5 $m^3\ s^{-1}\ km^{-2}$ (500 $ft^3\ s^{-1}\ mi^{-2}$) and 10 $m^3\ s^{-1}\ km^{-2}$ (1000 $ft^3\ s^{-1}\ mi^{-2}$) thresholds.

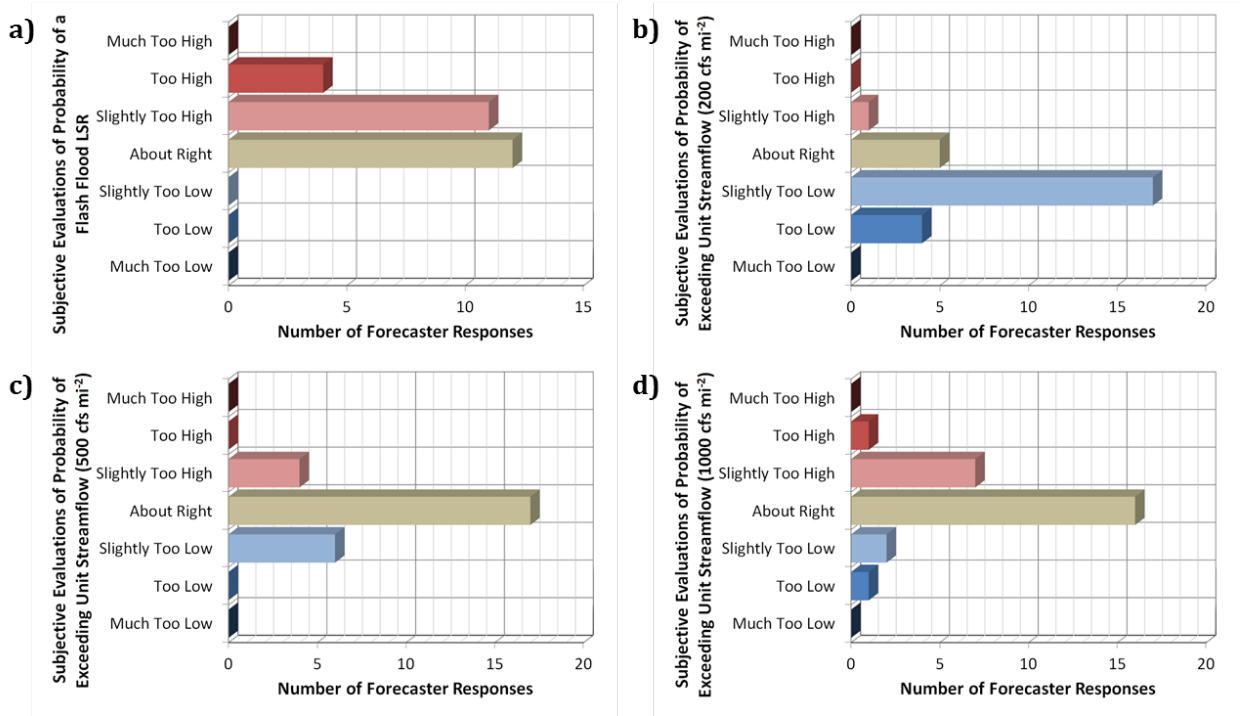


Figure 8. Subjective evaluation of the magnitude of values for the a) Probability of Receiving a Flash Flood LSR product and the Probability of Exceeding Maximum Unit Streamflow values for b) 200 $ft^3\ s^{-1}\ mi^{-2}$, c) 500 $ft^3\ s^{-1}\ mi^{-2}$, and d) 1000 $ft^3\ s^{-1}\ mi^{-2}$.

The most striking different amongst the probabilistic products were the perception of the biases in the Probability of Receiving a Flash Flood LSR and the Probability of Exceeding Maximum Unit Streamflow Value for the 2 m^3 s^{-1} km^{-2} (200 ft^3 s^{-1} mi^{-2}) threshold products. Both of these products were considered as proxies for determining the predictability of flash flooding. Discussions were had about how the probability values with the Probability of Exceeding Maximum Unit Streamflow Value for the 2 m^3 s^{-1} km^{-2} (200 ft^3 s^{-1} mi^{-2}) threshold product were viewed as too low, especially when compared to the FLASH CREST Maximum Unit Streamflow product and its deterministic values. It was concluded that the generation of the probabilities for the Probability of Exceeding Maximum Unit Streamflow Value product suite used a bias correction technique during post-processing, which determined that the deterministic unit streamflow values were too high when compare to observed unit streamflow values derived by USGS observations.

The follow-up discussions to these spatial and product magnitude evaluations yielded that the participants were using all available data and not just one preferred product to make their warning decisions. And while the attention was more focused on the Probability of Receiving a Flash Flood LSR and the Probability of Exceeding Maximum Unit Streamflow Value for the 2 m^3 s^{-1} km^{-2} (200 ft^3 s^{-1} mi^{-2}) threshold products, some forecasters noted that they were using the higher thresholds for the Probability of Exceeding Maximum Unit Streamflow Value product suite to determine the flash flood potential. More specifically, some forecasters were noting that they were using probabilistic values of 30-40% in the 5 m^3 s^{-1} km^{-2} (500 ft^3 s^{-1} mi^{-2}) threshold product to determine if a FFW was warranted. There were also instances of utilizing certain values with the 10 m^3 s^{-1} km^{-2} (1000 ft^3 s^{-1} mi^{-2}) threshold to help in the warning decision making process.

FFW Evaluations

A total of ten experimental FFWs were subjectively analyzed during the group evaluation sessions. These experimental FFWs were collocated with verified flash flooding. Three of the ten warnings did not have an associated operational FFW. For the experimental warnings that had a collocated operational FFW, there were no conclusive trends in the subjective comparisons, though more responses had a more favorable view of the experimental FFWs than the operational FFWs (Figure 9). Some comparisons considered the spatial extent of both FFWs, while acknowledging that some operational FFWs were more constrained in area due to local knowledge of the terrain and flashiness that the participants did not have. Other noted how the location of the flash flood threat within the FFW polygons (more centered within the polygon area versus near the polygon border).

The participants analyzed the probability values assigned within the experimental FFWs by the warning forecaster. The average minor flash flood probability value in the evaluated experimental FFWs was 68% and the average major flash flood probability was 19%. The participants evaluated the minor flash flood probabilities as being “About Right” or “Slightly Too Low” while the major flash flood probabilities had more responses as being higher than it should be (Figure 10). It should be noted that no verified flash flood events were deemed as “major” by the HMT-Hydro Experiment officers based on the aforementioned LSR criteria.

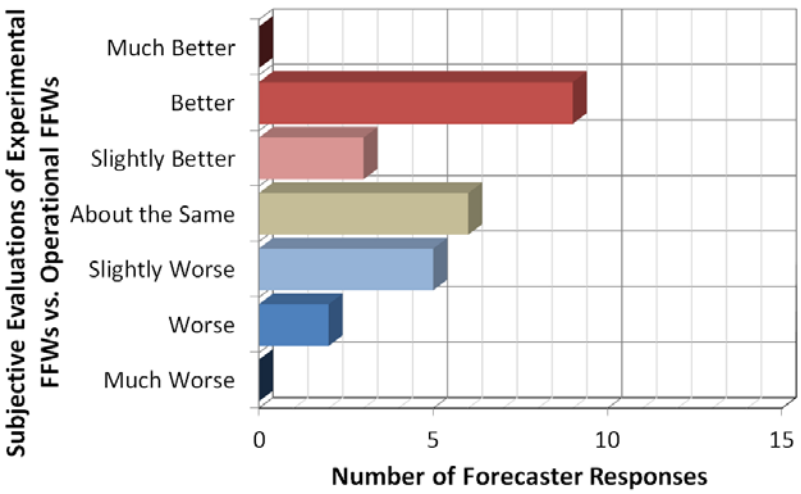


Figure 9. Subjective evaluation of the experimental FFWs when compared to collocated operational FFWs for areas with verified flash flooding.

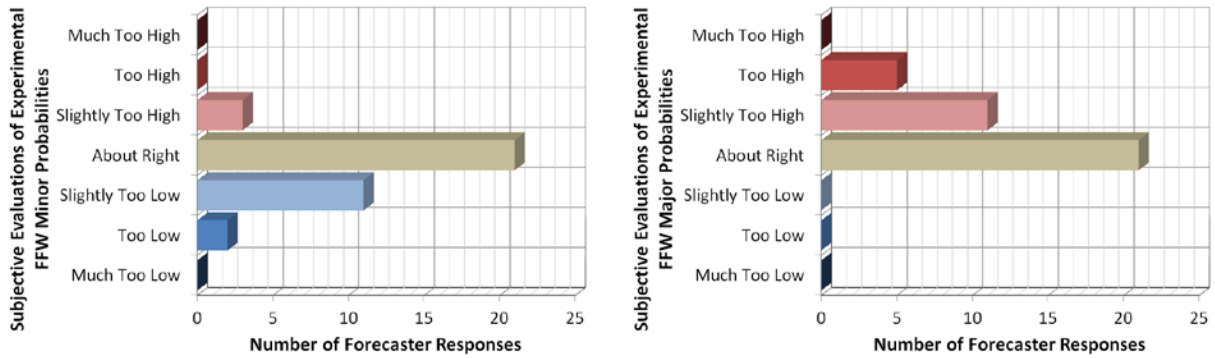


Figure 10. Subjective evaluation of the assigned minor (left) and major (right) flash flood probabilities in the experimental FFWs containing verified flash flooding.

A total of 57 experimental FFWs were issued throughout the 2018 HMT-Hydro Experiment. Fourteen of the experimental FFWs verified while 43 went unverified. The average minor (major) flash flood probabilities for the verified warnings were 73% (21%). The average minor (major) flash flood probabilities for unverified warnings were 74% (17%). The differences in the minor probabilities were negligible, while experimental FFWs that were verified did have a higher average major flash flood probability. A reliability diagram of all experimental FFWs showed that the participants generally over-predicted flash flooding with the minor probabilities except for when the probability of minor flash flood was 40% (Figure 11). Given that there were no major flash flooding during the real-time experimental warning operational periods, the reliability diagram showed the fraction of FFWs hitting to be zero.

It should be noted that the verification was conducted based on reports garnered by the local NWS offices, meaning that they would look for verification within their warning areas and not those considered by the HMT-Hydro Experiment. This would impact the verification statistics and the subsequent reliability plot.

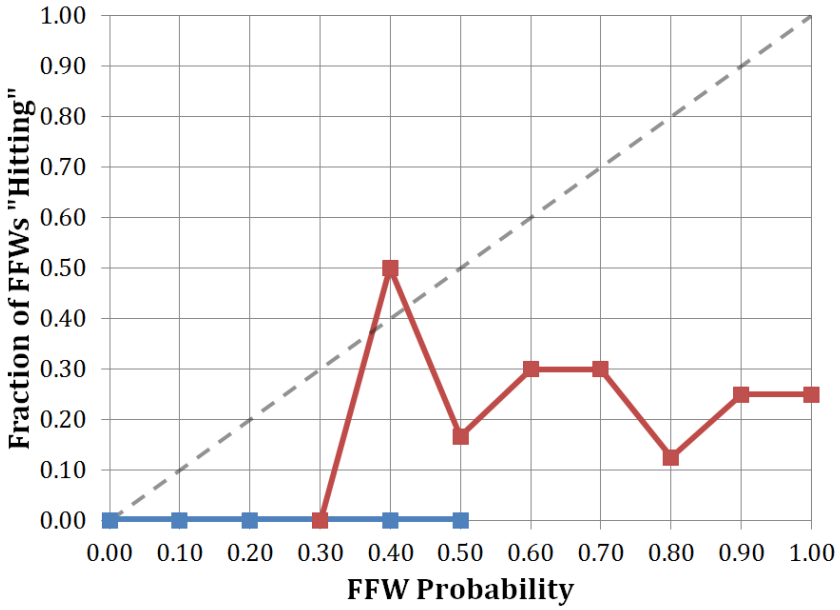


Figure 11. Objective assessment of the reliability of experimentally-issued flash flood warnings for major (blue) and minor (red) flash flood events.

