The 2014 Multi-Radar/Multi-Sensor (MRMS) HWT-Hydro Testbed Experiment

Final Report

02 January 2015

Prepared By

Robert A. Clark III[1] and Jonathan J. Gourley[2]

[1]*Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, OK*
[2]*NOAA/OAR/National Severe Storms Laboratory, Norman, OK*

Table of Contents

List of Tables

List of Figures

Introduction

From July 7 to August 1, 2014, the Hazardous Weather Testbed (HWT) at the National Weather Center (NWC) in Norman, OK, hosted the inaugural Multi-Radar/Multi-Sensor (MRMS) HWT-Hydro Testbed Experiment (hereafter called "HWT-Hydro"). The overarching purpose of the experiment was to explore the utility of the FLASH (Flooded Locations and Simulated Hydrographs) suite of experimental heavy rainfall and flash flood forecast and monitoring tools in the issuance of warm season experimental flash flood watches and warnings. During the four weeks of the experiment, 17 National Weather Service (NWS) participants from across the U.S. traveled to Norman and engaged in various experimental activities; detailed participant information can be found in Appendix A.

Activities included training on the suite of MRMS-FLASH tools, forecast shifts to issue experimental flash flood watches and warnings, daily sessions to evaluate experimental forecasts and the tools used to generate them, and "Tales from the Testbed" webinars to spread initial findings and recommendations to NWS local and regional offices. Researchers from the National Severe Storms Laboratory (NSSL) and the University of Oklahoma (OU) administered the project and the NWS Warning Decision Training Branch (WDTB) provided physical space and computing resources.

The inaugural HWT-Hydro experiment had four specific goals: 1) prepare the FLASH tools for transition to NWS operations through preparation of training materials, display of real-time products in AWIPS II (Advanced Weather Interactive Processing System), the NWS's primary forecasting and warning software suite, and conduct a system usability survey of the FLASH products; 2) ingest and use of a near-real-time flash flooding observation dataset incorporating multiple sources of information; 3) issuance of experimental flash flood watches between zero and six hours prior to the anticipated start of a flash flooding event; and 4) issuance of experimental flash flood warnings just prior to and during flash flooding events; 5) subjective evaluation of all experimental observations, tools, and forecast products. Every effort was undertaken to mimic the general operational organization of flash flood forecasting within the NWS. To this end, experimental staff coordinated HWT-Hydro activities with the second-annual Flash Flood and Intense Rainfall (FFaIR) Experiment conducted at the NWS Weather Prediction Center (WPC). This report discusses the activities of the HWT-Hydro Experiment and presents findings from it with a specific emphasis on operational impacts and recommendations for future investigation.

Experimental Activities and Schedule

Each of the four weeks of the experiment followed a similar schedule; participants arrived at the testbed Monday morning and departed Norman early on Friday afternoon. Table 1 is a general outline of the experimental schedule. Forecasters spent a total of thirty-eight hours per week in the testbed. Of that time, fifteen hours were spent in experimental forecasting shifts, 10 were spent collecting data via 3 survey instruments, and the rest in other activities.

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 scientific research received preference. Prior to their arrival in Norman, participants were given general information about the principal scientific goals of the experiment, but were not officially exposed to any experimental products or tools until the Monday afternoon training session. In this session, four separate presentations were given: a reiteration of the scientific goals of the project; detailed descriptions and usage examples of all constituents of the suite of FLASH tools; AWIPS II training that focused on the differences between it and AWIPS I; and an explanation of the survey and audio/visual recording data to be collected throughout the experiment.

Table 1. Weekly experimental schedule of HWT-Hydro. Gray shading corresponds to non-working hours.

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
8 AM					Evaluation
9 AM					Evaluation
10 AM					Webinar Prep.
11 AM	Facility Tour				Lunch Interview
Noon	Lunch	Lunch	Lunch	Lunch	Webinar
1 PM	Wx Briefing	Wx Briefing	Wx Briefing	Wx Briefing	Feedback Survey
2 PM	Training	Evaluation	Evaluation	Evaluation	
3 PM	Training	Evaluation	Evaluation	Evaluation	
4 PM	Forecasting	Forecasting	Forecasting	Forecasting	
5 PM	Forecasting	Forecasting	Forecasting	Forecasting	
6 PM	Forecasting	Forecasting	Forecasting	Forecasting	
7 PM		Forecasting	Forecasting	Forecasting	

Weather Briefing with WPC

One benefit of conducting HWT-Hydro during the month of July is the timing overlap with the FFaIR experiment at WPC (Barthold and Workoff 2014). HWT-Hydro coordinated with FFaIR in an attempt to mimic the operational cascade of responsibilities from heavy rainfall guidance from WPC down to the issuance of flash flood watches and warnings from local forecast offices. In the HWT-Hydro framework, FFaIR took on the role of a national center, providing daily guidance on synoptic-scale heavy rainfall potential, numerical weather prediction diagnostics, and probabilistic forecasts of various heavy rainfall and flash flooding parameters. HWT-Hydro participants took on the role of

a floating, national Weather Forecast Office (WFO), using FFaIR's guidance as a starting point. In general, FFaIR was responsible for forecasting heavy rainfall and flash flooding potential for timescales greater than six hours with HWT-Hydro taking on the responsibility for forecasting less than six hours prior to an event. The main conduit for interaction between the two experiments was a daily videoconference weather briefing from 1 – 2 PM CDT, Mondays through Thursdays. HWT-Hydro participants had the chance to ask questions of the FFaIR participants at each briefing and frequently took the opportunity to do so.

Experimental Forecast Shift

Experimental forecast shifts are, with evaluation sessions, the heart of the HWT-Hydro Experiment. These sessions were nominally slated to begin at 4 PM CDT Monday through Thursday, though if the training session was shorter than expected due to a lack of questions or if an evaluation session went quickly due to a lack of significant events to interrogate, the forecast shift began ahead of schedule. At the latest, forecast shifts ended at 8 PM CDT, though weekly experiment coordinators had wide latitude to dismiss participants early if weather conditions were not conducive to flash flooding. Within forecast shifts, participants were expected to issue experimental flash flood watches and warnings, as necessary, for any portion of the Lower 48 they believed flash flooding was impacting. Specific characteristics of these experimental watches and warnings are in the subsequent Experimental Datasets section of this report.

Evaluation Session

A forecast evaluation session took place at the beginning of each Tuesday – Friday session of the experiment. In this session, weekly experiment coordinators walked the participants through an online survey with questions about the relative ability of the forecast tools and the observations to properly diagnose the spatial extent and magnitude of the flooding that occurred. Forecasters were additionally asked to determine the ability of the forecast tools to detect flooding. Participants were given the opportunity to rank their forecasts and warnings relative to the operational equivalents, in terms of magnitude, uncertainty, and lead time. In consultation with the participants, experiment coordinators often chose to administer this survey multiple times in one evaluation session in order to collect independent responses for different regions. Experiment coordinators used the FLASH web interface (flash.ou.edu) to display experimental tools, products, and observations as well as operational flash flood warnings. The Iowa Environmental Mesonet's online archive of NWS text products was used to display operational flash flood watches, when necessary. All evaluation survey results were recorded online via the OU's Qualtrics software platform. Appendix B contains further details of the evaluation questionnaire.

System Usability Survey

A version of the System Usability Survey (SUS; Brooke 1996) was administered to participants at the end of the Monday forecast shift, after four hours of exposure to the FLASH tools in AWIPS II, and again at the end of the Thursday forecast shift, after a total of 16 hours of exposure to the tools. This survey was administered using the Qualtrics platform. The SUS consists of ten Likert (1932) items, each of which has five ordered response levels. Each item is scored from 0 (strongly disagree) to 4 (strongly

agree). The score for each negative item is determined by subtracting this coded value from five and then summing all five negative items. The positive score is determined by subtracting one from the coded value for each statement and then summing all positive items. The negative and positive scores are added together and multiplied by 2.5 to yield the final system usability value on a scale from zero to one hundred. The wording of the FLASH usability survey is nearly identical to Brooke's original survey. Appendix B contains the exact wording of the prompts that made up this implementation of the SUS.

End of Week Interviews

Friday morning, at the conclusion of the evaluation session, participants engaged in an audio-recorded free-form group interview in which they were prompted to talk about their week of experience in flash flood and heavy rainfall forecasting and monitoring. Forecasters additionally discussed their opinions about the various activities making up the HWT-Hydro Experiment, provided recommendations for future improvements to the FLASH suite of tools, and described their approach to uncertainty, probability, and confidence in flash flood forecasting. Different members of the HWT-Hydro staff conducted these interviews throughout the experiment. Results from these interviews will be reported in subsequent publications.

'Tales from the Testbed'

In association with the WDTB, participants were asked each week to prepare a short presentation on what they learned during their time in the testbed. The WDTB invited all NWS forecast offices, River Forecast Centers (RFCs), and regional centers to join these webinars, which took place Friday afternoons during the experiment. Participants used their webinar time to share tips for how to use various components of the FLASH tool suite in operations, to describe interesting flash flooding cases they encountered during their experimental shifts, and to answer questions about future development work on the FLASH suite and its constituents. Participants were instructed during the Monday training sessions to collect screenshots of interesting or important FLASH tools from AWIPS II as desired throughout the week. Experiment coordinators assisted participants in preparing webinar segments Friday mornings after the final evaluation survey. On average, between twenty and thirty different NWS offices from across the U.S. tuned into the webinar each week.

Feedback Survey

The final activity of each week of the experiment was a short online feedback survey administered via the Qualtrics system. This feedback survey gave participants a chance to expound on experimental activities including the amount of time assigned to each endeavor, the level of mental stress experienced during various activities, the physical setting and technical set-up of the testbed, and suggestions for improvement in future experiments. The feedback survey consists of 15 questions and one comment box. Participants were asked to rate the Monday introductory activities, the time allotted for four separate activities, and if they had the appropriate tools to issue experimental products and if discussion and evaluation helped to improve their forecasts. Forecasters ranked their workload during various activities and were asked if they would want to participate in the future or if they would recommend participation to their colleagues. The survey instruments are reproduced in full in Appendix B.

Experimental Datasets

Forecast Tools

Within AWIPS II, experiment participants had access to a range of operational NWS forecast guidance, including the regular runs of the GFS (Global Forecast System), NAM (North American Mesoscale), and RAP (Rapid Refresh) models. Forecasters also had the ability to view observed soundings from the NWS upper-air network of rawinsondes, surface observations from the national network of ASOS (Automated Surface Observing System) stations, and data from the GOES (Geostationary Operational Environmental Satellite) program. Forecasters had access to the European Centre's global forecast system but were limited to freely available, unencrypted model outputs. Local radar data from the Terminal Doppler Weather Radar (TDWR) or WSR-88D (Weather Surveillance Radar-1988 Doppler) networks was generally not available in AWIPS II due to bandwidth limitations. Outside of AWIPS II, forecasters could access additional tools via web browser or personal device.

A total of 35 experimental forecast tools were available within AWIPS II to the participants. Thirty-three of these are part of the suite of FLASH tools and an additional two are part of the suite of MRMS Hydro tools. The tools were segregated into six main categories:

1. Hydrologic models
2. Precipitable water
3. Quantitative precipitation estimates/forecasts (QPE/QPF)
4. Flash flood guidance (FFG)
5. Precipitation return periods (or "average recurrence interval"; ARI)
6. Radar

In category one, six tools were available from two separate hydrologic models. Two outputs from the Sacramento Soil Moisture Accounting (SAC-SMA) model (Burnash et al. 1973) were available: soil moisture and streamflow. Four outputs from the Coupled Routing and Excess Storage (CREST) model (Wang et al. 2011) were also available including streamflow and soil moisture. An additional two CREST model outputs, "CREST Max Return Period" and "HRRR-Forced CREST" (High Resolution Rapid Refresh), are sometimes known as "FLASH" and consist of a forecast of the return period of the maximum simulated streamflow at each grid cell between thirty minutes prior to the forecast valid time and 6 hours after the forecast valid time.

Category two consists of four tools: two surface – 300 hPa precipitable water analyses, one from rawinsonde observations and one from the RAP weather model. These analyses were compared to a gridded monthly precipitable water climatology developed by M. Bunkers (2014). In week one of HWT-Hydro, the precipitable water climatology was used to calculate the percentile of the analyzed precipitable water value at each point. At the request of experimental participants, from week two forward the climatology was used to calculate the standard deviation of the analyzed precipitable water value at each point relative to the monthly mean at that point.

QPE and QPF were provided from the MRMS suite of tools (Zhang et al. 2014) and the HRRR suite of tools, respectively. Flash flood guidance is provided in gridded format by RFCs across the U.S. (Clark et al. 2014). This mosaic is then compared to MRMS QPE or HRRR QPF and to produce grids of QPE-to-FFG ratio (also referred to

as FFG ratio). Precipitation average recurrence intervals consist of MRMS QPE grids compared to ARI grids from NOAA Atlas 14 (Perica et al. 2013). Atlas 14 analyses are not yet available for states in the Pacific Northwest, northern Intermountain West, or New England, as well as the states of New York and Texas. For New York and New England, alternative data from the Extreme Precipitation in New York & New England project is used (DeGaetano and Zarrow, 2010).

Finally, two MRMS radar reflectivity factor mosaics were provided to participants. The first, "MRMS Quality-Controlled Composite Reflectivity", consists of the maximum reflectivity factor value, regardless of vertical level, at each grid point. The second, "MRMS Seamless Hybrid-Scan Reflectivity", consists of the reflectivity factor at the lowest unblocked vertical level at each grid point. Appendix C contains names of and basic information about each of these tools.

Observations

During the experiment, four separate sources of flash flood observations were available to participants and staff: automated streamgage measurements collected by the United States Geological Survey (USGS), Local Storm Reports (LSRs) collected by NWS WFOs, unsolicited public geolocated smartphone or mobile phone reports from the mPING (Meteorological Phenomena Identification Near the Ground) project run by NSSL and OU (Elmore et al., 2014), and solicited public landline phone reports from the SHAVE (Severe Hazards Analysis and Verification Experiment) project also run by NSSL and OU (Ortega et al., 2009; Gourley et al. 2010). For the analyses presented in this report, a fifth observation data set has been included: flash flood events from the NWS *Storm Data* publication, which is heavily post-processed and not available until well after events have been initially reported. Figure 1 is a map of all USGS streamgage reports, LSRs, *Storm Data* reports, mPING reports, and SHAVE interviews collected during the experiment.

USGS streamgages are located on catchments of various sizes across the U.S. In order to qualify for inclusion in this observation database, a flash flood event recorded at a streamgage 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 (0.9 $m \bullet hr^{-1}$) of the stage (B. Cosgrove 2014, personal communication). Only streamgages with contributing drainage areas of less than 2,000 km^2 are considered. Twenty-five separate events were recorded during the entire HWT-Hydro Experiment, but only nine of these occurred during times covered by experimental shifts.

NWS Local Storm Reports are issued during or immediately after a given hazardous weather event (Horvitz 2012). They 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. Flash flooding LSRs will typically include a short description of the exact impact of the reported event in plain English. There were 473 flash flood LSRs across the U.S. during the four weeks of the HWT-Hydro Experiment, but only 250 of these occurred during times covered by experimental forecast shifts.

Closely related to LSRs are reports in the NWS publication *Storm Data* (MacAloney 2007). In contrast to LSRs, they can contain a range of times and also a spatial range. In general, *Storm Data* reports will be correlated with LSRs, but there are situations when a flash flood only comes to light days after an event and thus is absent

from the LSR database but present in the *Storm Data* database. There were 184 reports of flash flooding in *Storm Data* during the 16 experimental forecast shifts of HWT-Hydro.

mPING uses the recent proliferation of GPS-enabled smart phones and other mobile devices to crowd-source surface weather conditions. Users can identify the relative severity of the observed flood using a 1-4 integer scale, where "1" corresponds to the least risk to human life and limb and "4" corresponds to the greatest risk to life and limb. For example, a "1" flood corresponds to a river or creek out of its banks, or flooding in a yard, basement, or over cropland. A "4" flood requires homes, buildings, or cars to be swept away by floodwaters. There were 39 mPING reports during the four weeks of the experiment, but only 35 of these occurred during experimental shifts.

SHAVE is the fifth and final observational dataset used in conjunction with the experiment. Four undergraduate students were hired for the duration of the experiment and worked weekdays from roughly 12 to 9 PM CDT calling landline telephone numbers located in or near experimental flash flood warnings issued by participants. Each respondent was given a short telephone survey and on the basis of this survey the report was categorized into one of the four mPING impact classes or identified as a "No Flooding" report. During the experiment, students completed 1,254 phone interviews, of which 124 reported some sort of flash flooding impact at or near their location.