Participants provided data regarding their warning decision making process in their issuance of their experimental FFWs through the modified Hazard Services software (see Section IV). When asked about what probabilistic product primarily drove the warning decision making process, 29 of the 57 warnings were based on the Probability of Exceeding Maximum Unit Streamflow Value (2 $m^3 s^{-1} km^{-2}$) product while another 25 were based on the Probability of Receiving a Flash Flood LSR product (Table 1). The other three warnings utilized the Probability of Exceeding Maximum Unit Streamflow Value (5 $m^3 s^{-1} km^{-2}$) product as their primary tool for the warning decision. When broken down by FFWs that were verified versus unverified, a greater percentage of verified warnings utilized the Probability of Exceeding Maximum Unit Streamflow Value (2 $m^3 s^{-1} km^{-2}$) product as their primary product for the warning decision (Tables 2–3).

Table 1. Distribution of the primary probabilistic product used in the warning decision making process to for all issued experimental FFWs.

Product	Count	Percent
Probability of Receiving a Flash Flood LSR	25	43.8%
Probability of Exceeding Maximum Unit Streamflow Value (2 $m^3 s^{-1} km^{-2}$)	29	50.9%
Probability of Exceeding Maximum Unit Streamflow Value (5 $m^3 s^{-1} km^{-2}$)	3	5.3%
Probability of Exceeding Maximum Unit Streamflow Value (10 $m^3 s^{-1} km^{-2}$)	0	0.0%

Table 2. Distribution of the primary probabilistic product used in the warning decision making process to for issued experimental FFWs that were verified by a flash flood LSR.

Product	Count	Percent
Probability of Receiving a Flash Flood LSR	4	28.6%
Probability of Exceeding Maximum Unit Streamflow Value (2 $m^3 s^{-1} km^{-2}$)	9	64.3%
Probability of Exceeding Maximum Unit Streamflow Value (5 $m^3 s^{-1} km^{-2}$)	1	7.1%
Probability of Exceeding Maximum Unit Streamflow Value (10 $m^3 s^{-1} km^{-2}$)	0	0.0%

Table 3. Distribution of the primary probabilistic product used in the warning decision making process to for issued experimental FFWs that were not verified by a flash flood LSR.

Product	Count	Percent
Probability of Receiving a Flash Flood LSR	21	48.8%
Probability of Exceeding Maximum Unit Streamflow Value (2 $m^3 s^{-1} km^{-2}$)	20	46.5%
Probability of Exceeding Maximum Unit Streamflow Value (5 $m^3 s^{-1} km^{-2}$)	2	4.7%
Probability of Exceeding Maximum Unit Streamflow Value (10 $m^3 s^{-1} km^{-2}$)	0	0.0%

There were slight differences with the gridded probability values that forecasters were using as guidance to issue their warnings for the two primary products. The average gridded value of the Probability of Receiving a Flash Flood LSR product for verified (unverified) experimental FFWs was 87.5% (90.5%). The average gridded value of the Probability of Exceeding Maximum Unit Streamflow Value (2 $m^3 s^{-1} km^{-2}$) product for verified (unverified) experimental FFWs was 55.6% (53.5%). How the forecasters used the gridded probabilities did vary based on the analysis of verified vs. unverified FFWs and by what primary product did the participant consider. The majority of responses (57.9%) stated that the participant modified the assigned probability values from the gridded product (Figure 12); however, the deviation directions from the gridded probabilistic products were somewhat similar for both increasing (33.3%) and decreasing (24.6%) the assigned probabilistic value in the FFW.

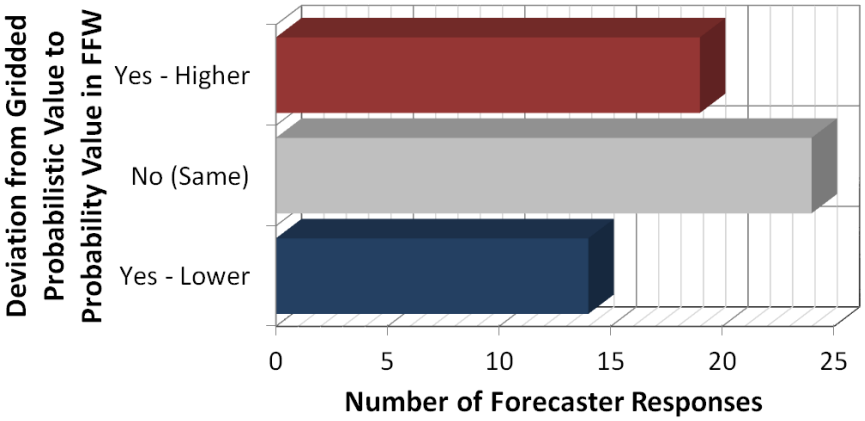


Figure 12. Number of responses stating whether the participant deviated from the gridded probabilistic value in their assigned minor flash flood probability value for all experimental FFW.

When looking at this from a perspective of verified versus unverified warnings, the majority of verified warnings (64.3%) did not have the assigned minor flash flood probability value deviate from the gridded product (Figure 13). When deviations did occur, there was more favorability to increasing the assigned minor flash flood probability. For unverified FFWs, the opposite occurred where 65.1% percent of the minor flash flood probabilities were adjusted. The distribution between the three responses ("Yes – Higher," "Yes – Lower," and "No") were nearly similar.

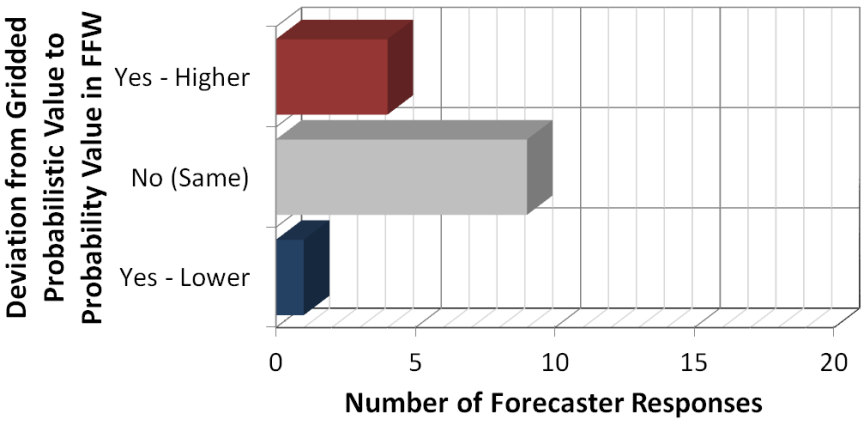


Figure 13. Number of responses stating whether the participant deviated from the gridded probabilistic value in their assigned minor flash flood probability value for only verified experimental FFW.

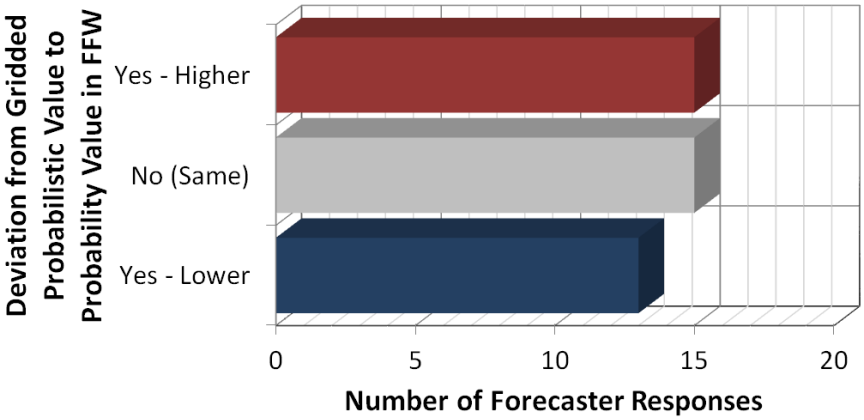


Figure 14. Number of responses stating whether the participant deviated from the gridded probabilistic value in their assigned minor flash flood probability value for only unverified experimental FFW.

Conducting a similar based on the primary probabilistic product used in the warning decision making process, a similar, yet opposite, distribution was seen in the deviation of the assigned minor flash flood probabilities. For the Probability of Receiving a Flash Flood LSR product, 44% of the warnings kept the probability provided by the gridded product while 40% of the experimental FFWs had assigned probabilities that were lower than the gridded product (Figure 15). Approximately 41.4% of the warnings kept the probability provided by the Probability of Exceeding Maximum Unit Streamflow Value (2 m^3 s^{-1} km^{-2}) product; however, 44.8% of the warnings had assigned probabilities that were higher than that gridded product (Figure 16). The human input into the assigning of probabilistic values reflected the earlier assessments on how the participants perceived the two products to be too high and too low, respectively.

The changes in the probabilistic values were also significant. The downward adjustment of the Probability of Receiving a Flash Flood LSR product for the minor flash flood probability assigned to the experimental FFWs was decreased by 22 points from 91% to 69% (Table 4). A similar increase was seen in the Probability of Exceeding Maximum Unit Streamflow Value (2 m^3 s^{-1} km^{-2}) product with an average 25.4 point increase in probability values.

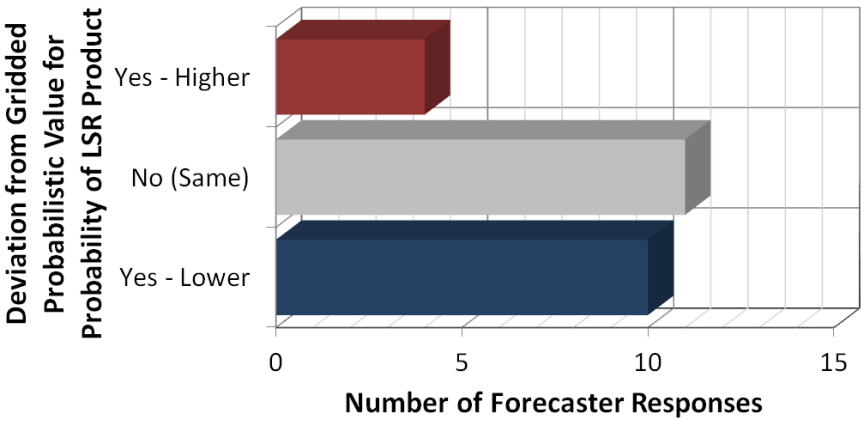


Figure 15. Number of responses stating whether the participant deviated from the gridded probabilistic value in their assigned minor flash flood probability value for experimental FFW that utilized the Probability of Receiving a Flash Flood LSR product as the primary product in the warning decision making process.

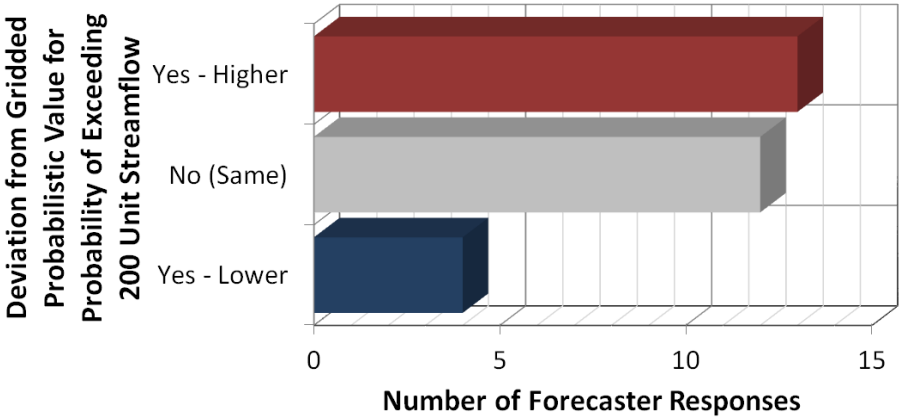


Figure 16. Number of responses stating whether the participant deviated from the gridded probabilistic value in their assigned minor flash flood probability value for experimental FFW that utilized the Probability of Exceeding Maximum Unit Streamflow Value (2 $m^3 s^{-1} km^{-2}$) product as the primary product in the warning decision making process.

Table 4. Average deviations of the gridded probabilistic values to the assigned minor flash flooding probability in experimental FFWs.

Product/Deviation	Gridded Value	Assigned Value	Change
Probability of Receiving a Flash Flood LSR			
Yes – Higher	87.5	92.5	+5.0
Yes – Lower	91.0	69.0	–22.0
Probability of Exceeding Maximum Unit Streamflow Value (2 $m^3 s^{-1} km^{-2}$)			
Yes – Higher	47.7	73.1	+25.4
Yes – Lower	52.5	45.0	–7.5

VI. Results: Archived Case Evaluations

A number of trends and decision making characteristics were found through the analysis of the participant data collection forms. This section here will focus on the analysis of the Falls Creek, OK event. A more detailed set of results will be discussed in the upcoming Warn-on-Forecast report.

The feedback provided by the forecasters going through the three conditions of data (described in Section III) allowed for a few different ways to analyze their actions. The first method was to determine where each forecaster focused their attention through the use of the three different datasets. Attention maps were created for each hour of the Falls Creek, OK case for the entire study domain (e.g., Figure 17). Initial evaluations focused on where the participants were examining the flash flood potential based on the deterministic data and the probabilistic data driven by QPE-forcing only (i.e., Conditions #1 and #2).

In general, the areas that drew attention to the forecaster that was defined by the deterministic data (Seamless Hybrid Scan Reflectivity, MRMS Radar-only QPE, QPE-to-FFG Ratio, QPE Average Recurrence Interval, and FLASH CREST Maximum Unit Streamflow) were larger than the areas using the probabilistic data. The deterministic data with QPE-only forcing drew the attention of participants to potential threat areas that covered an average of 5.1 counties for each analysis hour. An average of 3.5 counties per hour drew the attention of forecasters with the probabilistic data with QPE-only forcing.

There were several instances where the attention of the forecaster overlapped with both of the deterministic and probabilistic datasets. On average, 2.8 counties per hour garnered attention from participants with both the QPE-only forced datasets. This meant that the area that drew the attention of participants with only the deterministic data was an average of 2.3 counties per hour and 0.7 counties per hour with the probabilistic data. This signifies that the area of interest using the probabilistic data with the QPE-only forcing was more concise while introducing very few new areas that drew the attention of participants that the deterministic data did not.

The introduction of the Warn-on-Forecast QPF into the probabilistic data (i.e., Condition #3) on average expanded the area garnering the attention of participants by an additional 2.4 counties per hour. Most instances were immediately downstream of the current precipitation being observed at each hour; however, there were cases that increased probabilistic values driven by the Warn-on-Forecast QPF were not in proximity of current precipitation. Most of the instances of new counties being focused on by the participants were within two hours of the valid time of the Warn-on-Forecast model run. The primary product that caught the attention of the participants was the Probability of Receiving a Flash Flood LSR product with the Warn-on-Forecast QPF. This is likely attributed to the higher probability values that were described as “too hot” by forecasters in their subjective evaluations and subsequent group discussions; however, the trends throughout the attention maps showed that areas that were focused on with the Warn-on-Forecast data were eventually evaluated by forecasters using the QPE-only forced products.

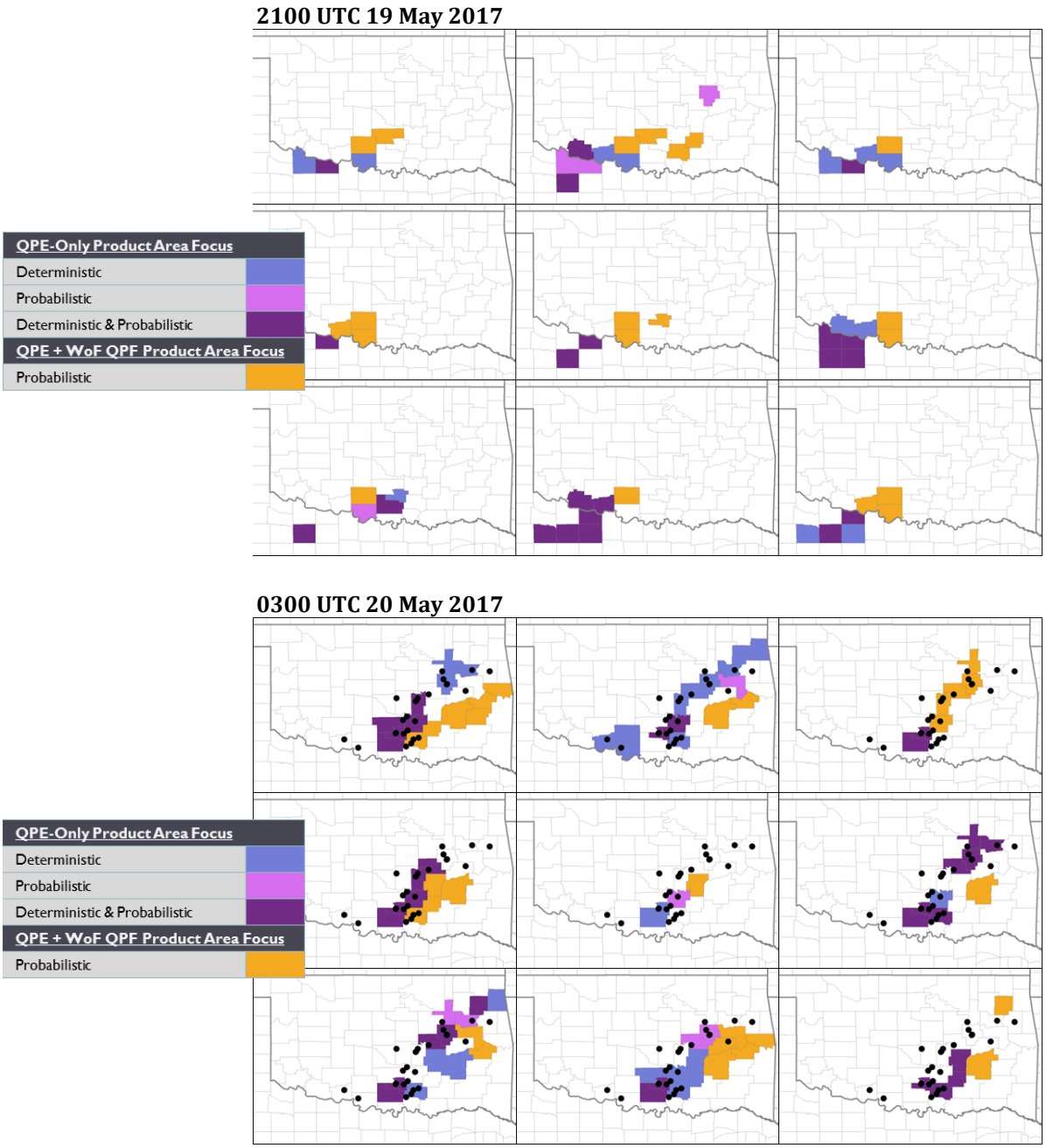


Figure 17. Attention maps for 2100 UTC 19 May 2017 (top) and 0300 UTC 20 May 2017 (bottom). Each map represents a participating forecast and the locations that were mentioned in their data collection forms. The cool colors represent the counties where forecasters had their attention with the QPE-forced data (deterministic, probabilistic, or both). The orange color represents where forecasters had their attention with the probabilistic data driven by Warn-on-Forecast QPF. The dots represented on the map for 0300 UTC 20 May 2017 are flash flooding LSRs from the start of the archived case through 0400 UTC.

Analysis also considered the change in actions by the forecasters as they moved through and analyzed the three different conditions of data. The area of interest that is demonstrated in this report focuses on the area of Murray and Carter Counties in south-central Oklahoma (Figure 18). In this location, numerous roads and state highways were flooded around the cities of Davis, Sulphur, and Pooleville, Oklahoma along with a report of water entering the basement of a local elementary school. Falls Creek rose approximately nine feet in one hour in response to the heavy rainfall, which flooded all bridges at a local, popular campground. Fortunately, all in attendance at the campground were evacuated prior to the rapid rise in height of Falls Creek.

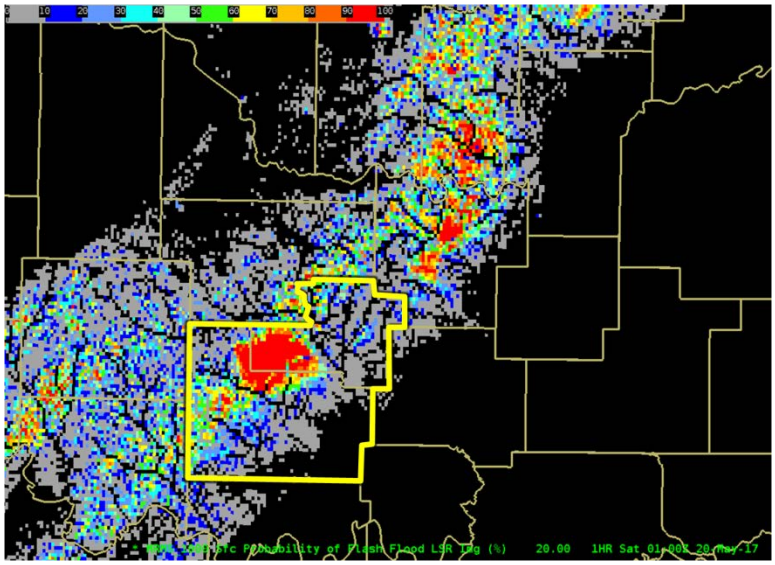


Figure 18. Location of Murray Carter Counties in south-central Oklahoma (yellow contour) highlighted on a map depicting the Probability of Receiving a Flash Flood LSR product valid at 0000 UTC 20 May 2017 showing a one-hour forecast using Warn-on-Forecast QPF for 0100 UTC 20 May 2017.