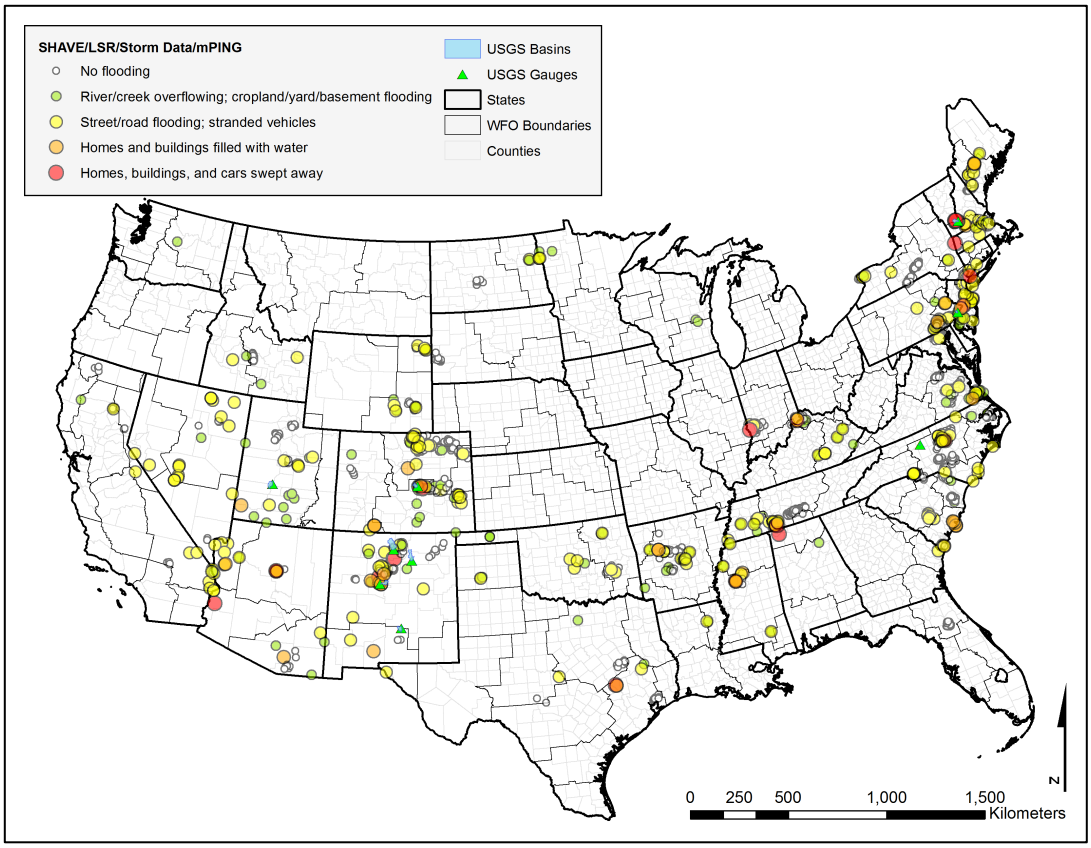

Figure 1. Flash flood observations from all four databases used during HWT-Hydro Experiment. The color coding corresponds to the impact classes denoted in the figure legend.

Figure 1 shows the mPING and SHAVE impact classes at the time of collection and the LSRs and Storm Data reports were grouped into the same classes during post-processing. Major flooding is defined as any class "3" or "4" impact (as well as any injury or fatality) and nuisance flooding is defined as any class "1" or "2" impact.

Products

In common National Weather Service 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, four types of products are considered: operational flash flood warnings, operational flash flood watches, experimental flash flood warnings, and experimental flash flood watches. Figure 2 is a map of all operational and experimental flash flood watches and warnings valid during any part of an HWT-Hydro experimental forecast shift.

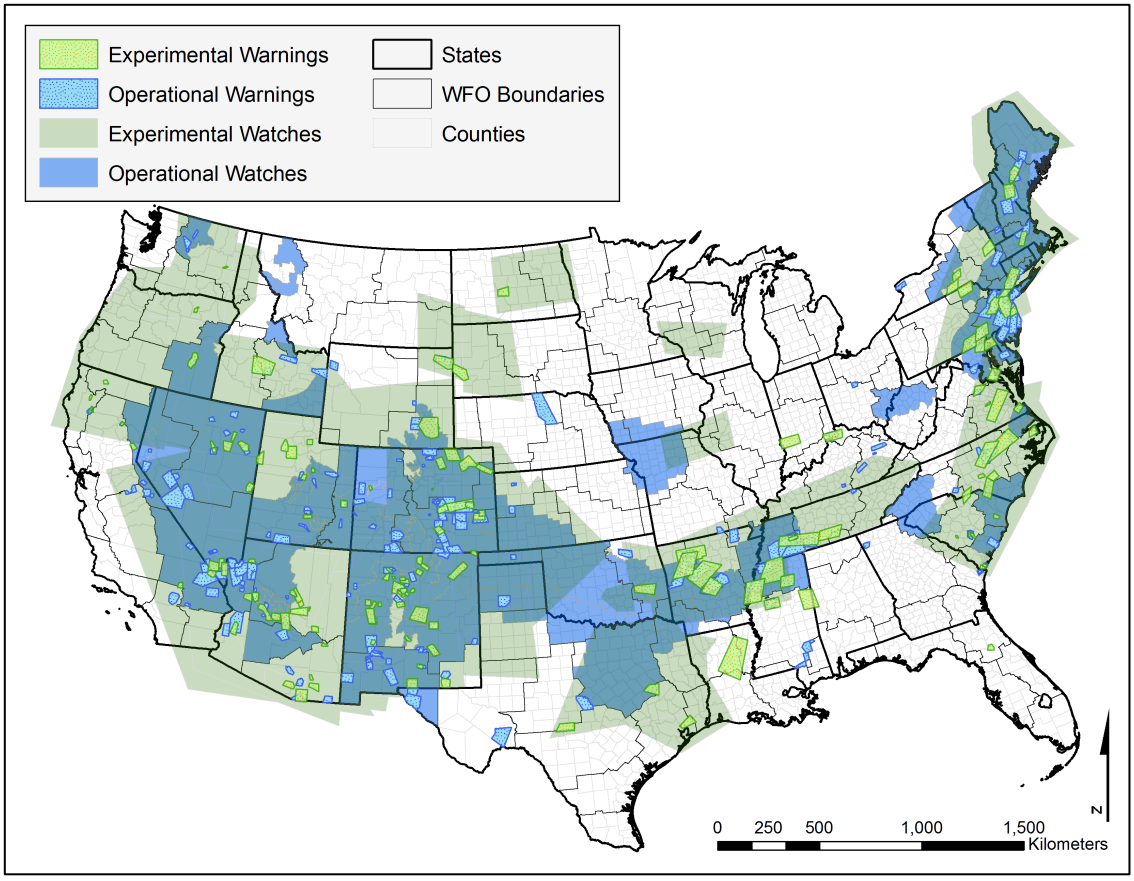

Figure 2. Experimental and operational flash flood watches and warnings valid during HWT-Hydro experimental shifts.

Operational flash flood warnings 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, but in the HWT-Hydro context those caused by heavy rainfall are of chief interest. These products 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 (or AWIPS II, depending on the office), called Warngen. The forecaster draws a polygon with as many vertices as needed to accurately encompass the threat. Based on this polygon, Warngen determines which counties and locations should be in the warning text, produces the appropriate text, and then disseminates the warning. During HWT-Hydro, 236 operational flash flood warnings were valid during any portion of an experimental shift.

Operational flash flood watches are used to alert the public that flooding is possible six to forty-eight hours before an event (Clark 2011). They are also issued at the WFO level and do not cross the boundaries of WFOs. These watches are not polygons in the same sense as *Storm Data* reports or operational flash flood warnings. Instead, watches cover a set of counties or parishes, or in areas with large counties, forecast zones defined at a sub-county level. They are generated operationally in the GHG (Graphical Hazards Generator) software program. Unlike warnings, watches can be issued before they officially enter into effect. Watches contain a generalized non-technical synopsis of the anticipated event. Operational flash flood watches are supposed to be issued when the forecaster's confidence in flooding occurring within two days is between 50-80%. Operational watches were processed to include only those *valid* (not just issued) during some portion of an HWT-Hydro Experimental forecast shift. There were a total of 79 such watches.

Experimental flash flood warnings work similarly to their operational counterpart but with some important differences. In the testbed, WFO boundaries are unimportant. Participants were told to act as a national forecast office; in other words, they were responsible for forecasting and monitoring conditions for flash flooding across the entire Lower 48 and so experimental warnings could cross WFO boundaries. The investigators modified the default Warngen templates to require forecasters to quantify their uncertainty about the magnitude of flooding expected in each polygon. The probability of minor flooding (corresponding to mPING impact classes "1" and "2") and the probability of major flooding (corresponding to mPING impact classes "3" and "4") could be 0%, 25%, 50%, 75%, or 100% for either, but 0% minor flooding forecasts were disallowed. Areas of the Warngen template normally used for plain English explanations of the flooding threat were converted by the investigators and used by the participants to identify those forecast tools prompting the warning and what thresholds within those tools led to the warning. Although forecasters could identify a variety of valid lengths for their experimental warnings (ranging from 30 min to 3 hrs), in practice all chose to use a valid length of 3 hrs. Appendix D contains an example of an experimental warning and watch template as well as additional information about how Warngen functioned within the testbed. In all, 153 experimental warnings were issued during the 16 experimental forecast shifts. The slowest day saw only one warning issued during the entire shift and, on the busiest day, 24 separate experimental warnings were issued.

Experimental flash flood watches contain elements of their operational equivalents as well as of experimental flash flood warnings. Forecasters used identical Warngen templates for both their watches and their warnings. A single drop-down menu allowed them to choose whether the product was a watch or a warning, but the available minor and major probability categories are identical. Like operational flash flood watches, experimental watches are larger in area and longer in time than warnings.

However, experimental watches are not county-based but are drawn with the same Warngen polygon methodology used for warnings. Participants could also draw watch polygons that spanned WFO boundaries, unlike in the operational realm. A final important difference concerns lead time: official NWS watches are valid somewhere 6-48 hrs prior to an event. In the testbed, watches are valid for six hours starting immediately from the time of issuance. Forecasters could issue watches at the beginning of an experimental shift, to catch flooding during that entire shift, or they could issue watches later in the shift to catch flooding they forecast to occur overnight after their shift had ended. There were 51 experimental watches issued during HWT-Hydro.

All of the flash flood observations, tools and experimentally issued watches and warnings were evaluated by the participants. Because participants and experimental coordinators could give multiple evaluation surveys for the same forecast shift (e.g., if the tools performed differently for flooding in the western U.S. and the eastern U.S.), there are a total of 27 separate group responses to the evaluation survey. The evaluation survey consists of these questions: three relating to the performance of the experimental observations, five relating to the forecast tools, and six relating to the experimental watches and warnings. Observations (LSRs, mPING, USGS, and SHAVE) are ranked against one another from 1 to 4 (where 1 is the best) on the separate aspects of areal extent, magnitude, and specific impacts of flash flooding. A similar set of questions is asked of the forecast tools (MRMS QPE, ARI, QPE-to-FFG ratio, and CREST Max Return Period). The watch and warning section of the survey consists of five Likert scale questions where experimental watches and warnings are compared to their operational counterparts in the realms of spatial accuracy, uncertainty estimates, and magnitude assignment.

Results

Figure 3 contains information about how frequently various types of meteorological and hydrologic knowledge were employed in the issuance of experimental flash flood watches while Fig. 4 is the equivalent for experimental warnings. In NWS operations, flash flood guidance and the Flash Flooding Monitoring Program (FFMP), a component of AWIPS, are central to this issuance process. Another important component is "local knowledge", best expressed as a forecaster's connection with local networks of hydrologic and built-environment expertise. In other words, as forecasters become familiar with a particular area, they know what drainage basins and neighborhoods are most susceptible to inundation from heavy rainfall. In some cases, this knowledge is specialized enough to identify particular intersections and buildings. By design, HWT-Hydro required forecasters to discard their local knowledge and their experience with FFMP and instead attempt to issue experimental products – similar to their operational equivalents – using new and experimental tools largely unfamiliar to them.

Because forecasters can consider more than one factor when issuing a watch, the numbers in Fig. 3 sum to far more than the total of 51 experimental watches issued during the course of the experiment. These numbers arise from a textual analysis of the descriptive component of the experimental flash flood watch text. Because watches are in effect for six hours, meteorological and hydrological factors that persist for those timescales are necessary to issue one. Precipitable water, which tends to vary (and to be observed) over large space and time scales, was the most frequently cited factor in issuing a watch. The HRRR model, which provides quantitative precipitation forecasts (QPFs), is more important in issuing a watch than nowcasting tools like QPE-to-FFG ratio or ARI. Some watches were issued shortly before the expected commencement of flooding impacts, which allowed for the use of "radar trends" and "MRMS QPE". Note that the "kinematics" category here includes any mention of properties related to the wind field. For the longer-range forecasting required for the issuance of flash flood watches, 15 out of 51 watches (29%) used the HRRR-Forced CREST. Fewer watches (9) used the QPE-Forced CREST (a.k.a. CREST Max Return Period). For watch issuance, the hydrologic model tools and those incorporating QPF forcing were used far more frequently than traditional QPE-forced ARI or FFG.

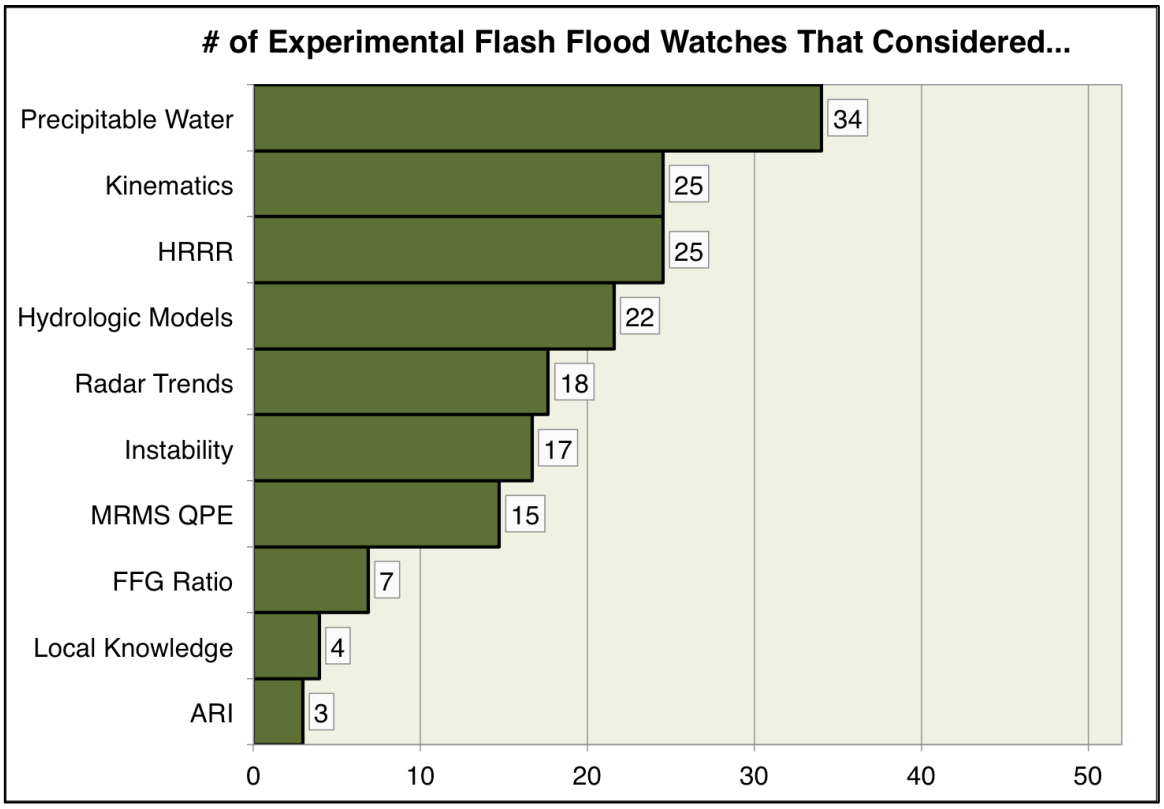

Figure 3. Histogram of tools and products analyzed prior to the issuance of experimental flash flood watches.

The products and tools analyzed prior to the issuance of experimental warnings differ from those used for watches. The most frequently used product for the issuance of warnings was the basic MRMS QPE product. Other traditional QPE-forced tools were also used frequently including QPE-to-FFG ratio and ARI. Larger-scale fields like atmospheric instability and precipitable water were infrequently used to issue warnings, as were any tools incorporating HRRR QPFs. Within the "Hydrologic Models" category in Fig. 4, QPE-Forced CREST was used in 35 experimental warnings, with HRRR-Forced CREST and Soil Moisture used more infrequently. In 7 out of the 153 cases, participants deemed the hydrologic model subset of tools unreliable.