The data collection forms for each participant were analyzed to determine 1) if the participant mentioned Murray and/or Carter Counties or the cities located within these counties and 2) what actions that the participants would have taken if working the event in real life. The actions were broken down into the following categories:

- Threat assessment and/or monitoring the situation
- Communications with partners and/or end users through any medium
- Flash flood watch issuance
- Flash flood advisory issuance
- Flash flood warning issuance
- Follow-up statement for issued flash flood warning
- Consideration or use of a "Flash Flood Emergency" in an issued flash flood warning

The attention to the area of Murray and Carter Counties and any subsequent consideration of actions by the participants when using the data provided under Condition #1 were generally minimal prior to 0100 UTC 20 May 2017 (Figure 19). Four of the nine evaluated participant data collection forms showed no action for the area of interest. Of the five

participants that highlighted this region prior to 0100 UTC 20 May, three were just assessing the region and monitoring for any potential impacts. Two participants considered some form of communication at 2200 and 2300 UTC 19 May, respectively. One forecast did issue a FFW at the 0000 UTC 20 May analysis time based on an earlier wave of convection producing rain over the area and the potential for more rain with the oncoming second wave of convection. All forecasters had a FFW issued for the area of interest at the 0100 UTC 20 May analysis time, which was prior to the first flash flood LSR being reported at 0155 UTC 20 May. Six participants did consider some follow-up statement for their issued FFW with one participant considering the use of a "Flash Flood Emergency" at 0300 UTC 20 May.

When moving to Condition #2, the first finding from the evaluation of the data collection forms was that the time that each forecast decided to issue a FFW for the area of interest did not change (Figure 20). The only differences found in regards to the FFW itself were that one additional participant considered some follow-up statement with the FFW and that the participant considering the "Flash Flood Emergency" statement considered it an hour earlier (i.e., at 0200 UTC 20 May). Prior to the issuance of the FFW by the participants, there were additional instances of threat assessment, monitoring, and other actions being considered for the area of interest. Three participants described the use of a Flash Flood Advisory prior to using a FFW for Murray and Carter Counties. All uses of the Flash Flood Advisory were considered at 2200 UTC 19 May. These participants described the higher probabilistic values from the Probability of Receiving a Flash Flood LSR product as being near or at 100% in this region with the forecasted first wave of convection as a catalyst for the advisory. It should also be noted that two participants did not specifically mention the area of interest prior to 0100 UTC. It is inconclusive to presume that these participants were not monitoring this region.

Through the follow-up group discussion at the end of the HMT-Hydro Experiment (see Section VII), participants described how the probabilistic data generally reinforced their decision that they made using Condition #1 (deterministic data with QPE-only forcing). This is represented in the studies for Murray and Carter Counties where no changes were made to the FFW issuance yet some earlier Flash Flood Advisories were taken as an action. There were some instances where the participants believed that they could reduce product coverage or magnitude with the probabilistic data. One participant stated that they "may not take as much action as I would with the deterministic products."

When comparing the given actions of participants in Condition #2 and Condition #3, there are multiple significant differences in responses. Beginning with the first hour of the archived case, attention was immediately drawn to Murray and Carter Counties at 2000 UTC 19 May (Figure 21). Seven of the nine evaluated participants began some threat assessment or monitoring for this region, with one participant beginning some communications action. This was in response to the forecast of the first round of convection as provided by the Warn-on-Forecast system. The majority of the attention at 2100 UTC 19 May was focused on counties to the west of Murray and Carter Counties; thus, no new actions were described in the data collection forms.

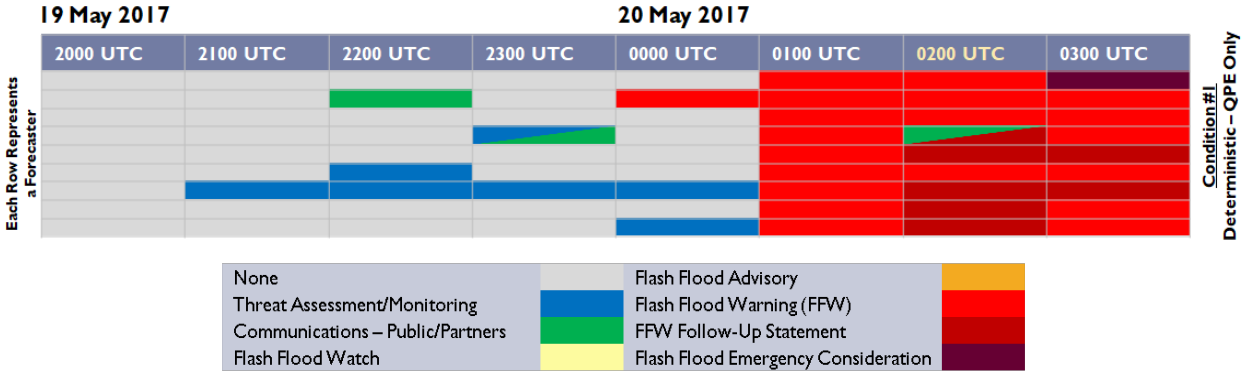


Figure 19. Timeline of all actions considered by the participating forecasters using Condition #1 for the area of Murray and Carter Counties in south-central Oklahoma from 2000 UTC 19 May 2017 to 0300 UTC 20 May 2017. Cells containing multiple colors describe multiple actions taken by the participating forecaster as specifically described by their entries in the data collection form. The column for 0200 UTC 20 May represents the period when the first flash flood LSR was reported (specifically at 0155 UTC).

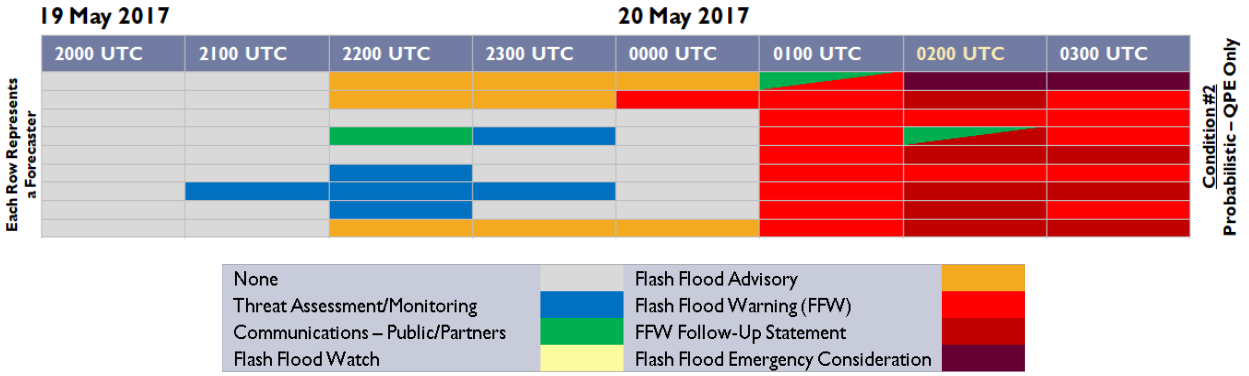


Figure 20. Same as Figure 19 except for Condition #2.

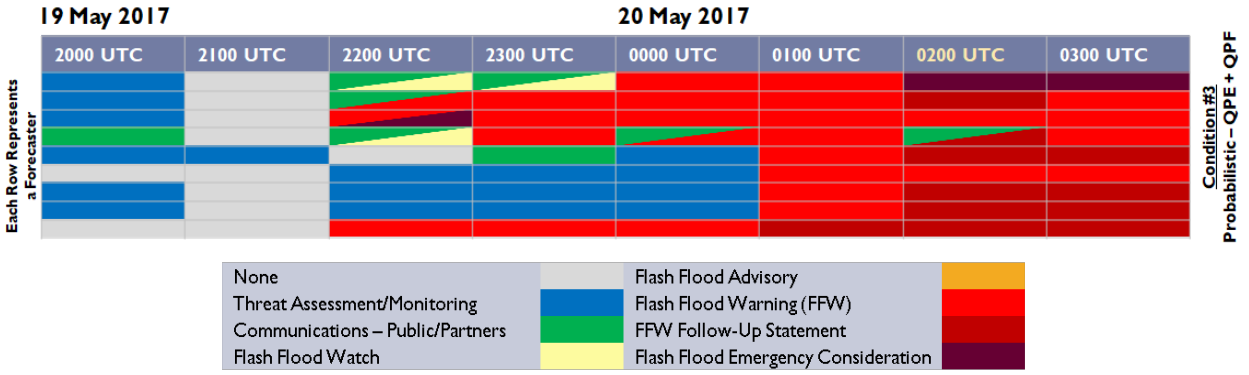


Figure 21. Same as Figure 19 except for Condition #3.

Murray and Carter Counties then received the attention of every participant at least once in 2200 UTC to 2300 UTC 19 May time frame with varying levels of action. Between 2200 and 2300 UTC, four participants called for the issuance of a FFW, with one forecaster proactively considering using the "Flash Flood Emergency" statement. Two forecasters considered the use of a short-fused Flash Flood Watch to communicate the potential flash flood threat in this region. By 0000 UTC 20 May, there were two very distinct sets of decisions portrayed by the participants. Five participants issued a FFW based on the Warn-

on-Forecast QPF-driven probabilistic data. Four participants did not issue a FFW and instead focused their attention on threat assessments with some communications. It is possible that these participants had not gained enough confidence in the Warn-on-Forecast system or were not comfortable issuing products before rain fell in the forecasted threat area. From 0100 UTC and beyond, there were little changes in action, with only two additional instances of follow-up statements for the FFWs between all of the participants.

All of the participants noted that the final hour of the probabilistic data driven by the Warn-on-Forecast QPFs appeared to be decreasing in magnitude. It was perceived that the threat of potential heavy rainfall and flash flooding might have decreased with time or that the Warn-on-Forecast system struggled with maintaining vigorous convection later in the forecast period. One likely explanation for the degraded probability values later into the Warn-on-Forecast period was the increasing in uncertainty, both spatially and in magnitude, that would limit the strength of the QPF and the subsequent probability values generated by the hydrologic model.

VII. Results: Group Discussions Summary

Approximately two hours were dedicated on each Friday to a group discussion on the probabilistic products and the short-term QPFs from the Warn-on-Forecast system. Eight questions were used to drive the conversation during these discussions. The first question asked about overall thoughts about working on the archived case studies. The results garnered during the first week differed from the other weeks due to changes in the cases and their length.

The Hurricane Harvey case was especially long in the first week, and based on participants' feedback, was reduced for the rest of the experiment. The overall feedback on the Hurricane Harvey case was mixed over the three weeks. Some participants felt that there was a lot to keep track of, resulting in it feeling overwhelming and tiring. Even with this case being shortened, participants still felt like it was repetitive after several hours. However, some participants noted that Harvey gave them an opportunity to view what a high-end event would look like in the various probabilistic products, which helped them build a baseline for subsequent events throughout the week. Comments regarding the Falls Creek, OK case focused on it being more challenging since it was not a "slam dunk" event and therefore required careful and thorough assessment of the products. The Falls Creek, OK case was also described as challenging due to the lack of geographical familiarity and not having the usual suite of data, which made normal decision making difficult. Finally, participants in the third week of the experiment found that working the archived events at the beginning of the week built confidence in regards to looking at the products during real-time operations later in the week.

When asked if one of the archived cases was more memorable, the first week participants agreed on the Davis case because the thunderstorm system was more complicated than the tropical system and therefore required them to assess all the products. Participants in the second and third weeks felt that Harvey pushed them outside of their comfort zones while also providing a good example of a high-end event. Considering what decision support services they could provide during such an event was also particularly challenging. A participant in the third week also felt that El Reno was memorable because such an event can happen anywhere and was therefore relatable.

Moving through the three conditions presented to them, participants felt that the Probability of Receiving a Flash Flood LSR was too "hot" while the suite of Probability of Exceeding Maximum Unit Streamflow Values products should be "tuned hotter" and did not appear sensitive enough. While one participant explained that forecasters learned to "calibrate the flaws [of the product] in their head" so that they can "quickly learn the signals," it was noted that improvements to the output of these products are needed to be made. Despite the skepticism of the product output, value was found when these products were used in combination with one another. This combined assessment was considered both a benefit and a challenge, since it enhanced the use of these products but also required the forecaster to process greater amounts of information; however, by using this approach, participants reported that the probability values influenced their decision making if they were previously on the fence about whether a warning was warranted or not.

Participants were asked how they envisioned the probabilistic Warn-on-Forecast condition contributing to the warning decision process and communication of flash flooding. Responses indicated that participants were in consensus that the three-hour forecast enabled them to direct their attention to threat areas that were not highlighted in the deterministic or probabilistic conditions and improved their overall situational awareness of the possible impending impacts. Some participants expected the Warn-on-Forecast guidance to increase their confidence in the geographic trend of the impact rather than the actual magnitude of the impact. Participants expected for this improved situational awareness to enhance the decision support services that they provide to both special end users and to the public. Some participants also felt that the Warn-on-Forecast guidance would support resource management, such as how many forecasters are needed on staff, whether their local NWS River Forecast Center needs to extend their operational hours for a certain event, and how workload should best be spread amongst the office personnel.

Participants were not in consensus on whether the probabilistic guidance driven by Warn-on-Forecast would dramatically change the lead time of their FFWs. One participant explained that they would not issue a warning if it was not raining at that location yet, while another expected warning lead time to vary geographically and by storm type. An example provided by one of the participants described how pulse-type thunderstorms have very short lifecycles, and therefore did not expect Warn-on-Forecast to provide much benefit to warning operations. In contrast, participants could see the benefit of Warn-on-Forecast for larger, longer-lived, and more predictable events (e.g., mesoscale convective systems).

Participants also discussed the watch product for flash flooding. Flash flood watches are generally issued at least a day ahead of an event; however, participants explained that it is sometimes very challenging to identify the threat areas with accuracy, especially during the convective season. Therefore, flash flood watches are sometimes not issued. To tackle this problem, some offices are opting to issue more targeted, short-fused flash flood watches. Participants believe that the probabilistic Warn-on-Forecast guidance will support the issuance of these short-fused watches, but it does raise questions about how responsibility for high-impact precipitation events should be shared between local NWS WFOs and the Weather Prediction Center (WPC). For watches that are issued at least one day in advance, participants felt that the Warn-on-Forecast guidance would provide them with new information that can be shared with end users closer to the event rather than having to "just regurgitate flash flood watch information" prior to FFW issuance.

The extent to which participants had trust or confidence in the products during the archived cases was also discussed. Participants listed a variety of reasons for higher confidence in the products:

- Increased exposure to the products (both probability-based products and the use of Warn-on-Forecast) throughout the week
- Considering the trends of the products rather than their actual values
- When observing a greater geographical clustering of higher probability values

- Evaluating probability products forced by Warn-on-Forecast QPFs nearer to the model initialization time (i.e., within two hours)
- When assessing convective cases rather than tropical cases
- Making flash flood watch versus warning decisions with Warn-on-Forecast

Confidence was lower after perceiving that the products were not calibrated appropriately to flash flooding threats and with not being able to verify trends with observations or verification. A participant in the second week noted that to build confidence in the output and to better comprehend the relative contributions of QPF and runoff, they wanted to see the convection itself rather than just the response. Another participant explained that learning to trust new tools is a "big deal" and will require viewing signals many times previously before feeling confident enough to "jump on them." It is likely that numerous aspects of the Warn-on-Forecast model will need to be demonstrated prior to forecasters establishing trust with the output and using it to make real-life actionable decisions. Finally, understanding how probabilities are calculated, and how they compare across products (e.g., what does a 50% mean in the Probability of Receiving a Flash Flood LSR product versus the Probability of Exceeding Maximum Unit Streamflow Values products), will be important for ensuring that forecasters can accurately interpret and apply them within their warning decision making process.

The final question of the group discussion focused on participants' opinions of the archived case experimental design. Most participants noted that they would have found it beneficial to work with another participant, so that they could interrogate the products together and maintain an open mind about what they were showing; however, many participants also acknowledged the importance of working alone for obtaining individual feedback. They also believed that working individually allowed for quicker completion of the cases than had they been working with others. Some participants felt that the cases were repetitive and suggested that future archived case sessions in the HMT-Hydro Experiment would benefit from a higher number of shorter cases that are more geographically diverse. Additionally, reduction in the repetition during data collection could be achieved if questions were rephrased to ask what is different now between the various conditions. One set of participants felt it would be beneficial to have group discussions at the end of each case rather than at the end of the week.

The list of questions and key findings from each question is provided in Appendix D.

VIII. Analysis, Other Findings, and Recommendations

A summary of each section of the HMT-Hydro Experiment is presented here, which includes some recommended actions to improve the design of future experiments. Some of the findings here stemmed from the end-of-week feedback survey. The questions and their rankings are provided in Appendix E. A summary of the written feedback from the end-of-week survey can be made available upon request.

For Probabilistic Tool Development

The suite of probabilistic products made available within the FLASH system provided utility in identifying the spatial coverage and the potential magnitude of the flash flood threat; however, some refinements are required based on the findings described within this report.

The Probability of Receiving a Flash Flood LSR product was perceived as "too hot" by the participants and was generally adjusted downward when the participants were using this particular product to assign a minor flash flood probability within a FFW. While it was one of the two more commonly used gridded probabilistic products to drive the warning decision making process, forecasters noted that because of its higher probabilities, it could be used as more of a "situational awareness" product instead. Further evaluation is still required on this product and should be derived in the same manner for the 2019-edition of the experiment.

The Probability of Exceeding Maximum Unit Streamflow Value (2 $m^3 s^{-1} km^{-2}$) product was perceived as "too cool" and was generally biased upward when the participants were using this particular product to assign a minor flash flood probability within a FFW; however, it should be noted that this product was the dominant product utilized for the warning decision of verified experimental FFWs. The higher threshold Probability of Exceeding Maximum Unit Streamflow Value products had a better perception in regards to the bias of the product; however, the lack of major flash flood events prevented a proper analysis of them. All of the Probability of Exceeding Maximum Unit Streamflow Value products were bias corrected based on comparisons of the FLASH CREST maximum unit streamflow values and those derived from USGS observations. It is likely that this bias correction reduced the probability values generated within these products. It is recommended that this particular product suite be recalculated at least without the bias correction component. This will make it consistent with the deterministic CREST products, for which most forecasters have had experience with.

The two forecast models made available during real-time experimental warning operations (HRRR and OU/CAPS) were not analyzed during the subjective evaluations. This is because the display of the products within the FLASH web site prevented proper analysis of the products. This is because each individual hour was not able to be viewed. Instead, the total accumulated forecast was seen on the page instead. In order to properly evaluate any real-time QPFs, a better methodology or viewing capabilities would have to be designed to allow for this.

For Real-Time Operations

Despite the lack of high-impact weather throughout the three weeks, the real-time experimental operations went rather smoothly. There was a learning curve with the issuance of FFWs using the Hazard Services software. It should be recommended that a demonstration should be given in the introduction and prior to the first operational period, which is similar to the archived case demonstrations that were conducted with success. Given the unpredictability of flash flooding, a more flexible schedule should be considered; however, the impacts of this with respect to the FFaIR daily briefings must be contemplated.

The use of the modified Hazard Information GUI in the Hazard Services software for FFW issuance should still be continued. The survey-style approach allowed for valuable feedback on the warning decision making process and how the experimental products have been utilized. However, the questions and available responses should be modified for the pursuit of improved feedback. This includes modified or different questions and the ability to select multiple responses.

One notable bug was found with the Hazard Services software. Participants were not able to draw warning polygons within counties that were split by NWS county warning areas. This came to light during an event in the southwestern CONUS with the inability to create a warning over San Bernardino County, California. Further testing allowed for the diagnosis of this issue, which was then reported to the software developers of the Hazard Services platform. It is recommended that this issue be fixed prior to the 2019 HMT-Hydro Experiment.

The 2015 HMT-Hydro Experiment explored the use of flash flood recommenders for an automated drawing of a first-guess warning polygon (Martinaitis et al. 2017). The experiment used a single product with a user-defined threshold to draw these recommended polygons. Given the use of probabilistic information, it is recommended that the reintroduction of the flash flood recommender should be considered with the complexity of its calculations to be determined. Similar analysis to the Martinaitis et al. (2017) should be conducted in regards to the use of fully automated versus manually drawn or edited polygons.

Since the inaugural HMT-Hydro Experiment in 2014, forecasters have issued experimental FFWs with user-assigned probabilities for the potential of minor and major flash flooding. This action continued in the 2018 HMT-Hydro Experiment. The combination of forecasters assigning minor/major flash flood probabilities with the multiple gridded probabilistic products developed within the FLASH system provided a platform to not only give the uncertainty of a potential hazard existing but the uncertainty of the severity of the potential hazard. This approach should continue to be explored with flash flooding and other storm-scale hazards and within the concepts of the FACETs paradigm.

For Archived Cases

The length of the archived cases was adjusted after the first week of the experiment, such that the 0–6 hour forecast provided by the Warn-on-Forecast system was reduced to a 0–3 hour forecast. We noticed that this change in forecast length reduced participants' workload and subsequent fatigue when working the cases. Furthermore, participants described the desire to have a larger number of shorter cases that would allow them to work greater variety of weather events (both in terms of storm morphology and geographic location). A shorter case was provided for the second and third weeks (El Reno, OK), which only had three hours of data (with each hour providing a 0–3 hour forecast). It is recommended that any new cases developed for the HMT-Hydro Experiment would have be of a duration greater than three hours but no more than 6–7 hours, with each hour providing a 0–3 hour forecasts from the Warn-on-Forecast system. The cases should also have a variety of geographic locations and storm morphology. Some domains also spanned large areas; thus, the focusing of the area of interest should also be considered.

There were multiple discussion points about the data collection process. Some participants felt that working independently was more isolating than what they would like for the archived cases, and instead suggested working in pairs and collaborating on their analyses. This topic will have to be considered by HMT-Hydro Experiment officers for future experiments, with careful consideration given for the tradeoffs between using individual versus group assessments of archived cases. Another topic that was mentioned was how challenging it was to track actions from one hour to the next, since the drawing of polygons were not considered in the experiment design. This issue will need to be explored to see whether creating polygons is feasible with the current case design. Modifications to the questions given to participants during the archived case data collection should also be reviewed to ensure they are yielding data that are relevant to the research questions posed.

To further expand upon these talking points, a more advanced data collection form could be used to better collect responses and reduce the fatigue of the cases. The current form was a modified spreadsheet with cells designated for written responses only. HMT-Hydro Experiment officers should investigate other means of conducting the data collection process, including a new medium to conduct the collection process, radio button and/or multiple-choice questions, and interactive maps to draw polygons or highlight threat areas, etc. Additionally, it should also be investigated to see if polygons can be drawn in the archived cases using the WES in a mode other than displaced real-time. Much of what has been observed and learned during this first experimentation of Warn-on-Forecast in the HMT-Hydro experiment will inform a more streamlined and efficient data collection process in subsequent experiments.