Fifty-four warnings mentioned impacts, with roads the most frequent concern. Forecasters identified explicit reasoning behind their nuisance/major warning probabilities in approximately 30 cases, and in one case even adopted the NWS Storm Prediction Center's "Particularly Dangerous Situation" wording for an experimental product. Participants also sought outside information in a few cases, including information from USGS streamgauges and law enforcement reports in areas with which they were familiar.

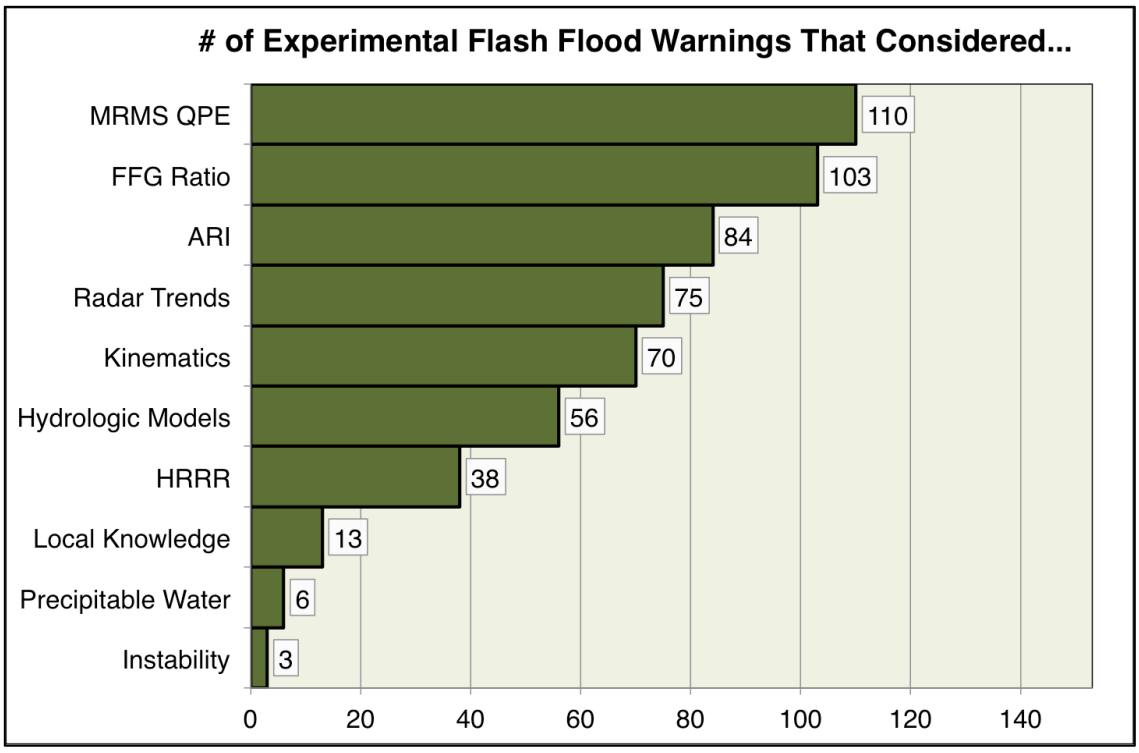

Figure 4. Histogram of tools and products analyzed prior to the issuance of experimental flash flood warnings.

Objective Skill of Experimental Watches and Warnings

The five flash flood observation datasets are used in combination to evaluate the experimental watches and warnings. A hit occurs when an observation falls inside a watch or warning polygon. If an observation falls outside a watch or warning polygon, a miss is recorded. Finally, watch or warning polygons that contain no corresponding observations are considered false alarms. In cases when more than one observation falls inside a single product polygon, multiple hits can be recorded. Contingency tables are populated (see Table 2) and the following statistics are computed: critical success index (CSI), false alarm rate (FAR), and probability of detection (POD).

Table 2. General structure of a contingency table used for evaluating flash flood watches and warnings.

		Observed?	
		Yes	*No*
Forecast?	*Yes*	Hit	False Alarm
	No	Miss	Correct Negative

The false alarm rate ranges from 0-1, where 0 indicates no false alarming as a result of the forecast and one indicates that all forecasts are false alarms. It is defined as the ratio of false alarms to the sum of hits and false alarms. The probability of detection also ranges from 0-1, but here a score of 1 is desirable, as that indicates that all observed events have been detected by the forecast (watch or warning). A POD of 0 indicates that no observed events are detected by the forecast. POD is defined as the ratio of hits to all

observations. CSI, sometimes called "skill", also ranges from 0-1, where 1 corresponds to high skill and 0 corresponds to no skill. CSI is calculated by dividing the number of hits by the sum of the hits, misses, and false alarms.

In Figure 5, the average of the individual CSI, POD, and FAR values for flash flood watches and warnings computed separately for the 5 observational datasets is plotted on the "Value" axis. Statistics for operational watches and warnings that were issued by the local NWS forecast offices during HWT-Hydro forecasting shifts are also shown for comparative purposes. Experimental warnings tend to have slightly lower skill than the operational equivalents, mainly as a result of lower detection ability. This is likely due to HWT-Hydro's attempt to cover the entire CONUS. An experimental shift could be staffed with as few as 2 forecasters – too few to cover the country on busy days, particularly those encountered during the experiment's 2nd week. The three metrics are similar between operational and experimental flash flood watches. Experimental watches tended to have a lower FAR, but this may be a function of the evaluation scheme, which does not penalize forecasters for drawing too-large polygons as long as at least one observation is contained within them. The average area of HWT-Hydro watches was twice that of the average operational watch, because forecasters had free reign to draw polygons without regard for County Warning Area (CWA) boundaries. Overall, these results suggest that despite unfamiliar surroundings and experimental tools, forecasters were able to produce reasonably successful experimental watches and warnings using the FLASH suite of forecast tools.

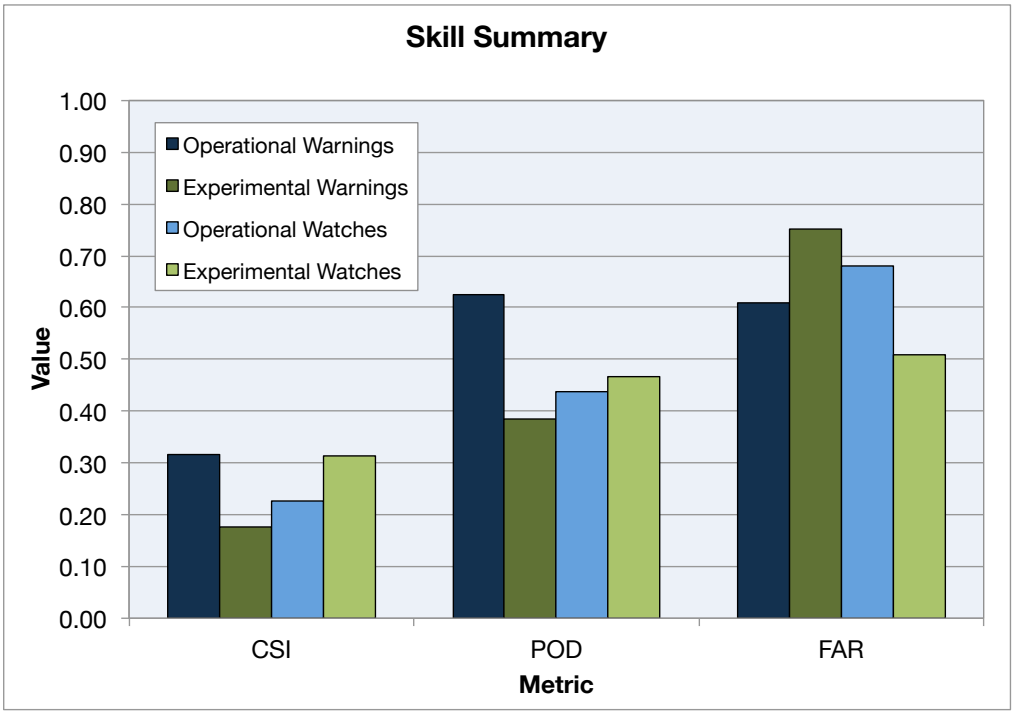

Figure 5. Summary of contingency table metrics for operational and experimental flash flood watches and warnings during HWT-Hydro.

Assigned Magnitudes and Probabilities to Experimental Watches and Warnings

Participants were required to select probabilities of "nuisance", or minor flooding, and major flooding for each experimental watch and warning they issued. Using a slightly modified analysis scheme from the preceding section, we computed the CSI, POD, and FAR of the experimental watches and warnings falling into each nuisance and major flooding probability bin. In this analysis, observations are treated equally and not segregated by reported mPING impact level. Figure 6 is a plot of experimental warning skill for forecaster-assigned nuisance events, broken down by probability assignment. As the probability assigned by the forecaster of a nuisance flood increases from 25% to 100%, the chance of that warning being a false alarm decreases by 40%. The POD changes little over that same comparison, but the decrease in false alarms is enough to double the CSI from less than 0.1 to more than 0.2. Forecasters are therefore able to use the experimental tools to determine how likely flooding is to occur within a particular flash flood warning. A similar FAR pattern exists for experimental watches, but there the skill improves from a 25% nuisance probability to a 75% nuisance probability before declining among 100% nuisance probability watches.

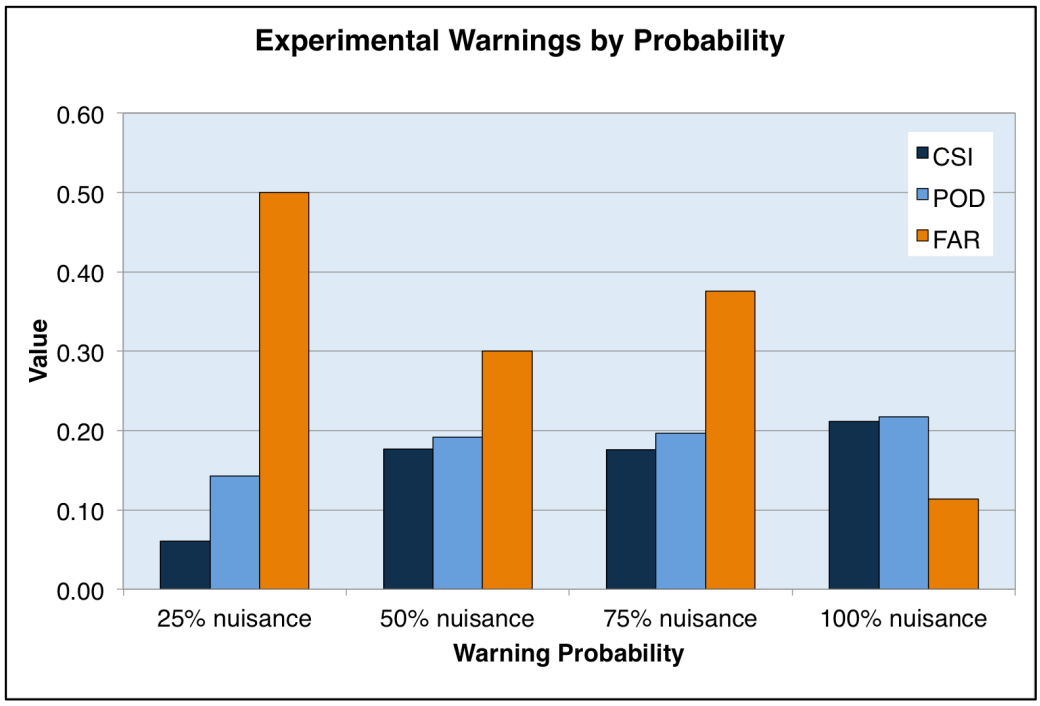

Figure 6. Skill of experimental warnings for nuisance flash floods, segregated by probability category.

Figure 7 provides the reliability of experimental, probabilistic warnings for nuisance category floods. Perfectly reliable probability forecasts (e.g., warnings) would line up along the 1:1 line along the diagonal. The curve is down and to the right of the 1:1 line for the 50%, 75%, and 100% nuisance probability categories, indicating over-forecasting (i.e., assigning high probabilities to events that did not occur that frequently). Figure 8 is the equivalent diagram for experimental, probabilistic watches for nuisance category floods. While warnings are more reliable at lower probabilities, watches are more reliable at moderate probabilities. In both cases, however, over-forecasting and

overconfidence is a problem at high probabilities and under-forecasting is a problem at low probabilities. This may reflect a need for additional training given that the present paradigm in flash flood warning is inherently deterministic in operations. In fact, participants frequently asked experiment staff for clarification on what the probabilities should correspond to for the nuisance and major flash flood categories.

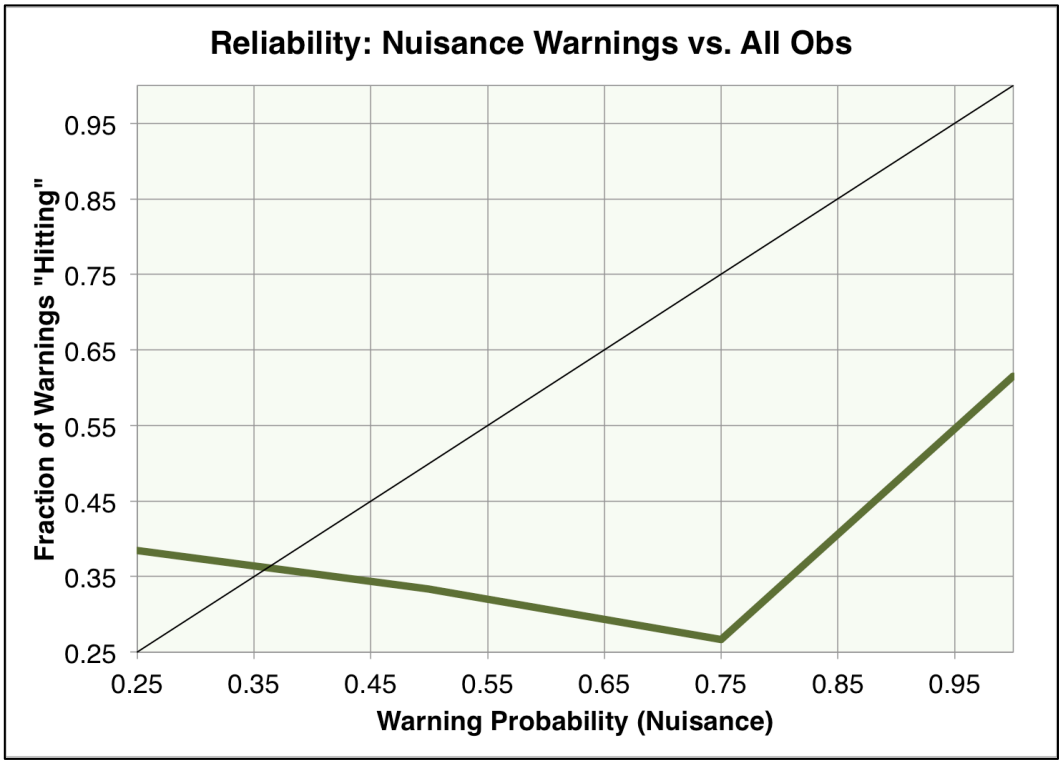

Figure 7. Reliability of experimental, probabilistic flash flood warnings for nuisance flash floods.

The reliability of watches and warnings for major category floods was determined by comparing the forecast products with the observations classified as a "3" or "4" on the mPING scale (N=57); these correspond to water in homes/building or homes/buildings/vehicles swept away. Experimental watches were reasonably skillful at forecasting major impacts at the 25% and 50% probabilities, but under-forecasting was a problem at the 0% probability level. Multiple participants expressed a desire to see a five or ten percent major probability choice included in the watch and warning templates instead of 0% major. For the experimental warnings, over-forecasting the major category flash floods was a problem at all probabilities greater than 0%. Of the warnings issued with 75% or 100% probabilities, less than one-third actually verified with at least one report of major impacts. There was also some slight under-forecasting at the 0% probability level. Reliability diagrams for the major probabilities are available in Appendix E.

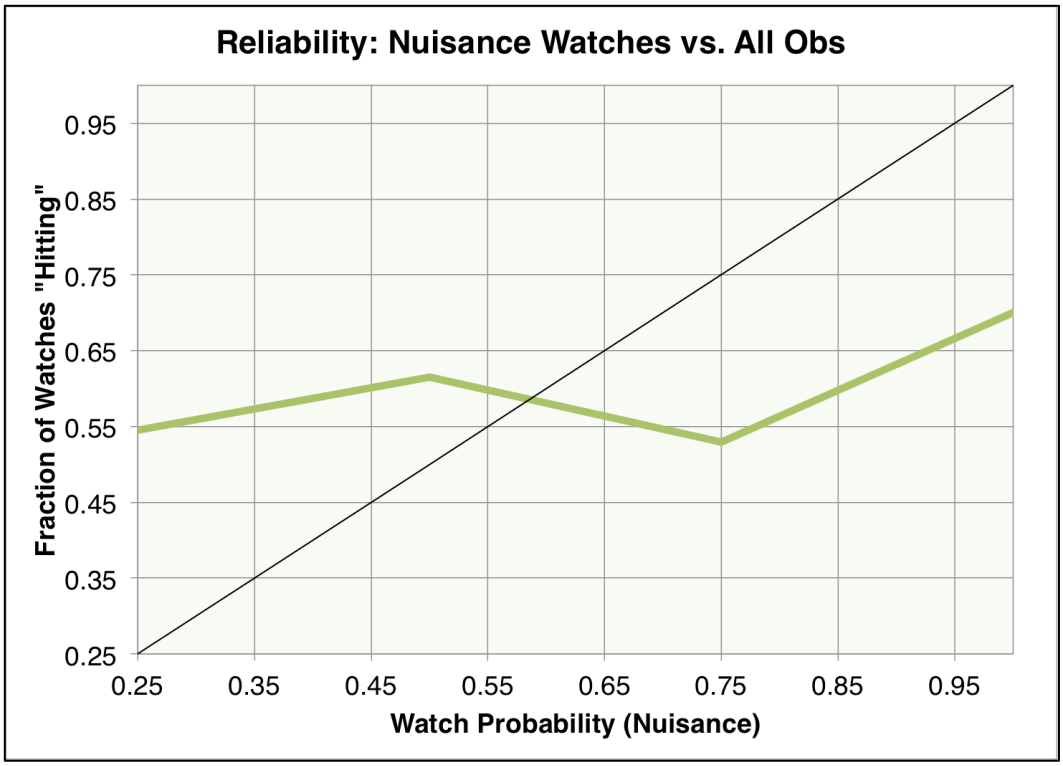

Figure 8. Reliability of experimental, probabilistic flash flood watches for nuisance flash floods.

Subjective Evaluation of Experimental Products and Tools

Participants in the daily evaluation sessions subjectively evaluated all experimental observations, tools, and watches/warnings they issued during the HWT-Hydro Experiment. A subjective evaluation involves qualitative human interpretation, which is useful in the context of evaluating the tools and forecast products given the known deficiencies in the flash flood observations. The results of the evaluation survey indicate that participants 1) preferred SHAVE reports and LSRs for flash flood observations, 2) found that each of the experimental tools contributed to their decision-making, and 3) felt that their own experimentally issued watches and warnings were comparable or better than the operational equivalent. The second point can be clarified by introducing an independent, objective evaluation of the tools. This study is underway and will be reported separately. The third point is partially supported using the objective evaluation in Fig. 5 for flash flood watches, but not warnings.

Among the types of flash flood observations, forecasters repeatedly said they preferred SHAVE observations and LSRs to mPING and USGS reports. The survey comments indicate that this is largely due to the unavailability of the latter two types with only 9 USGS observations and 35 mPING reports (see Table 9 in Appendix F). In the case of USGS streamgages a lack of gauged basins is to blame, while in the case of mPING increased adoption of the application by users may eventually elevate its ranking among the various observation types. Forecasters liked the density of the SHAVE reports and the inclusion of "no flooding" reports in the dataset, while indicating that if "no flooding" reports were also recorded in the LSR dataset the two would likely be of comparable utility. Figures of these results and additional results of the evaluation and other surveys are in Appendix F of this report.

Figure 9 shows results from the daily surveys about the tools' abilities to detect flooding events. In the figure, the percent of survey responses (N = 27) assigning each tool a particular ranking from 1-4 (where 1 corresponds to "Best" and 4 to "Worst") is plotted. The average rating with error bars equivalent to plus or minus one standard deviation of the population mean is plotted against the secondary ordinate. For each metric evaluated (i.e., detection, magnitude, and specific impact), there is no clear winner in terms of skill amongst the ARI, FFG, and MRMS QPE tools. The QPE-forced CREST tool has the lowest average ranking. The result that the most sophisticated tool for forecast flash flooding having the lowest average ranking was anticipated by the model development team. The QPE-based products have been developed for years and in some cases, decades. The version of the CREST model that was running during the experiment was the same "proof of concept" version that has been running for almost two years to demonstrate the ability to run a distributed hydrologic model in real-time across the CONUS at flash flood scale. but the standard error of the means of the other three tools overlap with one another.

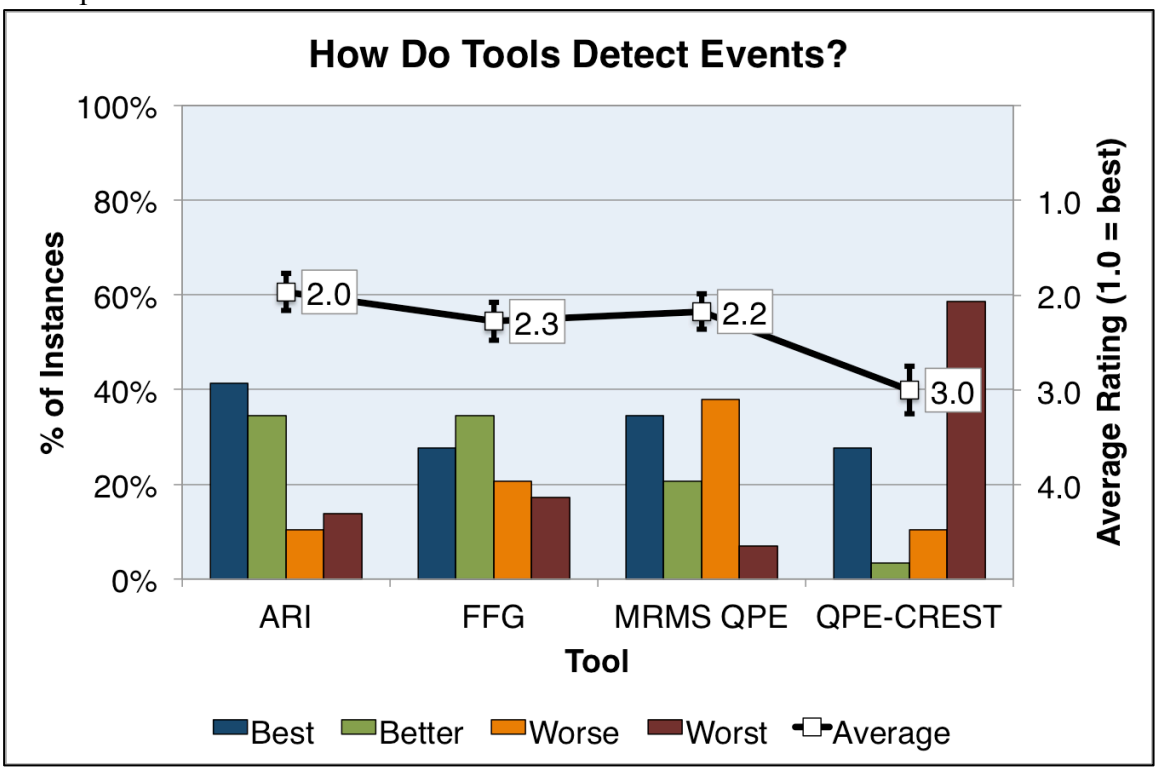

Figure 9. Subjective ranking of tools' ability to detect flooding.

The comments recorded by the experiment coordinators support the results shown in Fig. 9, with repeated mentions that the hydrologic model tools highlighted too many areas with very high values. However, no discernable trend among the comments related to ARI, MRMS QPE, or FFG exists, with equal numbers of respondents expressing support for or disdain for the three tools depending on the forecast shift. Participants noted that a combination of tools is often more useful than any one single tool. Forecasters said that the HRRR-forced CREST improved upon the lead time of the QPE-forced CREST in 52% of cases, but the comments suggest that the HRRR suffered from a

lack of run-to-run consistency, which severely limited the forecasters' confidence in the tool even on days when it would have provided significant lead time. On days when the HRRR did not provide lead time it suffered from the same general over-forecasting problem as the QPE-forced CREST product.

Forecasters generally found their experimental products equal to or slightly better than the parallel operational products, as indicated in Table 3. The differences between operational and experimental products in Table 3 (note that totals may not add up to 100% due to rounding) are most striking when considering the spatial accuracy of flash flood watches. No participants thought their watches worse than the operational equivalent, and nearly one-quarter thought their watches were better. The objective analysis buttresses this conclusion. The survey comments related to spatial accuracy and lead time express some of the pitfalls encountered during the experiment. In multiple cases events were ongoing at the start of experimental shifts so lead time comparisons were unfair or impossible. Other forecasters noted the difficulty in producing accurate polygons quickly during the beginning of a shift under stressful conditions and without appropriate environmental interrogation. One general trend in the comments indicates that forecasters felt the operational warnings were smaller and more precise; indeed, on average, experimental warnings were nearly twice as large in area as the average operational warning polygon. Sometimes this difference in area resulted in larger false alarm areas for the experimental warning, but, just as often, the experimental warning was the one that ended up catching an observation within its area.

Table 3. Subjective analysis of experimental watch and warnings' accuracy and lead time, evaluated in relation to the operationally issued products.

	Spatial Accuracy		**Lead Time**	
	Watches	**Warnings**	**Watches**	**Warnings**
Better or Much Better	23%	27%	17%	30%
About the Same	27%	31%	7%	20%
Worse or Much Worse	0%	23%	17%	20%
N/A	50%	19%	67%	30%

Forecasters were also asked to evaluate their own watches and warnings on the basis of the magnitude and uncertainty estimates they assigned to their products. These results are shown in Table 4. Note the high degree of similarity between the answers to the uncertainty question and the magnitude question. On multiple days, participants stated that they believed the uncertainty and magnitude sections of the survey were asking the same question. Participants also expressed concern about exactly what assigning, for example, a 50% probability to watch or warning meant. In general though, the comments suggest that there was confusion regarding these questions, and this likely explains the large discrepancy between forecaster-perceived skill versus objective reliability of the uncertainty and magnitude assignments of the experimental flash flood watches and warnings (see Figs. 7 and 8).

Table 4. Subjective analysis of experimental watch and warnings' magnitude and uncertainty assignments.

	Uncertainty		Magnitude	
	Watches	**Warnings**	**Watches**	**Warnings**
Too Low	14%	17%	14%	20%
About Right	48%	50%	45%	50%
Too High	10%	23%	14%	20%
N/A	28%	10%	28%	10%

Usability of the FLASH System

The System Usability Scale (SUS) is scored on a 0-100 point scale. Sauro (2011) states that the average SUS score across 500 different evaluations is 68. Other research successfully assigned adjectives to SUS scores (Bangor et al. 2009). Table 5 summarizes the results of the SUS administered to experimental participants.

Table 5. Results of the System Usability Survey for the FLASH product suite.

	Mean	**Median**	**Range**	**Std. Deviation**
All responses (N = 29)	71	73	48-95	13
Beginning of week (N = 13)	65	65	48-85	12
End of week (N = 16)	77	75	55-95	10

In general, the usability score of the FLASH system increased from the beginning of the week to the end of the week, with improvements in the mean, median, and range of scores. At the beginning of the week, system usability was slightly less than the average system, but increased above the average system by the end of the week. Using adjectives from Bangor et al. (2009), the FLASH system could be described as falling between "OK" and "Good" usability at the beginning of the week, and improving to "Good" usability by the end of the week. Sauro's (2011) archive of SUS results would place FLASH in the 40th percentile at the beginning of the week and is above the 80th percentile among all systems by the end of the week. Participants spent an additional 12-15 hrs with the system between responses, so this amount of hands-on training will significantly improve system usability. Histograms of the results of the survey can be found in Appendix F of this report.

Results of the Feedback Survey

Figure 10 is a graphical representation of the results from the survey questionnaire about the Monday introduction. On average (+/- one standard error of the mean), forecasters agreed that the Monday introductory section helped them to understand the experimental tools, the goals of the experiment, and how to use AWIPS II. All forecasters responded neutrally or positively to those three items, with the exception of one respondent who disagreed with understanding the experimental goals at the conclusion of the training.

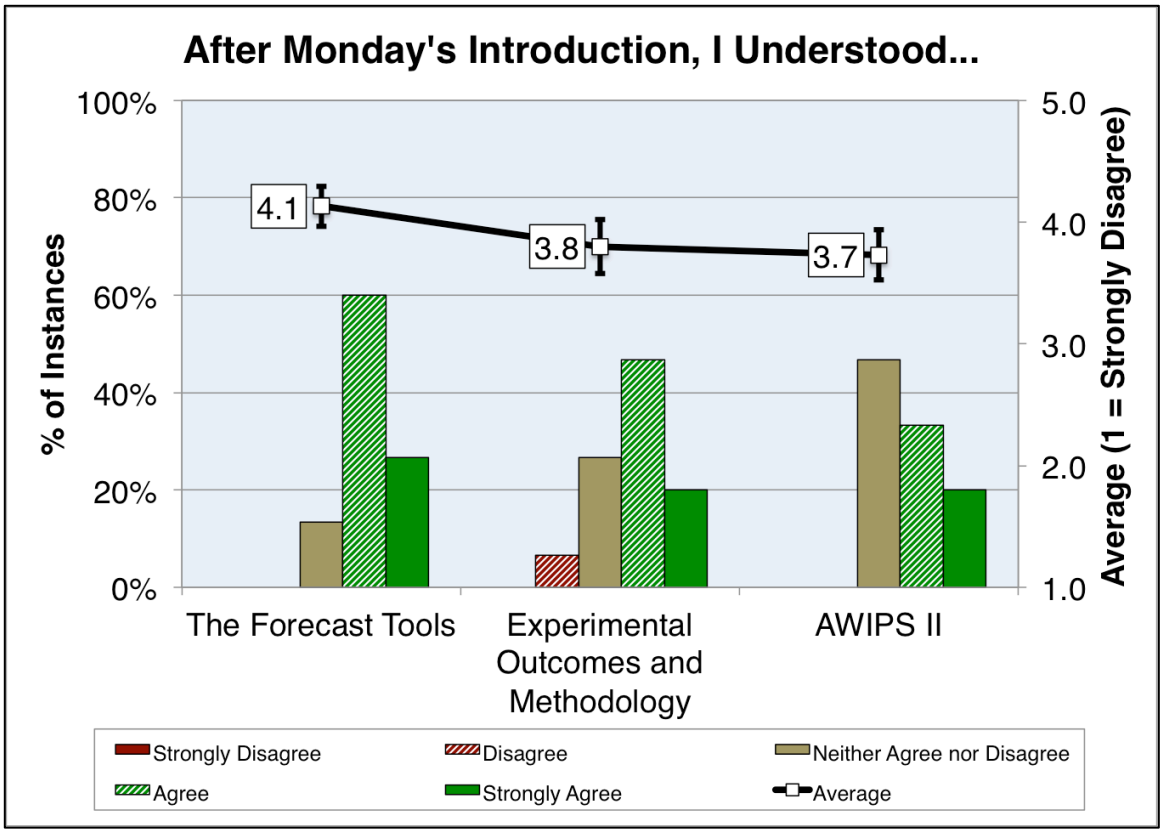

Figure 10. Results from feedback survey questionnaire about the Monday introduction session.

The majority of respondents also felt that the time allotted to the Monday introductory sessions, the evaluation sessions, and the 'Tales from the Testbed' webinar was appropriate. The biggest disagreement resulted when participants rated the evaluation sessions: only 60% of participants felt the appropriate amount of time was allotted to this activity, as 13% felt the sessions were too short and 27% thought them too long. Respondents generally stated that the experimental shifts, evaluation sessions, and webinar preparation sessions were associated with workloads similar to what they expected in the course of their regular duties. Here, the biggest disagreement with that consensus exists when considering experimental shifts, where only 53% thought them comparable to their regular workload. 27% thought the workload during the shifts was "somewhat higher" and 20% thought the workload was "somewhat lower".

Forecasters agreed that they had the tools they needed to issue experimental products throughout the week and that evaluation and discussion helped them improve their forecasts. Finally forecasters unanimously indicated they would recommend the experiment to colleagues, and all but one forecaster expressed interest in participating in future HWT-Hydro testbed experiments. Results from the other survey questionnaires are reported in Appendix F.