For Future HMT-Hydro Experiments

The time of the year of the HMT-Hydro Experiment allowed for close coordination with the FFaIR Experiment and avoids interfering with springtime severe convection studies by other experiments under the HWT umbrella. Running the HMT-Hydro Experiment in the summer also allows for the inclusion of monsoon-driven events in the southwestern

CONUS and the potential for impacts from tropical cyclones; however, some operational days were notably slow despite the fact that only two days per week were dedicated to real-time operations.

Given the inevitability of this and the inclusion of the Warn-on-Forecast system in the 2018-edition of the HMT-Hydro Experiment, two days of the experiment were dedicated to archived case studies. This schedule was beneficial to providing activities on the slower days, which included days during this year that had no flash flooding during the daily operation of the experiment. While the 2019 HMT-Hydro Experiment was initially written to be conducted using real-time operations only, the utilization of more archived case studies similar to the ones presented in 2018 should be highly considered.

NWS offices located in the southwestern CONUS had issued numerous operational FFWs in areas of wildfire burn scars. Modifications to the hydrologic properties in regions impacted by wildfires were not considered in the FLASH system and the 2018 HMT-Hydro Experiment; thus the HMT-Hydro Experiment participants were not able to issue experimental warnings for these specific areas. Ongoing research objectives at NSSL and other NOAA entities are focusing on implementing post-wildfire hydrologic impacts into NOAA's hydrologic modeling suite, including FLASH and the National Water Model (NWM). The coupling of the modified hydrologic properties with QPE and Warn-on-Forecast QPF forcings should be considered in future testing within HMT experiments, including the HMT-Hydro Experiment with regards to warning issuance for post-wildfire flash floods and debris flows.

Subjective evaluations conducted during the HMT-Hydro Experiment utilized a laptop containing the TurningPoint™ software and individual clickers used to collect, display, and archive forecaster responses. This laptop was borrowed from the NWS Warning Decision Training Division (WDTD). The use of the TurningPoint™ software allowed for participants to provide independent, anonymous feedback to evaluation questions and statements without other participants influencing responses (via open discussions during the scoring portion of the evaluation). The purchasing of a laptop with the TurningPoint™ software should be considered by the HWT for use in the HMT-Hydro Experiment and potentially for other experiments conducted within the HWT for evaluation purposes.

Acknowledgements

First and foremost, the officers of the HMT-Hydro Experiment would like to thank all of the participants for their hard work and contribution during the three weeks of this experiment. The insight gained throughout this process is invaluable to furthering the science and application of future products to improve flash flood prediction.

The principal investigators would like to thank everyone who helped make the 2018-edition of the HMT-Hydro Experiment a success. From the various data sets to the technical and managerial aspects of running the HWT, this would not be possible without your time and dedication to the project.

The HMT-Hydro Experiment is funded through the Hydrometeorology Testbed by the Office of Weather and Air Quality under NOAA Award Number NA17OAR4590281 for the project "Probabilistic Warn-on-Forecast System for Heavy Rainfall and Flash Flooding." Regional NWS headquarters/offices also provided some funding for certain participants.

References

Barthold, F., T. Workoff, B. Cosgrove, J. Gourley, D. Novak, and K. Mahoney, 2015: Improving flash flood forecasts: The HMT-WPC Flash Flood and Intense Rainfall Experiment. *Bull. Amer. Meteor. Soc.*, **96**, 1859–1866.

Clark, E., 2011: Weather Forecast Office hydrologic products specification. US Department of Commerce, National Weather Service, Instruction 10-922. [Available online at http://www.nws.noaa.gov/directives/sym/pd01009022curr.pdf.]

Clark, R. III, J. J. Gourley, Z. Flamig, Y. Hong, and E. Clark, 2014: CONUS-wide evaluation of National Weather Service flash flood guidance products. *Wea. Forecasting*, **29**, 377–392.

Elmore, K. L., Z. L. Flamig, V. Lakshmanan, B. T. Kaney, V. Farmer, H. D. Reeves, L. P. Rothfusz, 2014: MPING Crowd-Sourcing Weather Reports for Research. *Bulletin of the American Meteorological Society*, **95,** 1335–1342.

Gourley, J. J., and Coauthors, 2017: The FLASH project: Improving the tools for flash flood monitoring and prediction across the United States. *Bull. Amer. Meteor. Soc.*, **98**, 361–372.

Horvitz, A., 2012: Multi-purpose weather products specification. US Department of Commerce, National Weather Service, Instruction 10-517. [Available online at http://www.nws.noaa.gov/directives/sym/pd01005017curr.pdf.]

MacAloney, B., 2007: Storm Data preparation. . US Department of Commerce, National Weather Service, Instruction 10-1605. [Available online at http://www.ncdc.noaa.gov/stormevents/pd01016005curr.pdf.]

Martinaitis, S. M., and Coauthors, 2017: The HMT Multi-Radar Multi-Sensor Hydro Experiment. *Bull. Amer. Meteor. Soc.*, **98**, 347–359.

Perica, S., D. Martin, S. Pavlovic, I. Roy, M. St. Laurent, C. Trypaluk, D. Unruh, M. Yekta, and G. Bonnin, 2013, cited 2014: NOAA Atlas 14: Precipitation-frequency atlas of the United States, Volume 9, Version 2.0. US Department of Commerce, National Weather Service, Hydrologic Design Studies Center. [Available online at http://www.nws.noaa.gov/oh/hdsc/PF_documents/Atlas14_Volume9.pdf.]

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–637.

APPENDIX A: HMT-Hydro Participants and Staff

A total of ten participants were a part of the 2018 HMT-Hydro Experiment. Nine participants were from local NWS Weather Forecast Offices (WFOs) or River Forecast Centers (RFCs), while one participant was affiliated with the National Water Center in Tuscaloosa, Alabama.

Name	Affiliation	Week #
James Brotherton	NWS WFO San Diego, CA	1
Emily Carpenter	NWS WFO Tucson, AZ	1
Alana McCants	NWS West Gulf RFC (Fort Worth, TX)	1
Patrick Ayd	NWS WFO Bismarck, ND	2
Randy Bowers	NWS WFO Norman, OK	2
Peter Corrigan	NWS WFO Blacksburg, VA	2
Kate Abshire	NWS WRSB (Silver Springs, MD)	3
Whitney Flynn	NWS National Water Center (Tuscaloosa, AL)	3
Glenn Lader	NWS WFO Tucson, AZ	3
Jennifer Vogt Miller	NWS WFO Albany, NY	3

The officers of the 2018 HMT-Hydro Experiment were responsible for the facilitating of all operational activities during each week. At least one or two HMT-Hydro Experiment officers were in attendance through all daily activities, while other officers focused on the technical aspects, logistics, or evaluations of specific products during the experiment.

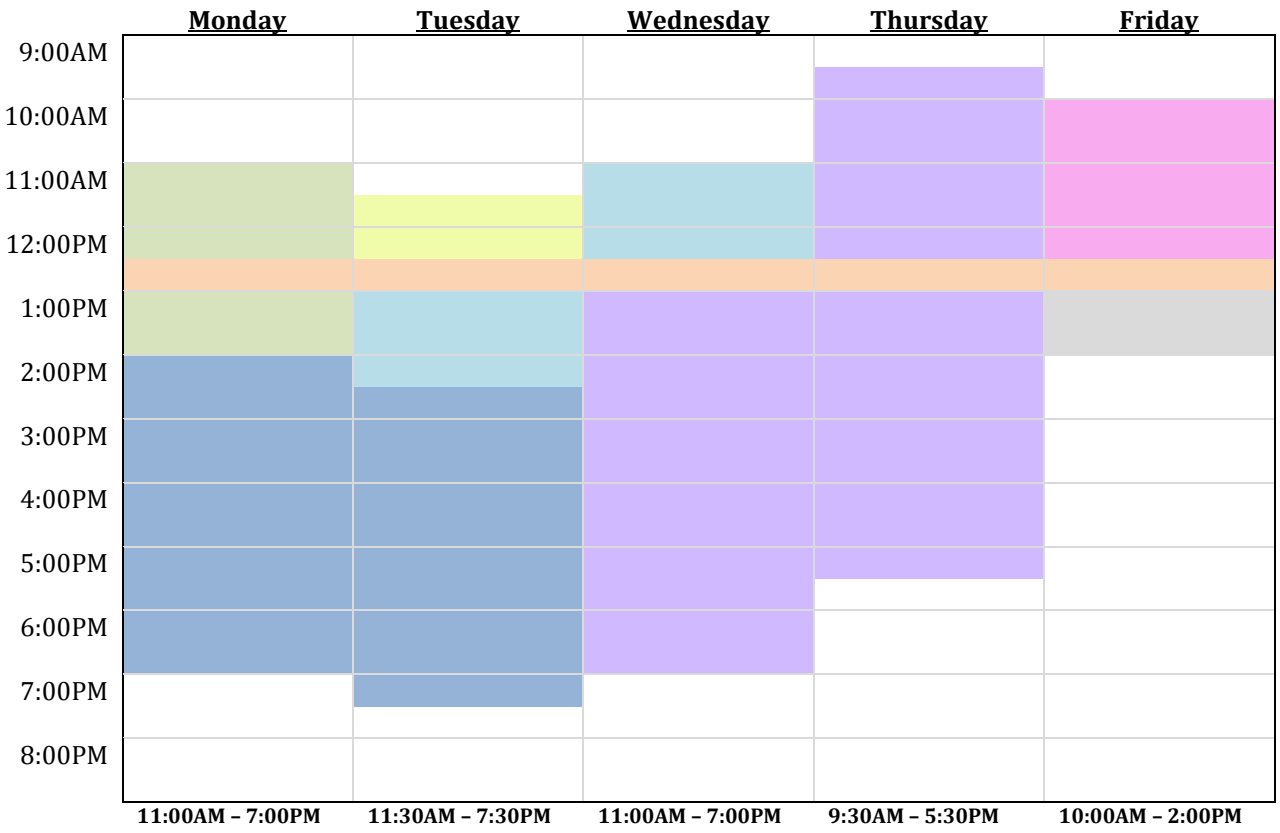
Name	Role	Affiliation
Jonathan J. Gourley	Principal Investigator	NOAA/OAR/NSSL
Steven Martinaitis	Principal Investigator	OU/CIMMS
Katie Wilson	Warn-on-Forecast Coordinator	OU/CIMMS
Nusrat Yussouf	Warn-on-Forecast Coordinator	OU/CIMMS
Pamela Heinselman	Warn-on-Forecast Coordinator	NOAA/OAR/NSSL
Humberto Vergara-Arrieta	Real-Time Operations Coordinator	OU/CIMMS
Andres Vergara-Arrieta	Real-Time Operations Coordinator	OU/CIMMS
Tiffany Meyer	HWT Information Technology Coordinator	OU/CIMMS
Kodi Berry	Executive Officer – HWT	OU/CIMMS

APPENDIX B: Weekly Schedules of HMT-Hydro Experiment

This appendix details the base schedule that was kept for each of the three active weeks of the HMT-Hydro Experiment. Minor modifications were made based on the current weather during the experiment, and are not reflected in the base schedules shown.