There were 27 separate comments left as part of the feedback survey; these can be broadly divided into four categories: technology, logistics, data, and experimental goals. Six of the comments related to technological concerns: three were variations of a request

to have one AWIPS II workstation per participant. This particular concern was rectified beginning with the third week of the testbed when two supplemental workstations were obtained from the WDTB. The other three technological concerns expressed frustration at the high latency of the AWIPS II software. Six additional comments concern logistics (one of these was simply praise for the administration of the experiment). One participant asked for a larger operations area and the other four comments were all variations on attempting to provide more training and explanation remotely prior to the start of the testbed. The comments in the experimental goals section provide insight into how forecasters reacted to the experimental products framework used during the testbed. One felt that the evaluation procedure was too cumbersome and another felt that the evaluation survey should have explicitly included the number of observations in a product or the distance of observations from a product. The two other comments were related to the experimental products themselves: one questioned about whether the experimental warnings were meant as a replacement for operational warnings alone or both operational warnings and advisories and the other questioned if the probabilities used for nuisance and major flooding were appropriate. The final category has the largest number of comments and they concern requests for additional data sources in the testbed. Of these ten comments, four asked for local radars and/or FFMP outputs. Two other comments were from participants asking for ensemble QPF tools to be developed for future experiments. Another two participants requested additions of tools used in the testbed to the FLASH web interface. The final two comments were from those asking for additional sounding data as well as 30-min accumulation periods for the MRMS QPE, ARI, and FFG ratio tools.

'Tales from the Testbed' Takeaways

Participants were given wide latitude when preparing the 'Tales from the Testbed' webinars. Most chose to present their impressions from one or two cases during their week in the testbed, although at least one participant presented a case study from an archived event in his home area. At the end of each webinar, each participant presented between one and three 'takeaways' from their time at HWT-Hydro. Despite the wide range of cases selected by participants, a few commonalities exist in these takeaways. Participants in all four weeks stated that *the experimental tools used in HWT-Hydro improved their ability to diagnose the potential severity of an event in near real-time compared to the tools used in operations*. There was also general agreement that, of the experimental tools available, rainfall ARI did the best job of precisely focusing in on areas at greatest risk of experiencing flooding impacts. Forecasters mostly agreed that the CREST Max Return Period and HRRR-Forced CREST tools increased the potential lead time of watches and warnings but at the expense of too-high return periods and larger possible false alarm areas. On a related note, in more than one webinar, participants noted that the FFG and ARI tools were good at focusing in on specific areas of flooding impact but often did so later than the CREST tools and thus reduced lead time. Those participants recommended using FFG and ARI in tandem with the hydrologic model tools. Forecasters were unanimous that environmental and situational awareness is critical before using any of the experimental tools in a watch or warning context. Finally, participants expressed their need to frequently diagnose potential issues with experimental tools and recommended that all users consider doing the following: monitor the quality of HRRR initializations, reference MRMS QPE to ground truth or River

Forecast Center precipitation estimates, and consider if using CREST, ARI, FFG, or MRMS QPE is appropriate in particular regions of the U.S.

Analysis and Recommendations

For Operations

Based on the analysis of the 'Tales from the Testbed' webinars, every effort should be made to bring the ARI tools into NWS operations following the MRMS transition path to NCEP Central Operations. Participants also expressed concern at inconsistencies between regions and offices regarding current flash flood watch procedures. The additional lead time provided by the HRRR-Forced CREST could be used to issue watches six hours or more prior to an event and RAP precipitable water analyses and anomalies should be standard tools for issuing flash flood watches. Experimental tools will make impact categorization in flash flood warnings and watches possible. Watches could additionally include information about the expected number of specific types of reports in and near a watch box and they should, like severe convective watches and most types of warnings, be issued as simple polygons, not county outlines. These polygons could more accurately be drawn around the observed or forecast meteorological and hydrological conditions. The NWS must also explore ways of standardizing the definition of flooding (or perhaps removing altogether the distinction between "flood" and "flash flood") and then determine how to better observe these dangerous phenomena. Promotion of the mPING project via NWS text products, NWS social media, and the NWS website should be undertaken, as this app allows the wisdom of the crowds to be leveraged into appropriately identifying and classifying flash flood events largely independent of watches and warnings, unlike the *Storm Data* publication or LSRs.

For Tool Development

The CREST and Sacramento hydrologic model-based reanalysis should be recreated using recently estimated physically-based *a priori* parameters, improved kinematic wave routing scheme and associated parameters, and MRMS precipitation as forcing. Note that the streamflow ARIs are computed from a model reanalysis that used StageIV precipitation as forcing, which differs considerably in scale from the MRMS inputs that are used in real-time during the experiment. This scale inconsistency likely explains some of the high biases in streamflow ARIs that were noted by the participants. Research should also explore the use of partial-duration timeseries in place of the annual maximum streamflow currently used to calculate these Log-Pearson III parameters. Particular attention needs to be paid to arid regions and areas with poor radar coverage. Log-Pearson III parameters may need to be extrapolated from other regions to these data-sparse areas in the Intermountain West.

Forecasters were also interested in seeing more ensemble forecast information. One possible way to achieve this is to replicate the CREST model methodology with the SAC-HTET (Sacramento-Heat Transfer and Evapotranspiration) hydrologic model. Both models could be implemented alongside one another in future iterations of the suite of FLASH tools. Along these lines, a time-lagged QPF ensemble from the HRRR is also recommended. This should include SAC-HTET and CREST outputs for two or three consecutive QPF solutions from the HRRR. Additional precipitation forcings at the flash

flood scale, such as those from the Warn-On-Forecast project should be considered in the future.

The most popular forecast tool, on average, was ARI. Grids for the northwest U.S. and Texas should be created as soon as possible and scientific review on the grids for New York and New England should continue. Until NOAA Atlas 14 can be completed for the Lower 48, the FLASH team should consider processing other sources of data to create stopgap grids for the areas currently missing from the ARI tool. The FLASH team should explore methods of presenting the entire depth-duration curve at each grid point. This will allow forecasters to choose the accumulation period of their choice while simultaneously reducing the number of separate raster grids that users of the FLASH web interface or AWIPS II have to search through.

Finally, though flash flood observations were an important part of HWT-Hydro, more improvement is required. The skill of experimental products and of experimental tools is still too dependent on the observational dataset used. LSRs and *Storm Data* are biased toward operational warning products. SHAVE is biased toward experimental products. USGS streamgages and mPING reports are too sparse for use in forecast evaluation. NSSL and OU should continue to promote mPING to the public and should consider interfacing with broadcast media partners and the NWS to accomplish this. Additionally, the mPING app could be modified, with user permission, to accept push notifications from the iOS or Android ecosystems when flooding is expected in the user's area in an effort to get more users to report flooding impacts. In essence, this would work similarly to the SHAVE project but without the added manpower required for polling residents via landline, and without any bias toward specific text products.

For Future Iterations of HWT-Hydro

The inaugural HWT-Hydro Experiment was held in the month of July; June or July is recommended for future experiments. The summer allows for the inclusion of monsoon-driven events in the Desert Southwest (over three-quarters of the experimental shifts in HWT-Hydro had some sort of activity in this area). The summer also allows for close coordination with the FFaIR experiment and avoids interfering with springtime severe convection studied by other experiments under the HWT umbrella.

Despite a number of flash flooding events during the 2014 HWT-Hydro Experiment, some days were notably slow. This is inevitable in this sort of research and the experiment administrators should develop at least one – and probably two – displaced real time AWIPS II flash flood simulations for this eventuality. These simulations should showcase positive and negative aspects of the experimental tools and should require the length of an experimental shift to complete. The opposite problem was observed on three separate days this year. Occasionally, the atmosphere was simply too active for 2-4 forecasters to tackle the entire Lower 48. On days when this is expected, in consultation with FFaIR, the experiment coordinators should delineate strict domain boundaries and require the testbed forecasters to remain within those. This would allow for the objective exclusion of operational products and observations outside of those domains on the affected days.

Although HWT-Hydro has no control over USGS streamgage siting, LSR or *Storm Data* report collection, and little control over mPING, it can determine how best to operate SHAVE in relation to flooding calls. The inaugural experiment asked callers to call within experimental warning polygons and to blanket each polygon with a dense

network of reports. Callers were also asked to test at least a few points very near, but just outside, warning polygons. This method subjects SHAVE to the same dependence on warnings products that plague LSRs and *Storm Data*. Instead, future iterations of HWT-Hydro should not allow callers to see any experimental or operational products. Callers should be provided with multiple experimental and operational forecast tools and told to call in and near areas they believe could be experiencing flooding. Future SHAVE callers should be instructed to stop calling on a particular flooding event after a handful (less than five) reports have been received. Similarly, if five "no flooding" reports are received for a particular event, SHAVE should move on to another area if one exists, and return to the original "no flooding" area if events later warrant. Finally, SHAVE callers should explicitly record the time of the report as given by the interviewee and should avoid "day-after" calling on events when possible. These modifications would reduce any bias toward experimental products in SHAVE reports and allow the callers to cover larger areas than at present.

Nearly every day of the experiment, multiple flooding events occurred either just before or (more commonly) just after the conclusion of an experimental shift. Therefore, future HWT-Hydro Experiments should be conducted with two daily experimental shifts. Given participants' overwhelming desire to recommend the experiment to colleagues (or to participate again), the investigators should anticipate receiving plenty of interest from forecasters in future years, pending available funding. Multiple shifts should mimic operations at NWS forecast offices. One possible implementation would have two forecasters arriving in the late morning and completing a two-hour evaluation session before a five-hour forecast shift. Then, their eighth and final hour would mark the arrival of the second shift of two forecasters. An hour-long weather briefing and situational awareness session with all four forecasters would follow; this could possibly be coordinated with the FFaIR experiment. Then the first shift would depart and the second shift would begin experimental forecast operations. Their final two hours would be spent in their own evaluation session. Given the large number of HWT-Hydro staff members present at the first year of the experiment, there should be plenty of people available to keep the experiment operating. This requires fewer high-powered (and thus, expensive) AWIPS II workstations (see Appendix G). Lastly, HWT-Hydro should, to the extent possible, conduct all training remotely prior to the arrival of participants. The need for in-depth descriptions of AWIPS II and the experimental tools will decrease as more sectors of the NWS become familiar with those systems.

Acknowledgements

HWT-Hydro was funded by NOAA/OAR/Office of Weather and Air Quality (OWAQ) under the NOAA cooperative agreement, NA11OAR4320072. Regional NWS headquarters also provided some funding for certain participants. The principal investigator is Dr. Jonathan J. Gourley of NSSL. Co-investigators are Elizabeth Argyle, Race Clark, Zac Flamig, and Brandon Smith of the Cooperative Institute for Mesoscale Meteorological Studies (CIMMS) at the University of Oklahoma. The experimental coordinator is Steve Martinaitis of CIMMS and OU. The logistics coordinator is Race Clark (CIMMS/OU) and the technical coordinator is Zac Flamig (CIMMS/OU). Weekly coordinators Elizabeth Argyle (CIMMS/OU), Ami Arthur (CIMMS/OU), Jess Erlingis (OU), Maria Moreno (CIMMS/OU), and Brandon Smith (CIMMS/OU) ensured smooth

experimental operations. Several dedicated undergraduate students from the OU School of Meteorology made cold calls to members of the public as part of the SHAVE project.

Faye Barthold (WPC) and Tom Workoff (WPC) ensured excellent coordination between HWT-Hydro and FFaIR. Ed Clark of the NWS Office of Climate, Water, and Weather Services provided important guidance, as did Dave Novak and Wallace Hogsett of the WPC. Brian Cosgrove of the NWS Office of Hydrologic Development created the USGS streamgage event identification methodology. The Hazardous Weather Testbed coordinator is Gabe Garfield of CIMMS and OU. The NWS WDTB graciously provided HWT-Hydro with office space and several AWIPS II workstations. Andre Reddington (WDTB) was extremely helpful in getting AWIPS II workstations set up and running and Tiffany Meyer (WDTB) assisted in answering AWIPS II questions. Finally, the experiment would not have been possible without the enthusiastically whole-hearted participation of the NWS forecasters from around the country.

References

Bangor, A., P. Kortum, and J. Miller, 2009: Determine what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, **4**, 114-123.

Barthold, F., and T. Workoff, 2014: The 2014 Flash Flood and Intense Rainfall Experiment. Hydrometeorological Testbed at the Weather Prediction Center. [Available online at http://www.wpc.ncep.noaa.gov/hmt/FFaIR_2014_final_report.pdf.]

Brooke, J., 1996: SUS - A 'quick and dirty' usability scale. *Usability Evaluation in Industry*, P. Jordan, B. Thomas, B. Weerdmeester, and I. McClelland, Ed., Taylor & Francis, 189-194.

Bunkers, M., 2014, cited 2014: Upper-Air Climatology Plots. US Department of Commerce, National Weather Service Rapid City, SD Weather Forecast Office. [Available online at www.crh.noaa.gov/unr/?n=pw.]

Burnash, R., R. Ferral, and R. McGuide, 1973: A generalized streamflow simulation system – conceptual modeling for digital computers. US Department of Commerce, National Weather Service and State of California, Department of Water Resources.

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.

DeGaetano, A. and D. Zarrow, 2010, cited 2014: Extreme Precipitation in New York & New England: Technical documentation & user manual. Northeast Regional Climate Center, Cornell University. [Available online at http://precip.eas.cornell.edu/docs/xprecip_techdoc.pdf.]

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., J. M. Erlingis, T. M. Smith, K. L. Ortega, and Y. Hong, 2010: Remote collection and analysis of witness reports on flash floods. *J. Hydrol.,* **394,** 53-62, doi: 10.1016/j.jhydrol.2010.05.

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.]

Likert, R., 1932: A technique for the measurement of attitudes. *Archives of Psychology*, **140**, 1-55.

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.]

Ortega, K.L., T.M. Smith, K.L. Manross, K.A. Scharfenberg, A. Witt, A.G. Kolodziej, and J.J. Gourley, 2009: The severe hazards analysis and verification experiment. *Bull. Amer. Meteor. Soc.*, **90,** 1519-1530.

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.]

Sauro, J., 2011, cited 2014: Measuring usability with the System Usability Scale (SUS), Measuring Usability, LLC. [Available online at http://www.measuringu.com/sus.php.]

Reed, S., J. Schaake, and Z. Zhang, 2007: A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *J. Hydrol.*, **337**, 402-420.

Wang, J., Y. Hong, L. Li, J. J. Gourley, S. Khan, K. Yilmaz, R. Adler, F. Policelli, S. Habib, D. Irwn, A. Limaye, T. Korme, and L. Okello, 2011: The coupled routing and excess storage (CREST) distributed hydrologic model. *Hydrolog. Sci. J.*, **56**, 84-98.

Zhang, J., K. Howard, S. Vasiloff, C. Langston, B. Kaney. Y. Qi, L. Tang, H. Grams, D. Kitzmiller, and J. J. Levit, 2014: Initial operating capabilities of quantitative precipitation estimation in the Multi-Radar/Multi-Sensor system. *28th Conference on Hydrology*, Atlanta, GA, Amer. Meteor. Soc., 5.3. [Available online at https://ams.confex.com/ams/94Annual/webprogram/Manuscript/Paper240487/28Hydro_MRMS_QPE_IOCs_Zhang%20et%20al.pdf.]

Appendix A: HWT-Hydro Participants and Staff

Table 6. HWT-Hydro participants

Week	Name	Affiliation
1	Jonathan Brazzell	NWS Lake Charles LA
1	Chris Legro	NWS Gray ME
1	Mike Moneypenny	NWS Raleigh NC
1	David Ondrejik	NWS Middle Atlantic RFC
2	Laura Belanger	NWS Peachtree City GA
2	Amanda Schroeder	NWS Fort Worth TX
2	Jeff Waldstreicher	NWS Eastern Region HQ
2	Britt Westergard	NWS Albany NY
3	Anthony Anderson	NWS Arkansas-Red Basin RFC
3	Greg Hanson	NWS Burlington VT
3	Scott Lincoln	NWS Lower Mississippi RFC
3	Scott Watson	NWS Pleasant Hill MO
3	Jeff Zogg	NWS Des Moines IA
4	Ray Christensen	NWS Elko NV
4	Tom Clemmons	NWS Flagstaff AZ
4	Chris Horne	NWS Greenville-Spartanburg SC
4	Jennifer Palukci	NWS Albuquerque

Table 7. HWT-Hydro staff

Name	Title	Affiliation	Week
J. J. Gourley	Principal Investigator	NOAA/OAR/NSSL	All
Elizabeth Argyle	Co-Investigator/Weekly Coordinator	OU/CIMMS	1; 4
Race Clark	Co-Investigator/Logistics Coordinator	OU/CIMMS	All
Zac Flamig	Co-Investigator/Technical Coordinator	OU/CIMMS	All
Brandon Smith	Co-Investigator/Weekly Coordinator	OU/CIMMS	All
Steve Martinaitis	Experiment Coordinator	OU/CIMMS	All
Gabe Garfield	HWT Coordinator	OU/CIMMS	All
Ami Arthur	Weekly Coordinator	OU/CIMMS	3
Jess Erlingis	Weekly Coordinator	OU	1; 4
Maria Moreno	Weekly Coordinator	OU	2

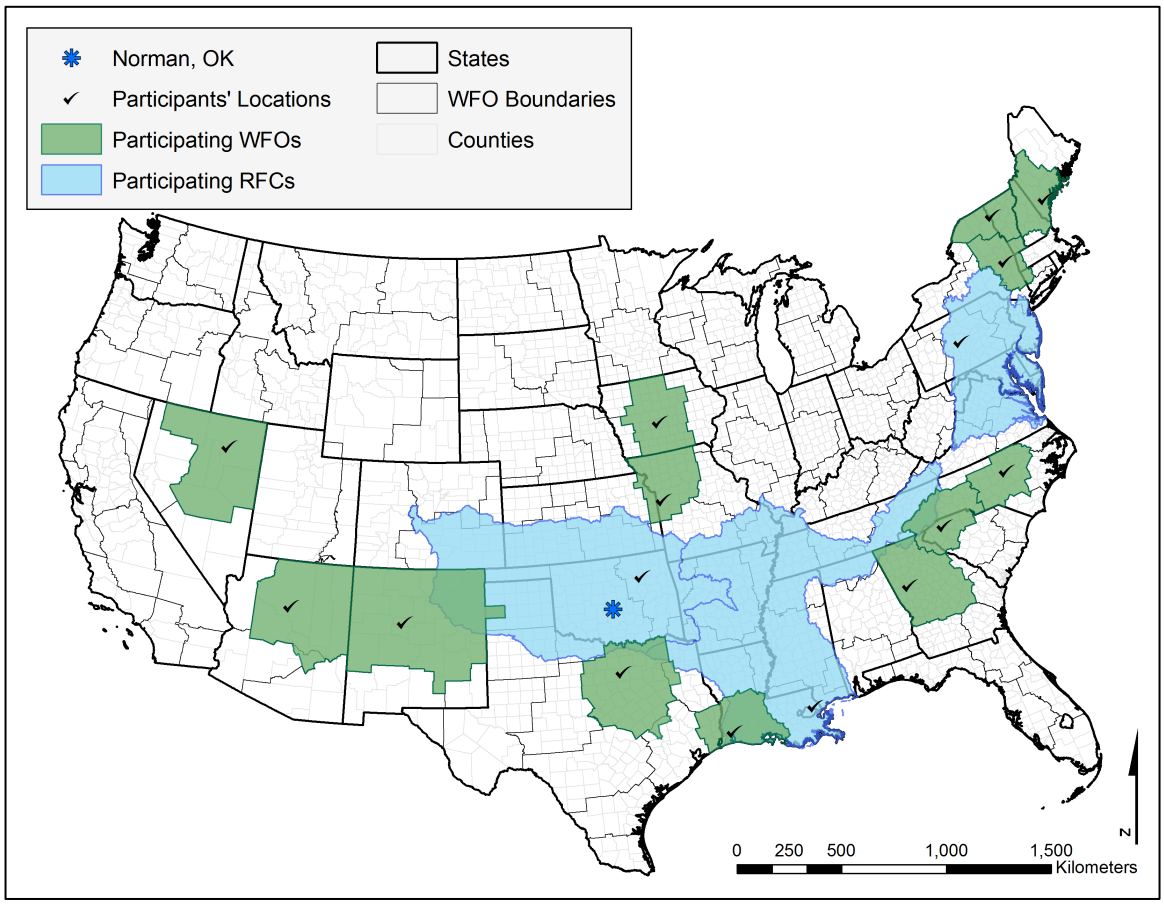

Figure 11. Map of HWT-Hydro participants' locations.

Appendix B: HWT-Hydro Survey Instruments

The evaluation survey instrument is reproduced in full below:

1.1 Please enter the forecast data (YYYYMMDD):
1.2 Please enter the forecast region you will be evaluating:
2.1 Of the flash flood observational data sets, rank from 1-4 (with [1] being the best) how the NWS local storm reports, mPING citizen-scientist reports, USGS streamflow, and SHAVE targeted public observations provide the most useful information about the **areal extent** of flash flooding. If two data sets provided the same information, then assign them the same ranking.
    a. Local storm reports
    b. mPING citizen scientist reports
    c. USGS streamflow
    d. SHAVE targeted public observations
2.2 Comments:
3.1 Of the flash flood observational data sets, rank from 1-4 (with [1] being the best) how the NWS local storm reports, mPING citizen-scientist reports, USGS streamflow, and SHAVE targeted public observations provide the most useful information about the **magnitude** of flash flooding. If two data sets provided the same information, then assign them the same ranking.
    a. Local storm reports
    b. mPING citizen scientist reports
    c. USGS streamflow
    d. SHAVE targeted public observations
3.2 Comments:
4.1 Of the flash flood observational data sets, rank from 1-4 (with [1] being the best) how the NWS local storm reports, mPING citizen-scientist reports, USGS streamflow, and SHAVE targeted public observations provide the most useful information about the **specific impacts** of flash flooding. If two data sets provided the same information, then assign them the same ranking.
    a. Local storm reports
    b. mPING citizen scientist reports
    c. USGS streamflow
    d. SHAVE targeted public observations
4.2 Comments:
5.1 Of the experimental flash flood monitoring and short-term prediction tools, rank from 1-4 (with

[1] being the best) how the MRMS QPE, QPE recurrence intervals, QPE-to-flash flood guidance ratios, and FLASH runoff recurrence intervals **detect the event** (consider hit/miss/false alarm). If two products provided the same information, then assign them the same ranking.

a. MRMS QPE
b. QPE recurrence interval
c. QPE-to-FFG ratio
d. FLASH runoff recurrence interval

5.2 Comments:

6.1 Of the experimental flash flood monitoring and short-term prediction tools, rank from 1-4 (with [1] being the best) how the MRMS QPE, QPE recurrence intervals, QPE-to-flash flood guidance ratios, and FLASH runoff recurrence intervals **accurately represent the spatial extent** of flooding. If two products provided the same information, then assign them the same ranking.

a. MRMS QPE
b. QPE recurrence interval
c. QPE-to-FFG ratio
d. FLASH runoff recurrence interval

5.3 Comments:

7.1 Of the experimental flash flood monitoring and short-term prediction tools, rank from 1-4 (with [1] being the best) how the MRMS QPE, QPE recurrence intervals, QPE-to-flash flood guidance ratios, and FLASH runoff recurrence intervals **reveal the magnitude** of flooding. If two products provided the same information, then assign them the same ranking.

a. MRMS QPE
b. QPE recurrence interval
c. QPE-to-FFG ratio
d. FLASH runoff recurrence interval

7.2 Comments:

8.1 How did the skill of the HRRR-forced FLASH compare to the QPE-forced FLASH? Consider detection, false alarming, spatial accuracy, and magnitude with the forecasts.
Select one: (Much worse, Worse, About the same, Better, Much better, n/a)

8.2 Comments:

9.1 Assess how much **lead time** was provided from the HRRR-forced FLASH compared to the QPE-forced FLASH. Consider detection, false alarming, spatial

accuracy, and magnitude with the forecasts. Mark a [-1] if the HRRR-based products led to a degradation compared to the QPE-based products. Please enter your lead time estimate into the box:

9.2 Comments:

10.1 Using all available flash flood observations, rate the **spatial accuracy** of the experimental flash flood watches and warnings vs. those that were issued operationally.
Experimental watches were:
Select one: (Much Worse, Worse, About the Same, Better, Much Better, n/a)
Experimental warnings were:
Select one: (Much Worse, Worse, About the Same, Better, Much Better, n/a)

10.2 Comments:

11.1 Using all flash flood observations and tools, rate the **uncertainty estimate** that was given to the issued flash flood watches and warnings. Recall that a low probability event should occur about 25% of the time, a medium about 50% of the time, and a high about 75% of the time.
Uncertainty estimates in watches were:
Select one: (Too Low, About Right, Too High, n/a)
Uncertainty estimates in warnings were:
Select one: (Too Low, About Right, Too High, n/a)

11.2 Comments:

12.1 Using all flash flood observations and tools, rate the **magnitude (nuisance vs. major)** that was given to the issued flash flood watches and warnings. Major floods can be validated with reports of homes/buildings with water in them, homes/buildings/vehicles swept away, rescues, evacuations, injuries, or fatalities.
Magnitude estimates in watches were:
Select one: (Too Low, About Right, Too High, n/a)
Magnitude estimates in warnings were:
Select one: (Too Low, About Right, Too High, n/a)

12.2 Comments:

13.1 Using all available flash flood observations, rate the **lead time** of the experimental flash flood watches and warnings vs. those that were issued operationally.
Experimental flash flood watches were:
Select one: (Much worse, Worse, About the same, Better, Much better, n/a)
Experimental flash flood warnings were:

Select one: (Much worse, Worse, About the same, Better, Much better, n/a)

The system usability survey is reproduced in full below:

Please enter today's date:
Please enter your participant ID:

(for each, pick one: Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree)

1. I think that I would like to use the FLASH products frequently.
2. I found the FLASH products unnecessarily complex.
3. I thought the FLASH products were easy to use.
4. I think that I would need assistance to be able to use the FLASH products.
5. I found the various functions in the FLASH system were well integrated.
6. I thought there was too much inconsistency in the FLASH system.
7. I would imagine that most people would learn to use the FLASH products very quickly.
8. I found the FLASH system very cumbersome/awkward to use.
9. I felt very confident using this system.
10. I needed to learn a lot of things before I could get going with the FLASH products.

The feedback survey is reproduced in full below:

1. The Experiment Introduction on Monday Afternoon:
   (for each, pick one: Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree)
   a. The introduction helped me to understand experimental flash flood products sufficiently.
   b. I understood the anticipated outcomes and methodology after the presentations.
   c. The introduction was effective in giving me more familiarity with AWIPS2 and procedure loading.
2. With regard to each activity, the time allotted to each was:
   (for each, pick one: Far too Little, Too Little, About Right, Too Much, Far too Much)
   a. Introduction Session

b. Experimental issuance of flash flood watches and warnings
c. Evaluation and discussion of the prior day's tools/watches/warnings
d. 'Tales from the Testbed' webinar

3. Please indicate your level of agreement or disagreement with the following statements:
(for each, pick one: Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree)
a. In the forecasting sessions, I was given the tools that I needed to issue flash flood watches and warnings.
b. The evaluation and discussion sessions helped me to improve my forecasts as the week progressed.

4. In terms of workload, please indicate the levels you felt across the whole week during each of the primary sessions:
(for each, pick one: Much lower than average, Somewhat lower than average, About average, Somewhat higher than average, Much higher than average)
a. Experimental issuance of flash flood watches/warnings (forecasting sessions)
b. Evaluation and discussion sessions
c. Webinar preparation session

5. Was the material provided before the experiment helpful in understanding and preparing for the experiment?
(select one: Not at all helpful, Somewhat helpful, Neutral, Somewhat helpful, Very helpful)

6. Would you consider participating in this experiment again in the future?
(select one: Yes, No, Undecided)

7. Would you recommend participating in this experiment to colleagues?
(select one: Yes, No, Undecided)

8. Comments or suggestions for improvement:

Appendix C: Experimental Tools Used in HWT-Hydro

Table 8. List of experimental tools available in AWIPS II

Tool Name (as shown in AWIPS II)	Tool Category	Additional Versions Available	Units
CREST Max Return Period	Hydrologic Model		Year
HRRR-Forced CREST	Hydrologic Model		Year
CREST Soil Moisture	Hydrologic Model		%
CREST Streamflow	Hydrologic Model		$m^3 \bullet s^{-1}$
SAC-SMA Soil Moisture	Hydrologic Model		%
SAC-SMA Streamflow	Hydrologic Model		$m^3 \bullet s^{-1}$
MRMS Radar-Only QPE	QPE/QPF	4: Instantaneous rate, 1-, 3-, and 6-h	in or $in \bullet hr^{-1}$
HRRR QPF	QPE/QPF	3: 1-, 3-, and 6-h	in
MRMS Radar-Only QPE to FFG Ratio	FFG	4: 1-, 3-, 6-h, and maximum of any	%
HRRR QPF to FFG Ratio	FFG	3: 1-, 3-, and 6-h	%
Precipitation Return Period	Precipitation Return Period	6: 1-, 3-, 6-, 12-, 24-h, and maximum of any	Year
Precipitation Return Period (Forecast)	Precipitation Return Period	3: 1-, 3-, and 6-h	Year
Precipitable Water Analysis	Precipitable Water	2: RAOBs or RAP	in
Precipitable Water Standard Anomalies	Precipitable Water	2: RAOBs or RAP	Unitless
MRMS Quality-Controlled Composite Reflectivity	Radar		dBZ
MRMS Seamless Hybrid-Scan Reflectivity	Radar		dBZ

Appendix D: Experimental Watch/Warning Templates; Warngen

Below is Figure 12, an example of the experimental product text issued during the HWT-Hydro Experiment.

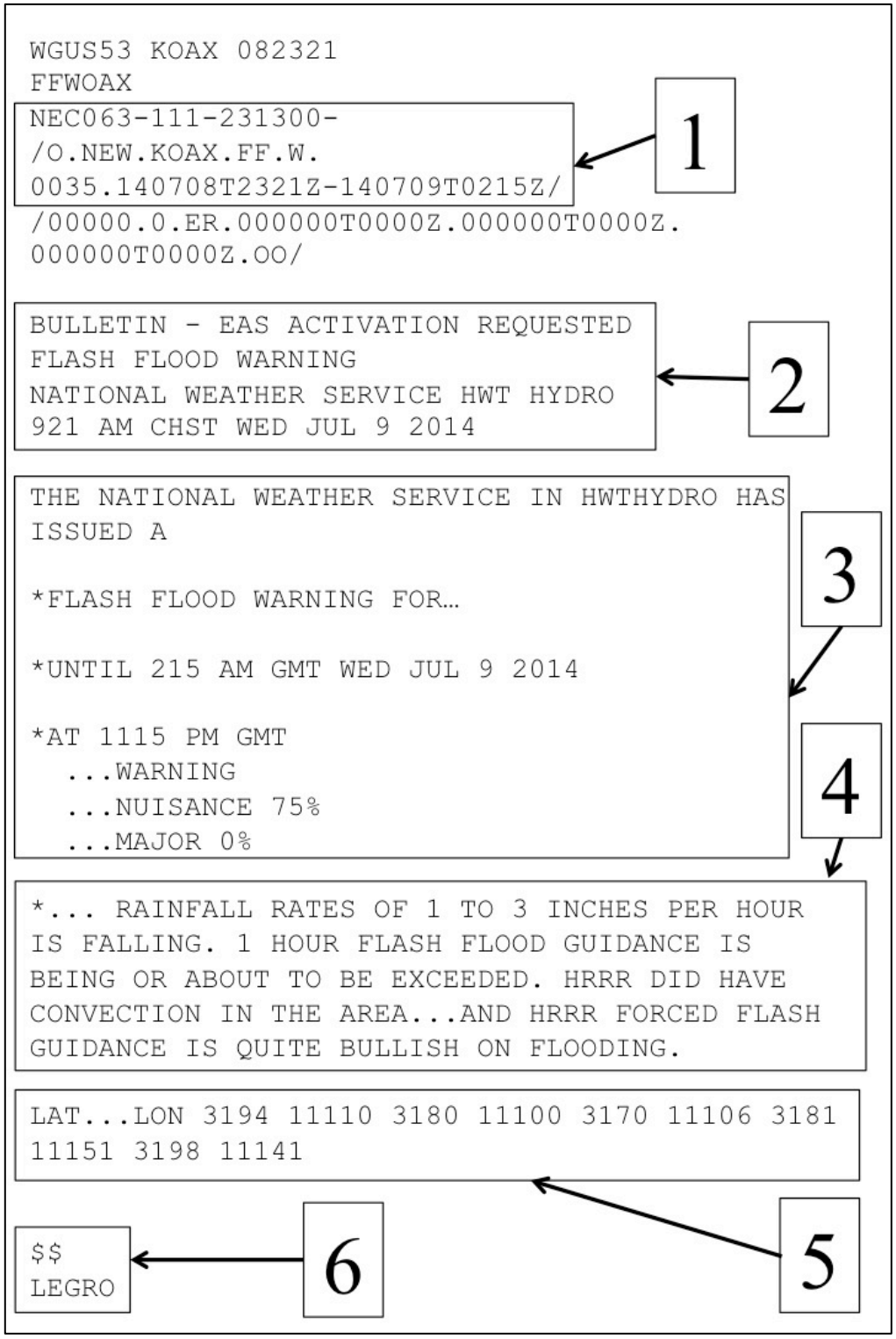

Figure 12. Example of experimental warning text.

Box 1 refers to the VTEC (Valid Time Event Code) header, which contains the product status type ("NEW"), the issuing office ("KOAX"), the phenomena type ("FF": flash flooding), the phenomena type code ("W": warning), the unique sequential product ID number ("0035"), and the time range in UTC for which the product is valid ("140708T2321Z-140709T0215Z": [Start year][Start month][Start day]T[UTC start time]Z-[End year][End month][End day]T[UTC end time]Z). Warngen generates the VTEC header automatically based upon the office where the product was generated and the start and end times selected by the forecaster. "OAX" is the three-letter WFO code corresponding to the Omaha, Nebraska office, which is where the initial deployment of AWIPS II took place. Therefore, research builds of the AWIPS II software, like in HWT-Hydro, tend to use this WFO as their default office.