Week 1 Schedule (25-29 June)

Real-Time Operations: Monday, Tuesday
Archived Case Studies: Wednesday, Thursday

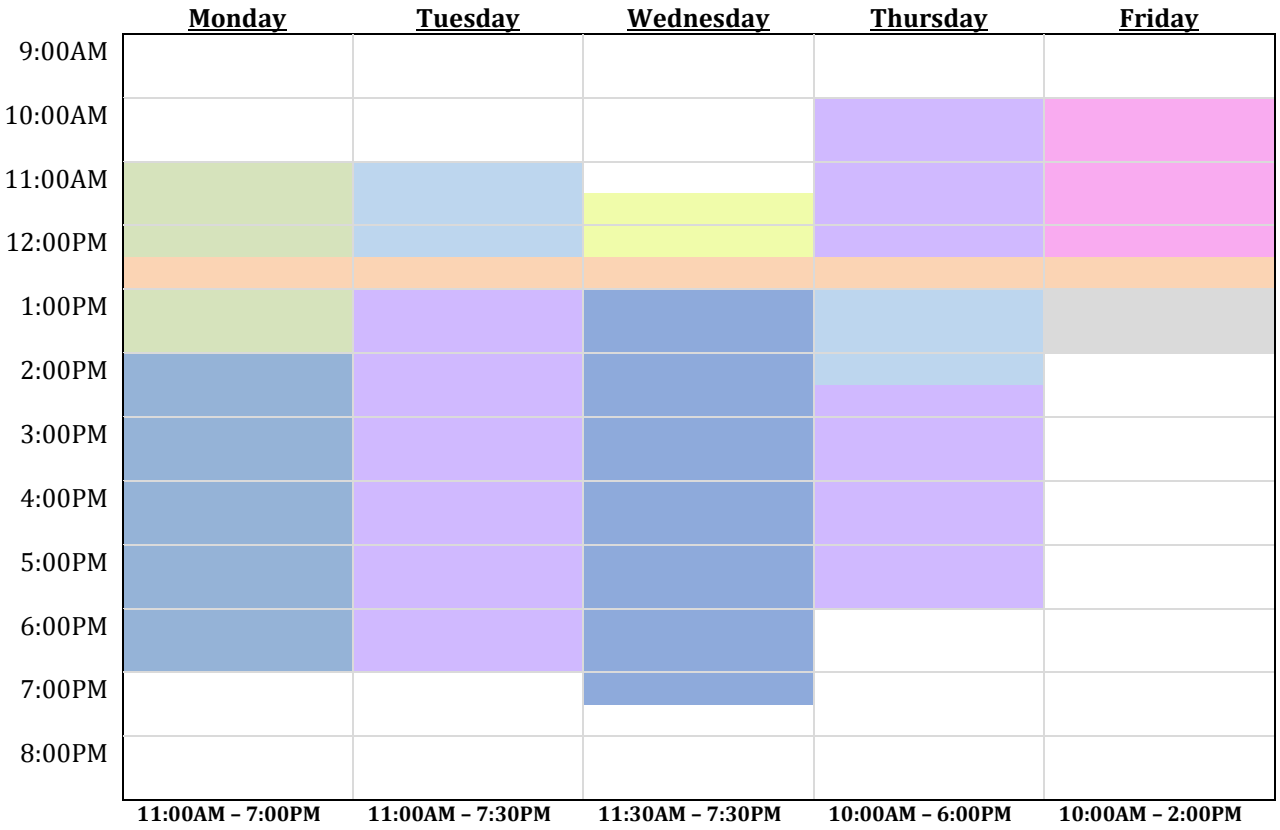


Legend:

- Introductory Remarks and Testbed/Product Training
- FFaIR Experiment Daily Weather Briefing
- Real-Time Experimental Warning Operations
- FFaIR Experiment 6-hr PFFF Collaboration
- Group Evaluation of Previous Day Real-Time Operations
- Archived Case Studies
- Group Discussion on Flash Flood Probabilities and Warn-on-Forecast QPFs
- End-of-Week Survey and Closing Remarks

Week 2 Schedule (9-13 July)

Real-Time Operations: Monday, Wednesday
Archived Case Studies: Tuesday, Thursday

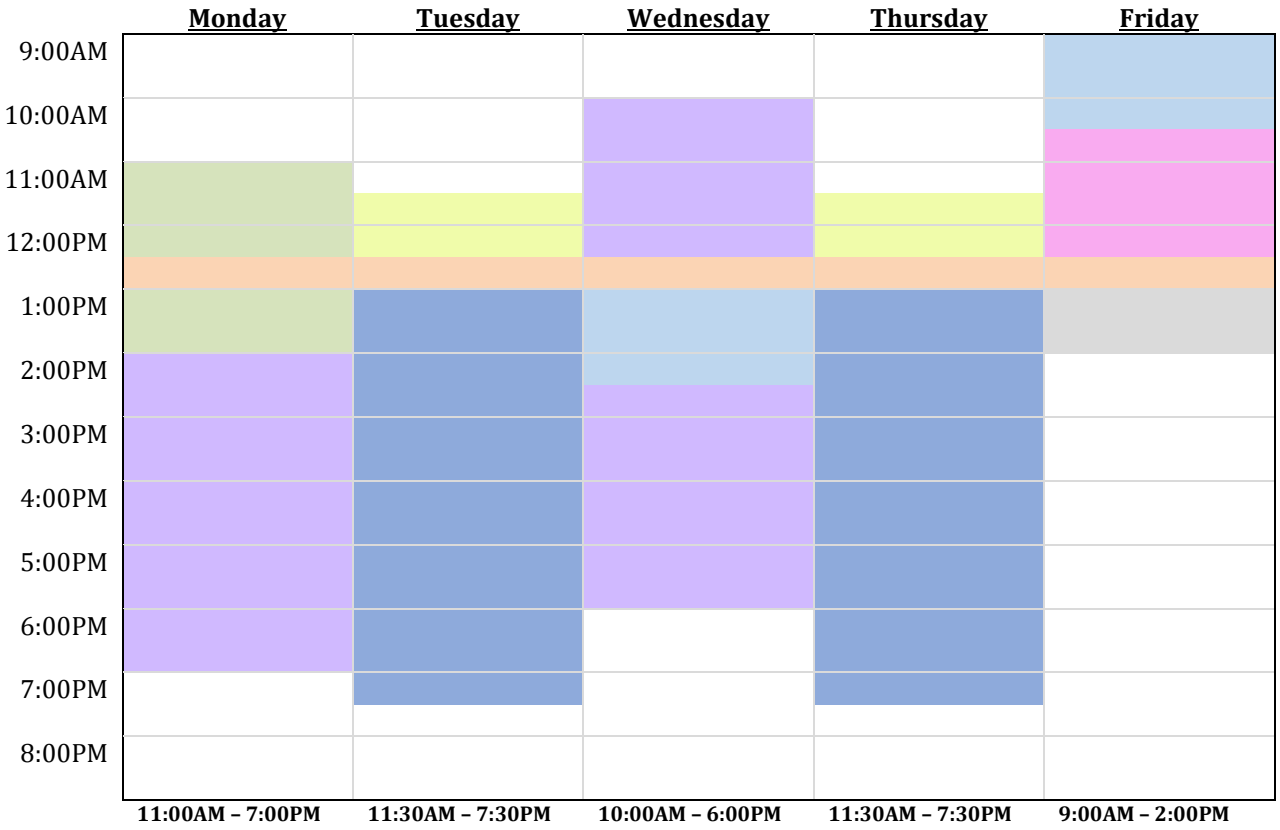


Legend:

- Introductory Remarks and Testbed/Product Training
- FFaIR Experiment Daily Weather Briefing
- Real-Time Experimental Warning Operations
- FFaIR Experiment 6-hr PFFF Collaboration
- Group Evaluation of Previous Day Real-Time Operations
- Archived Case Studies
- Group Discussion on Flash Flood Probabilities and Warn-on-Forecast QPFs
- End-of-Week Survey and Closing Remarks

Week 3 Schedule (16-20 July)

Real-Time Operations: Tuesday, Thursday
Archived Case Studies: Monday, Wednesday



Legend:

- Introductory Remarks and Testbed/Product Training
- FFaIR Experiment Daily Weather Briefing
- Real-Time Experimental Warning Operations
- FFaIR Experiment 6-hr PFFF Collaboration
- Group Evaluation of Previous Day Real-Time Operations
- Archived Case Studies
- Group Discussion on Flash Flood Probabilities and Warn-on-Forecast QPFs
- End-of-Week Survey and Closing Remarks

APPENDIX C: Products Used in HMT-Hydro Experiment

Subjective evaluations of the real-time operations focused on the gridded flash flood probabilities and the experimental warning issued. Archived cases utilized the same probabilistic products with additional focus on the Warn-on-Forecast QPFs. The table below summarizes the products and observations that were available for both the archived cases (AC) and the real-time operations (RT) in the 2018 HMT-Hydro Experiment.

Product	Provider	Description	AC	RT
Flash Flood Observations				
Local Storm Reports	NWS	Operational reports of flash flooding used to validate warnings		X
mPING	NSSL	Citizen-scientist reports of flash flooding defined by four levels of severity		X
Streamflow	USGS/NWS/NSSL	Measurement of streamflow that have exceeded flood stage or a nominal return period flow (e.g., 5-yr return) in small, gauged basins		X
Quantitative Precipitation Estimations and QPE Comparison Products				
MRMS QPE (Radar-Only)	NSSL	Precipitation estimates from radar-only algorithm using reflectivity only; Derives instantaneous rates and multiple accumulation periods	X	
MRMS QPE (Dual-Pol Synthetic)	NSSL	Precipitation estimates from radar-only algorithm using various dual-polarization variables; Derives instantaneous rates and multiple accumulation periods		X
QPE-to-FFG Ratio	RFCs/WPC/NSSL	Compares a 1, 3, and 6-h rolling sum of MRMS QPE to most recently issued 1, 3, and 6 h FFG*	X	X
QPE Average Recurrence Interval	NWS/NSSL	Compares various MRMS QPE accumulations from 30-min to 24-h to precipitation frequencies from NOAA Atlas 14**	X	X
Quantitative Precipitation Forecasts				
Warn-on-Forecast	NSSL	Ensemble QPFs provided to FLASH on a 3-km resolution, 900x900 km domain every hour for a lead time of 0–3 h or 0–6 h (case depending)	X	
HRRR Forecasts	GSD	QPFs provided to FLASH on a 3-km resolution CONUS domain every hour for a lead time of 0–6 h		X
CAPS Forecasts	OU	Ensemble QPFs provided to FLASH on a 3-km resolution CONUS domain once a day at 0000 UTC		X
Hydrologic Modeling Products				
Max Streamflow	NSSL	Maximum streamflow forecast during an interval spanning up to 12 hours after valid time	X	X
Max Unit Streamflow	NSSL	Maximum unit streamflow forecast during an interval spanning up to 12 hours after valid time	X	X
Soil Moisture	NSSL	Analysis of soil saturation	X	X
Probability of LSR	NSSL	Gridded probabilities of receiving a local storm report (on a scale of 0.00 to 1.00)	X	X
Probability of Exceeding Unit Streamflow Values	NSSL	Gridded probabilities of exceeding defined maximum unit streamflow values to determine hazard magnitude (on a scale of 0.00 to 1.00)	X	X

* RFCs typically update FFG at synoptic (0000, 1200 UTC) and sub-synoptic (0600, 1800 UTC) times, but the FLASH server queries all RFCs once an hour for FFG updates. During heavy rainfall events, some RFCs produce intermediate FFG products and hourly queries ensure that FLASH catches these intermediate FFG issuances. The FFG product displayed in FLASH is a national mosaic. There are different methodologies used to produce FFG across the country (including gridded and lumped FFG as well as the flash flood potential index), so discontinuities in FFG values across RFC boundaries may exist. Since the FFG values are being obtained from the RFCs, no locally forced FFG values from NWS forecast offices are included in this national FFG mosaic.