Box 2 refers to the bulletin box of the warning text. The product type ("FLASH FLOOD WARNING") is generated automatically by Warngen based on the forecaster's selection of product type. The time and date ("921 AM CHST WED JUL 9 2014") is also generated automatically. A timezone code in the Warngen product template can be altered to change the timezone in the text as desired. (For the curious reader, "CHST" is UTC+10 and stands for "Chamorro Standard Time".)

Box 3 refers to the product description. For HWT-Hydro, the name of the issuing office was hardcoded to read "HWTHYDRO", but can be coded to change dynamically to correspond with the issuing office in the VTEC header. The first asterisked line is also hardcoded for HWT-HYDRO such that the text of both experimental watches and warnings say "FLASH FLOOD WARNING" on that line. The second asterisk contains the expiration time and date of the product in GMT (Greenwich Mean Time). This is also the first line over which the participants had any control; here they could select the following product valid times: 30 minutes, 1 hour, 90 minutes, 2 hours, 150 minutes, 3 hours, and 6 hours. The third asterisk marks the product type and the uncertainties assigned to each impact. Forecasters were required to select "WATCH" or "WARNING" from a dropdown menu in Warngen. Then they were required to select 25%, 50%, 75%, or 100% for probability of nuisance flooding and 0%, 25%, 50%, 75%, or 100% for probability of major flooding via radio buttons in Warngen.

Box 4 corresponds to the fourth asterisk. Forecasters were required to enter text in this box but were free to choose the content included. Forecasters were encouraged to type out their thoughts leading up to the decision to issue an experimental warning or watch.

Box 5 is automatically generated by Warngen and consists of latitude and longitude pairs in decimal degrees multiplied by one hundred. Finally, forecasters were required to sign their products and could do so via a name, number, or some other alias (box 6). The format and content of the automatically-generated product text can be controlled by editing the Warngen templates (written in the Apache Velocity Java language) included in the AWIPS II source code (see Appendix G for more details).

Appendix E: Additional Objective Skill Results

In an effort to maintain a reasonable manuscript length, many analyses related to the objective skill of the experimental products are presented in this appendix:

- Fraction of operational watches containing at least one operational warning: 0.37
- Fraction of operational watches containing at least one experimental warning: 0.14
- Fraction of experimental watches with at least one operational warnings 0.55
- Fraction of experimental watches with at least one experimental warning: 0.73
- Average number of experimental warnings in an experimental watch: 3.1
- Average number of operational warnings in an operational watch: 2.0
- Average number of experimental warnings in an operational watch: 1.0
- Average number of operational warnings in an experimental watch: 2.8
- Fraction of experimental warnings outside experimental watches: 0.25
- Fraction of operational warnings outside operational watches: 0.62
- Fraction of operational warnings outside experimental watches: 0.57
- Fraction of experimental warnings outside operational watches: 0.73

Table 9. Watch and warning skill by observation dataset

		Watches		Warnings	
	N	**Op.**	**Exp.**	**Op.**	**Exp.**
CSI (LSRs)	250	0.40	0.39	0.45	0.18
CSI (Storm Data)	184	0.29	0.47	0.32	0.17
CSI (mPING)	35	0.12	0.21	0.07	0.03
CSI (SHAVE)	124	0.21	0.34	0.20	0.45
CSI (USGS)	9	0.04	0.05	0.01	0.01

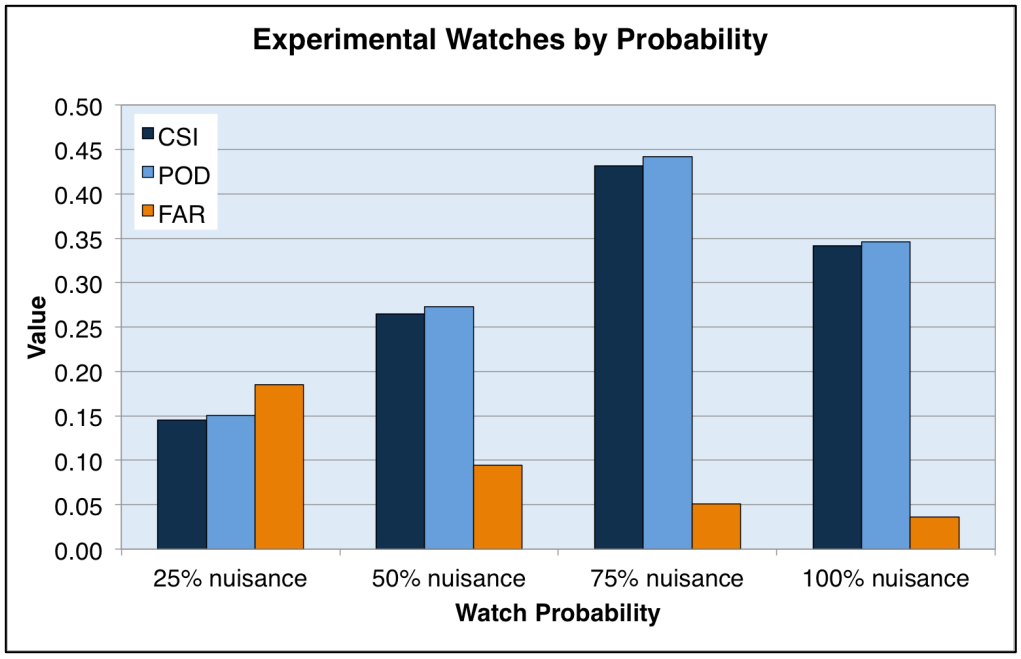

Figure 13. Skill of experimental watches for nuisance floods, segregated by assigned probability.

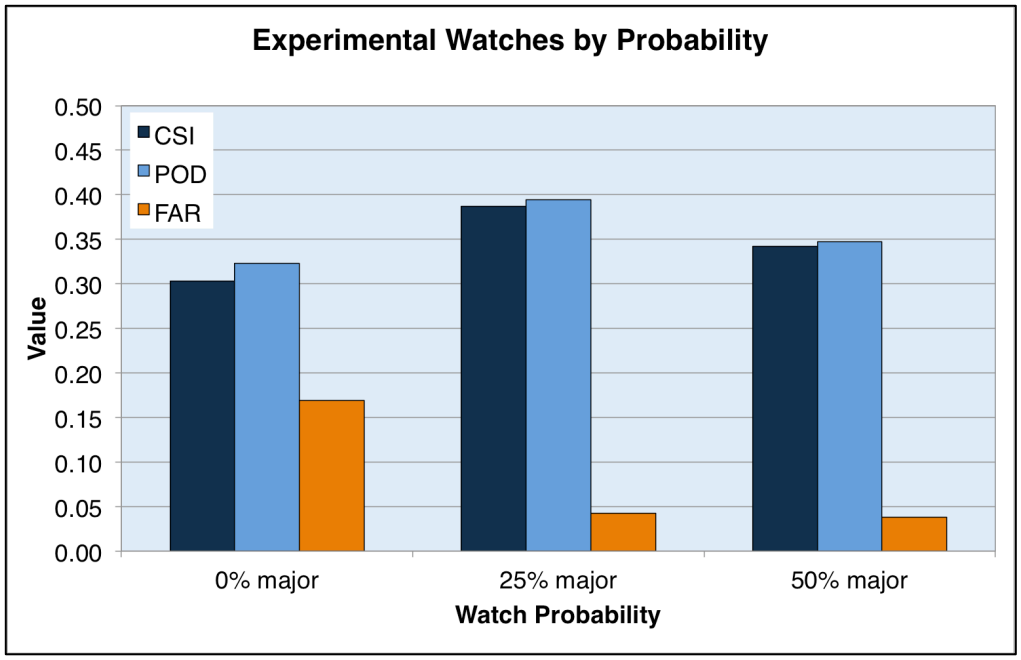

Figure 14. Skill of experimental watches for major floods, segregated by assigned probability.

The skill of experimental warnings by nuisance probability appears in the main text as Figure 6.

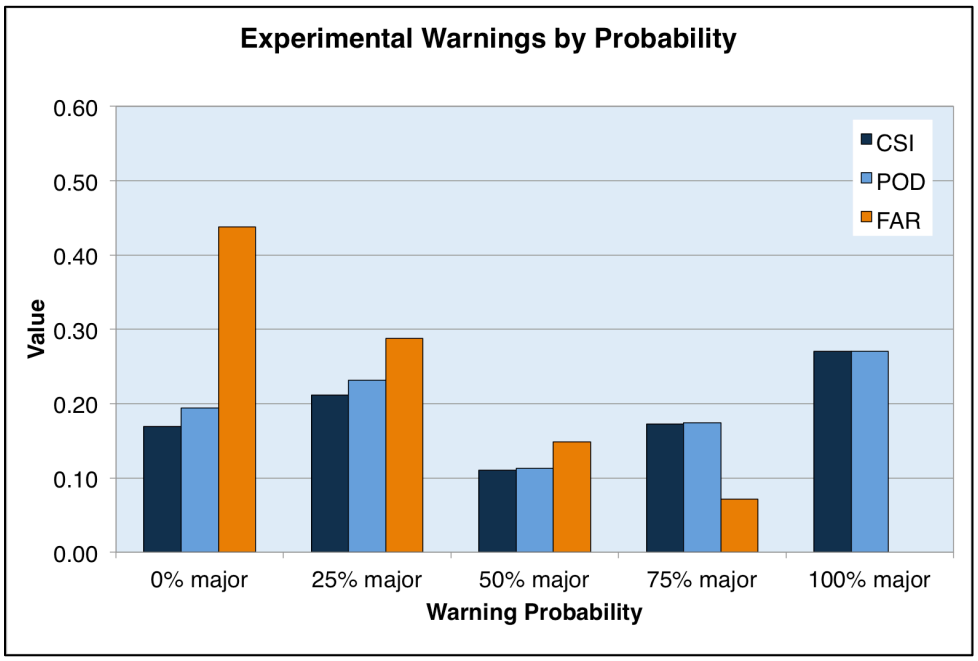

Figure 15. Skill of experimental warnings for major floods, segregated by assigned probability.

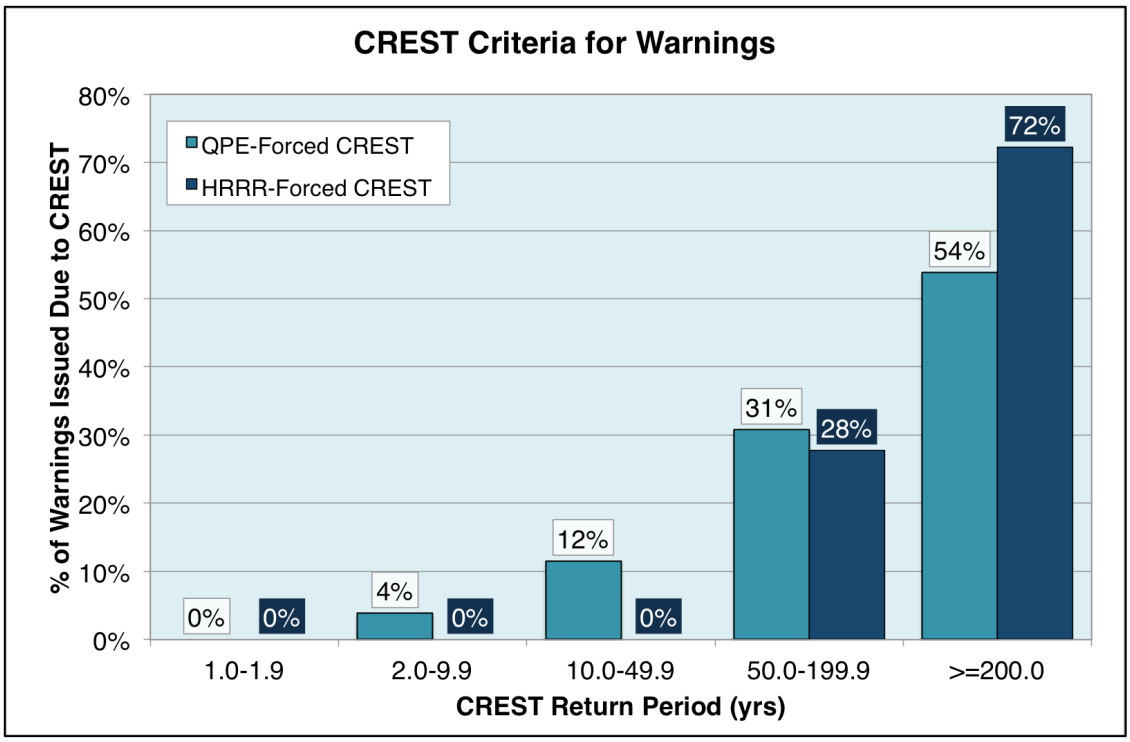

Figure 16. CREST model outputs associated with experimental warnings.

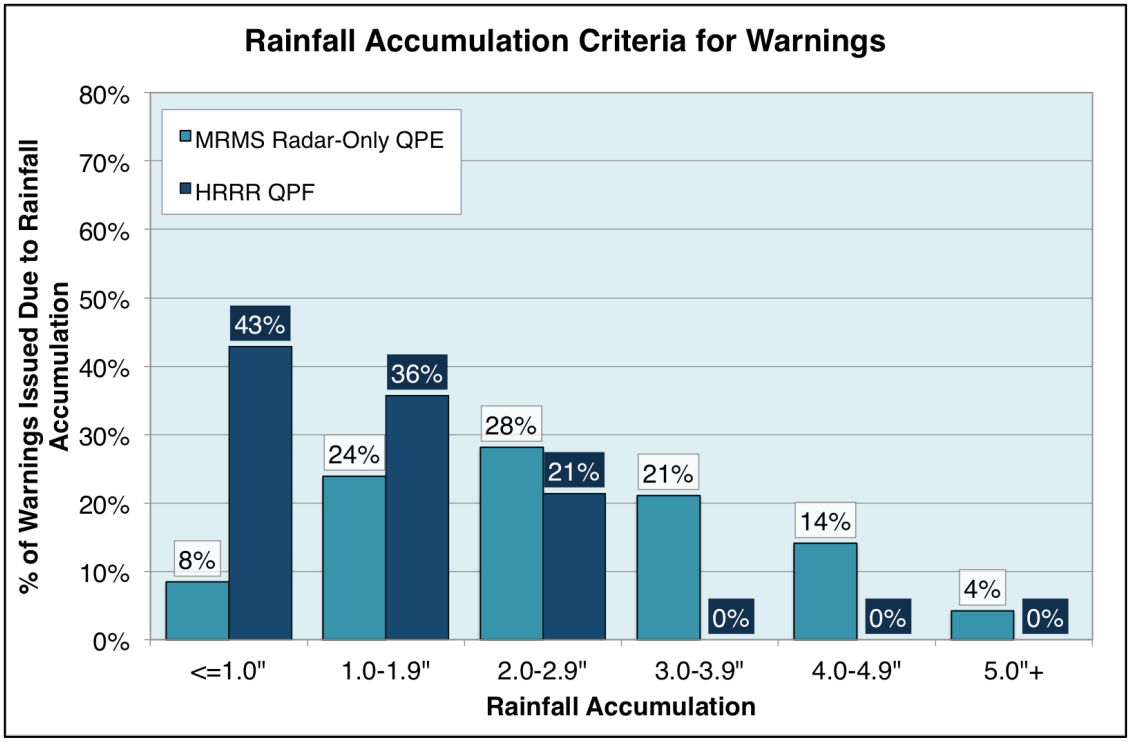

Figure 17. MRMS rainfall accumulations associated with experimental warnings.

In Table 10, the precipitable water anomaly is the number of standard deviations above the mean monthly precipitable water at a point.

Table 10. Other tool outputs associated with experimental watches and warnings.