** NOAA Atlas 14 does not yet include precipitation frequency estimates for the Northwestern United States. Precipitation frequency values were derived by NSSL for use in this product until the official grids are published.

APPENDIX D: Group Discussion Questions and Key Findings

This section will list the questions asked by the HMT-Hydro Experiment officers and PIs during the Friday morning group discussion. Below each question are the key responses and takeaways from the combined three weeks.

Question #1: Overall, how did you feel working the archived cases?

- Hurricane Harvey case was overwhelming and tiring (from first week)
- Hurricane Harvey case was worth seeing, especially with a high-end event looks like in the products (from week three)
- Falls Creek, OK case was more challenging, especially with the QPE-forced being more "on the fence" with making a warning decision
- Would have liked some geographical familiarity
- Interested in using a full suite of data and products since normal decision making not possible
- The archived cases helped build confidence with looking at products later on in the week (notable for real-time operations)
- Would like to work in pairs or group with case

Question #2: Was one of the cases particularly memorable? And if so, why?

- Hurricane Harvey:
  - Case was more of a slam dunk
  - Pushed forecasters outside of their comfort zones and had to consider the data more carefully
  - How Warn-on-Forecast handled the eyewall well with landfall but not the banding features
  - Good example of a high-end event
- Falls Creek, OK Event:
  - Need to look at all products for this event, since thunderstorm-based events more complicated than tropical system
  - Interesting to see how Warn-on-Forecast captured the Pooleville/Falls Creek ahead of time
- El Reno, OK Event:
  - More relatable event with how it progressed and its convection; Can happen anywhere

Question #3: As you moved through the three conditions, did you find value in the additional products?

- Warn-on-Forecast would vary with how it changed your level of concern after looking at the deterministic and QPE-only forced products
- Situations where Warn-on-Forecast added confidence in earlier product issuance and can give situational awareness of what hot spots to look at (including the consideration of short-term staffing needs)

- Warn-on-Forecast would focus attention on areas that were not looked at with the QPE-only forced conditions
- Probability of Exceeding Maximum Unit Streamflow Values products were too low while the Probability of Receiving a Flash Flood LSR product was too high with large spatial coverage
- Probabilistic products would help sway warning decisions
- Would like to see the convection and output itself from Warn-on-Forecast than just the response in the probability products from the hydrologic models (to help build trust in the Warn-on-Forecast system and downstream hydrologic output)
- One forecaster noted difficulties in issuing a warning for areas where it has not started raining yet

Question #4: What were some of the benefits and challenges associated with the products that provided probabilistic guidance?

- Probabilistic information helped decision making when on the fence with issuing a FFW
- Probability values need to be more sensitive and the color scales need to be tuned hotter (i.e., more reds/purples/pinks)
- Forecasters are used to dealing with probabilities; Quickly learn the signals and would learn to calibrate flaws of the model in their head in the real world
- Learning to trust new tools is a big deal and would have to figure out how many times do signals need to be shown before the product is trusted
- Warn-on-Forecast QPF forcing in the products have potential to add more lead time and improve specificity of information communicated to end users
- It was described as a benefit and a challenge to use the Probability of Receiving a Flash Flood LSR product and the Probability of Exceeding Maximum Unit Streamflow Values products in combination with one another instead of a standalone use of one or the other

Question #5: How do you foresee the third condition (probabilistic products with Warn-on-Forecast QPF) impact the warning decision process and communication of flash flood threats during operations?

- Helps identify new threat areas for NWS River Forecast Center decision support services (DSS), coordinate with end users, and determine potential staffing needs for "off-times" since only 16-hour operational days
- Beneficial for warning decision making and DSS, notably with communicating with core partners, social media posts, and determining short-term workloads
- Since flash flood watches can be issued a day or two in advance and can be quite broad in area, it can assist with focusing in on areas and tuning message to areas of greater threats
- Potential for Warn-on-Forecast to aid in the issuance of flash flood watches
    - Difficulties with the NWS directive on watch issuance (24-48 hours in advance) while it is challenging to define a watch area with forecasted convective events

    - Sometimes flash flood watches are not issued in the convective season due to the challenges with defining the proper threat area
    - Given a trend for some offices to issue short-fused, targeted flash flood watches, it raises questions on how responsibilities and guidance with be shared amongst local and national NWS offices within 24 hours of the event
- Focus attention on areas that were not showing a current indication of a flash flood threat
- Had more confidence in geographical trends in the data but not the magnitude of the probability values forced with Warn-on-Forecast QPF
- Given the hesitancy to issue a warning in an area where it has not started to rain yet, the question is raised on whether the warning needs to be redesigned or to have a product other than a warning that is more suitable for this period (notable with communicating the associated probabilistic information)
- Extent to which Warn-on-Forecast will impact warning lead time will vary geographically and by storm type (e.g., pulse storms versus long-duration mesoscale convective systems)

Question #6A: What was your confidence/trust like when observing, understanding, and interpreting these products? (Gridded probability values)

- Confidence increased with increased use of products; however, confidence did not always match the shown probability values
- Confidence increased with greater geographical clustering of higher values
- Unsure if products are calibrated appropriately to flash flooding threat
- Lower values with the Probability of Exceeding Maximum Unit Streamflow Values products were perceived as not capturing the event, which reduced trust
- Lack of real-time feedback made it difficult to trust what you were seeing
- Unsure if probabilities for different products mean the same thing (i.e., does 50% in one product mean the same as 50% in another product)
- Interpretation of the products may be subjective (i.e., how does one define minor flash flooding)

Question #6B: What was your confidence/trust like when observing, understanding, and interpreting these products? (Warn-on-Forecast QPF input)

- Confidence was greater near the time of the Warn-on-Forecast model initialization
- Greater confidence with the convective-based cases than the tropical cases
- Feeling that greatest application is for DSS and short-term watch issuance, not necessarily multi-hour warning lead time
- Trust is similar with other convective allowing models that confidence will increase with increased use and comparing trends with observations (i.e., “don’t take any run as gospel”)
- Expected more run-to-run variation in Warn-on-Forecast; Seeing the same signal in multiple runs undermined some confidence
- Trusted trends more than the magnitudes of the values; however, lacked trust in the evolution of the forecast and how to warn farther downstream

- Could potentially gain confidence if can visualize the relative contributions of new (forecasted) rain versus what is currently driving runoff in output

Question #7: What did you think of the experiment design chosen for the archived case reviews (e.g., working independently, restricted data, the data collection process, and viewing the three conditions)?

- Some desire to work with other participants to discuss/collaborate with and to "maintain an open mind about what you are seeing"
- Prefer to be able to use Hazard Services to issue products instead of typing out every action
- Could be beneficial to debrief after each case
- Want more scenarios that are shorter and more geographically diverse
- Having the restricted data sets was good for focusing on the experimental products
- Need to minimize the repetition in the data collection process

Question #8: Any other comments?

- Unsure if wanted to utilize the flash flood emergency wording
- Schedule of events was questioned (specifically why the archived case days were starting when they were)
- Liked the small number of participants during the testbed experiment for easy contribution and access to researchers
- Need to consider what happens to product decision making when probabilities decrease
- Probability of Receiving a Flash Flood LSR worked more like a situational awareness tool than one for warning decision making

APPENDIX E: Responses from Feedback Survey

With regard to the HMT-Hydro Experiment introductions on Monday:

Field	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree	Average Value
The introduction helped me to understand experimental flash flood products sufficiently.	1	0	0	7	2	3.90
The introduction helped me to understand the experiment quantitative precipitation forecasts sufficiently.	0	1	1	6	2	3.90
I understood the anticipated outcomes and methodology after the presentations.	0	1	2	2	5	4.10
The introduction was effective in giving me more familiarity with the AWIPS capabilities during the experiment.	1	0	2	5	2	3.70

On a ranking from 1 to 5 where 1 is "Strongly Disagree" and 5 is "Strongly Agree"

With regard to the time allotted for each activity:

Field	Far Too Little	Too Little	About Right	Too Much	Far Too Much	Average Value
Introduction session	0	2	5	1	1	3.11
Real-time experimental flash flood warning operations	0	2	7	0	0	2.78
Case evaluations with the Warn-on-Forecast QPFs	0	1	6	2	1	3.30
Evaluation and discussion of the tools/warnings from the prior day	0	2	8	0	0	2.80
FFaIR daily briefings	0	0	5	4	1	3.60
FFaIR probabilistic flash flood forecast collaborations	0	0	8	1	1	3.30
End-of-week discussion on WoF QPFs and probabilistic grids for flash flood prediction	0	0	10	0	0	3.00

On a ranking from 1 to 5 where 1 is "Far Too Little" and 5 is "Far Too Much"… 3.00 average is ideal

Please indicate your level of agreement or disagreement with the following statements regarding the real-time experimental flash flood warning operations:

Field	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree	Average Value
In the forecasting sessions, I was given the tools that I needed to issue flash flood warnings.	0	0	0	8	2	4.20
The evaluation and discussion sessions helped me to improve my forecasts as the week progressed.	0	1	3	5	1	3.60
The FFaIR briefings and FFaIR collaboration gave me sufficient situation awareness to start the day.	0	0	2	6	2	4.00
The FFaIR briefings gave me all the information I needed to identify areas at risk.	0	3	1	6	0	3.30

On a ranking from 1 to 5 where 1 is "Strongly Disagree" and 5 is "Strongly Agree"

Please indicate your level of agreement or disagreement with the following statements regarding the archived case evaluations:

Field	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree	Average Value
In the archived cases, I was given the tools that I needed to evaluate the flash flood threat.	0	0	1	7	2	4.10
The AWIPS procedures were beneficial to the evaluation process.	0	0	0	5	5	4.50
The data collection form allowed me to provide detailed information on the experimental products.	0	0	0	7	3	4.30

On a ranking from 1 to 5 where 1 is "Strongly Disagree" and 5 is "Strongly Agree"

In terms of workload, please indicate the levels you felt across the whole week during each of the primary sessions:

Field	Much Lower than Average	Somewhat Lower than Average	About Average	Somewhat Higher than Average	Much Higher than Average	Average Value
Experimental real-time flash flood warning operations	1	4	4	1	0	2.50
Tools/warning evaluation and discussion sessions	0	1	9	0	0	2.90
Archived case evaluations	0	0	1	7	2	4.10
FFaIR briefings and FFaIR collaborations	0	3	5	1	1	3.00

On a ranking from 1 to 5 where 1 is "Much Lower than Average" and 5 is "Much Higher than Average"… 3.00 average is ideal

Was the material provided before the experiment helpful in understanding and preparing for the experiment?

Field	Not at All Helpful	Somewhat Not Helpful	Neutral	Somewhat Helpful	Very Helpful	Average Value
Forecaster Response	0	1	1	6	2	3.90

On a ranking from 1 to 5 where 1 is "Not at All Helpful" and 5 is "Very Helpful"

Would you consider participating in this experiment again in the future?

Field	Yes	No	Undecided
Forecaster Response	9	0	1

Would you recommend participating in this experiment to colleagues?

Field	Yes	No	Undecided
Forecaster Response	10	0	0