Category	Percent of Products from that Category
Warnings: ARI	
1.0 – 1.9 yrs	1%
2.0 – 9.9 yrs	8%
10.0 – 49.9 yrs	22%
50.0 – 199.9 yrs	49%
>= 200.0 yrs	19%
Warnings: MRMS QPE-to-FFG Ratio	
<= 100%	18%
101% to 150%	58%
151% to 100%	23%
> 201%	9%
Warnings: Rainfall Rate	
< 1.0"/hr	9%
1.0 – 1.9"/hr	38%
2.0 – 2.9"/hr	34%
3.0 – 3.9"/hr	16%
>= 4.0"/hr	3%
Watches: Precipitable Water Anomaly	
<= 0.50 SD	3%
0.51 – 1.00 SD	3%
1.01 – 1.50 SD	6%
1.51 – 2.00 SD	44%
2.01 – 2.50 SD	29%
2.51 – 3.00 SD	9%
> 3.01 SD	6%

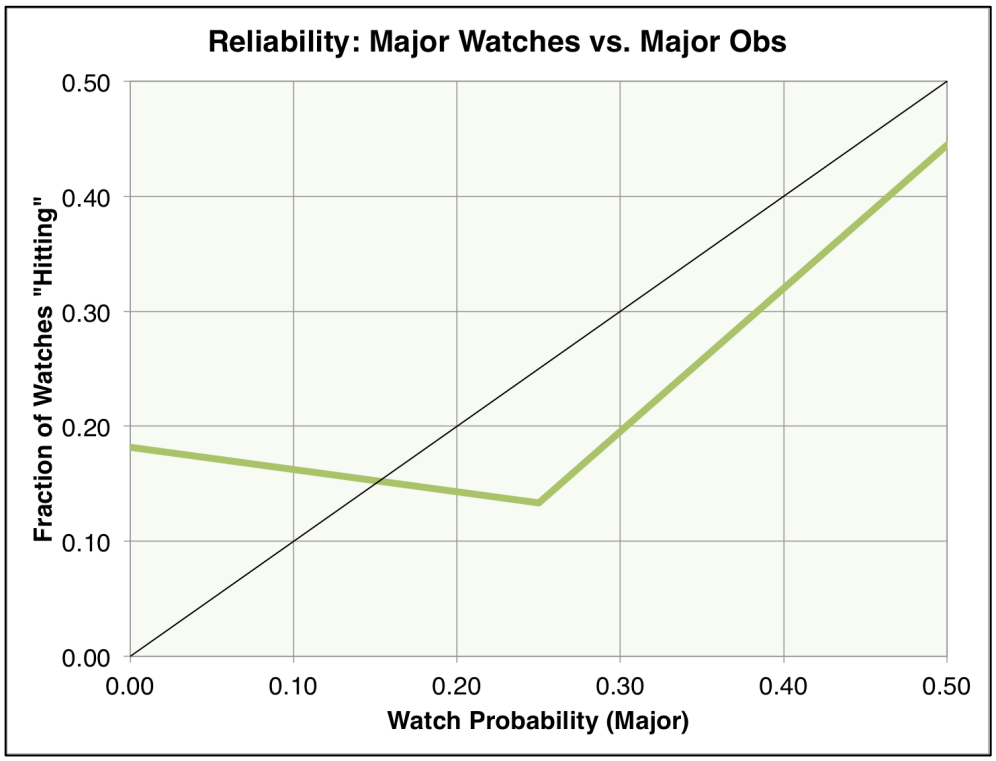

Figure 18. Reliability of probabilistic flash flood watches for major flash floods.

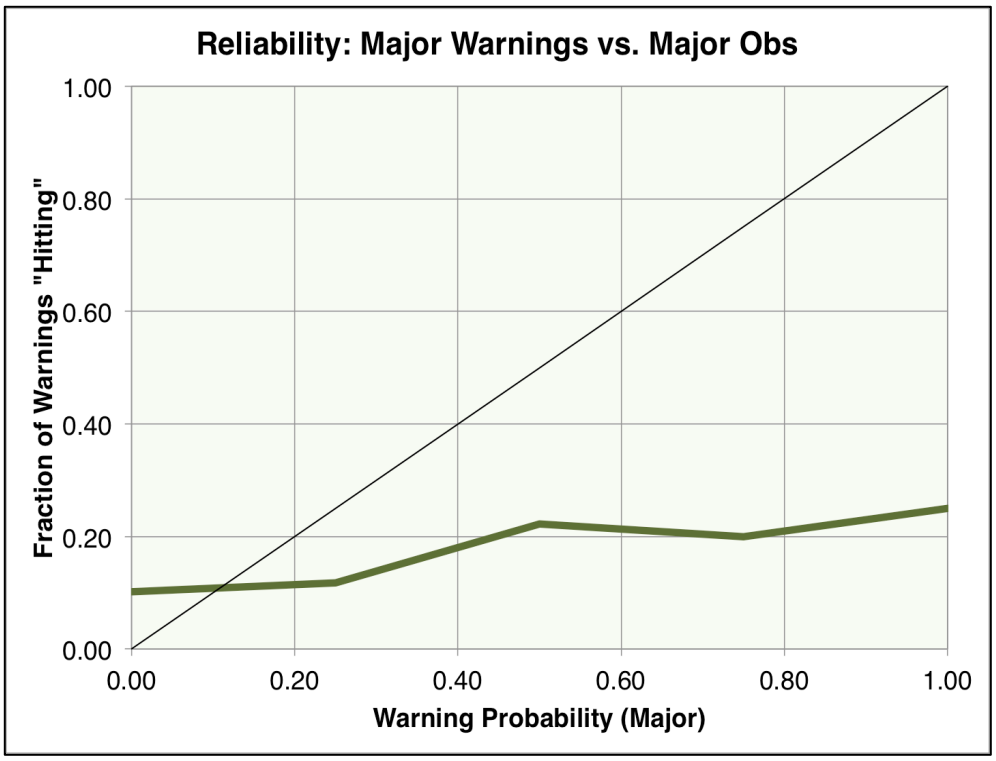

Figure 19. Reliability of probabilistic flash flood warnings for major flash floods.

Appendix F: Additional Survey Results

Five figures from the evaluation survey are presented below.

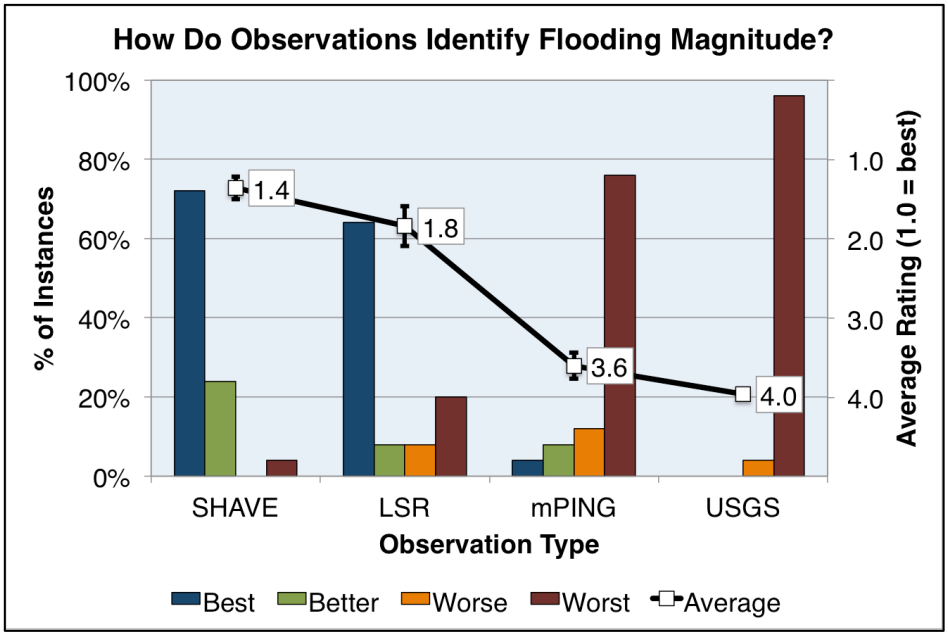

Figure 20. Subjective evaluation of flash flood observations in terms of identifying flood magnitude.

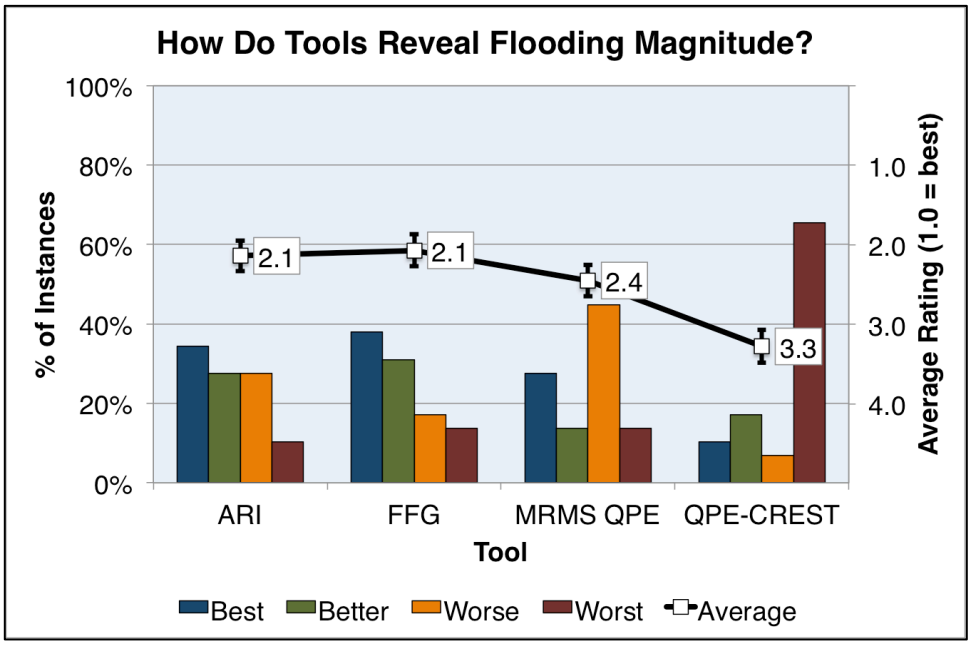

Figure 21. Subjective evaluation of flash flood tools in terms of identifying flood magnitude.

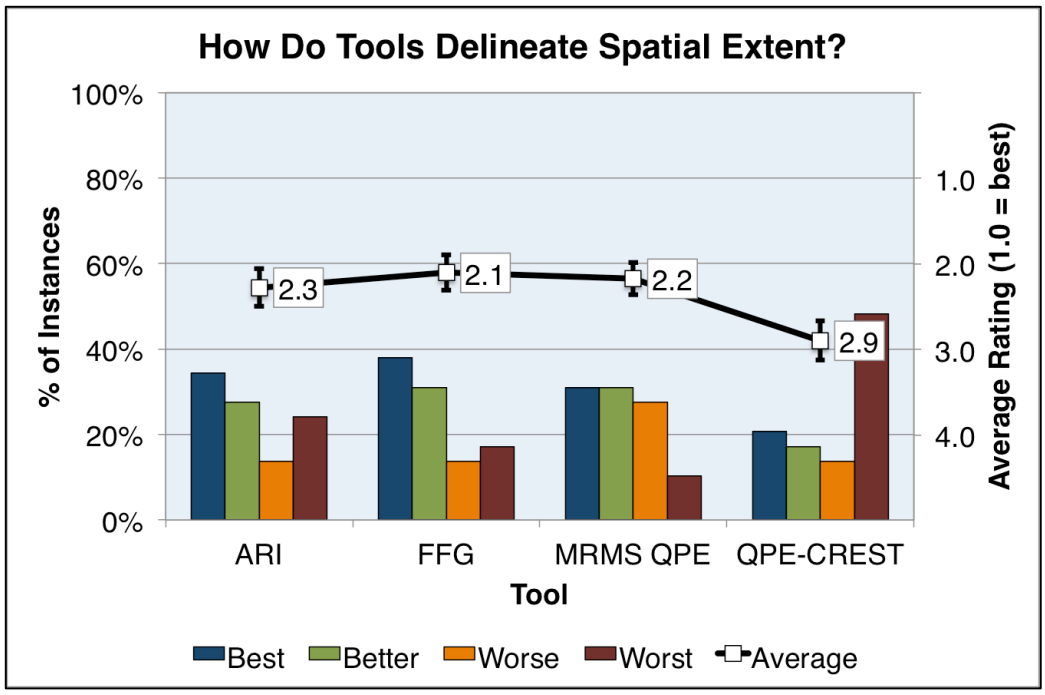

Figure 22. Subjective evaluation of flash flood tools in terms of delineating spatial extent.

In the feedback survey, participants were asked if they agreed with certain statements.

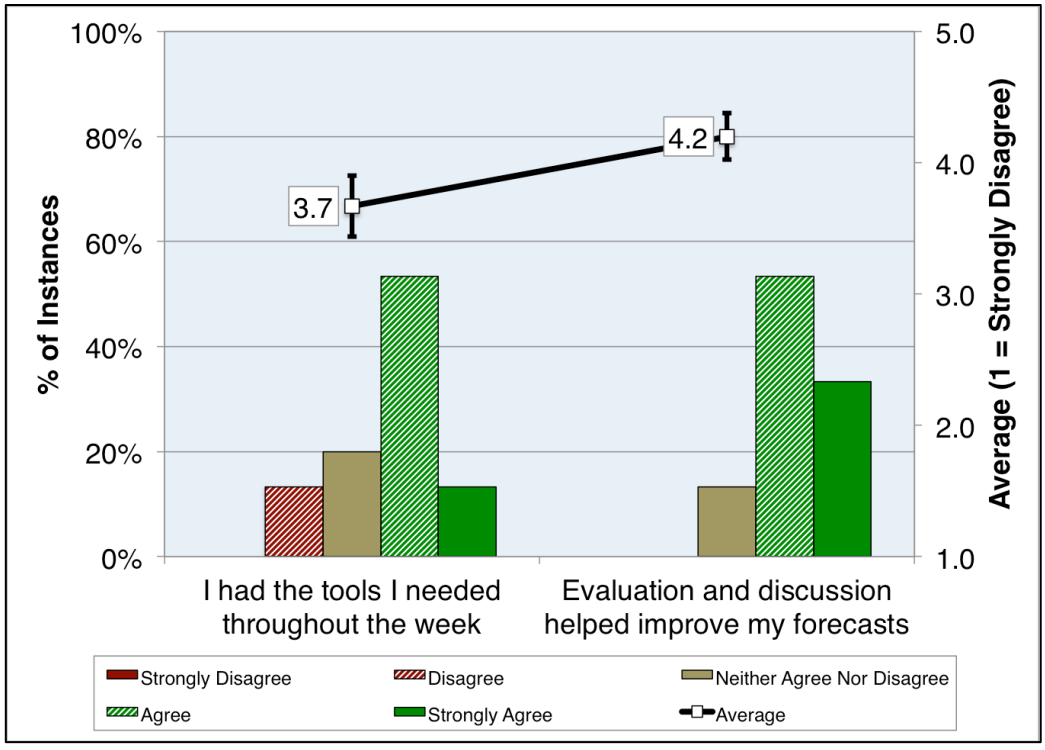

Figure 23. Selected results from feedback survey.

Finally, histograms from the beginning-of-week and end-of-week system usability surveys are presented below. Note the improvement in the summary statistics throughout the week.

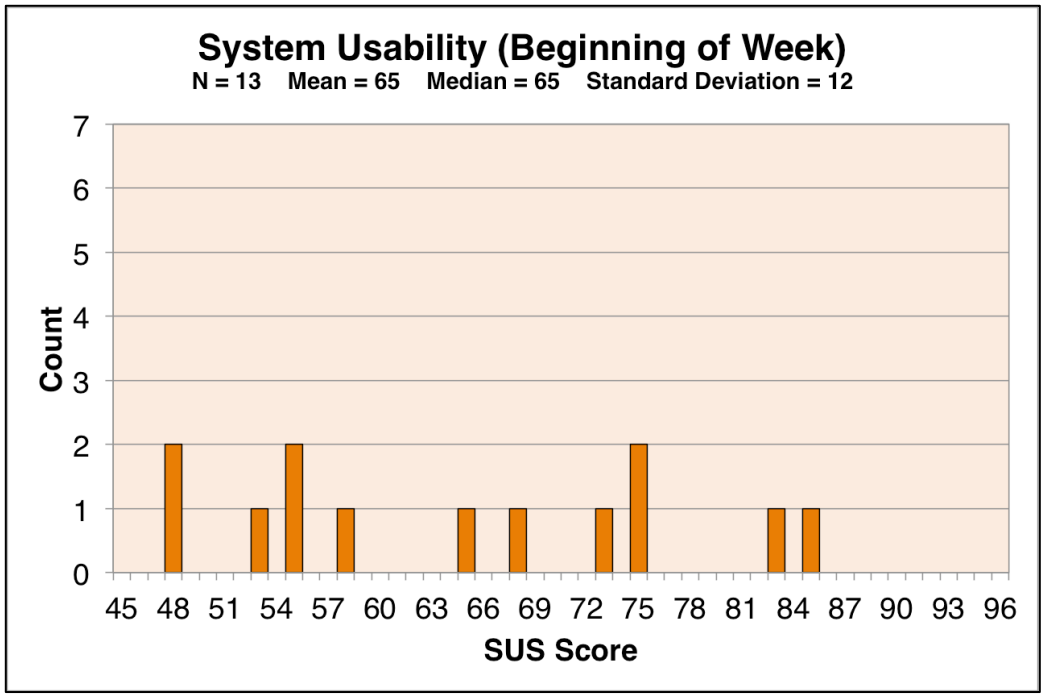

Figure 24. Histogram of System Usability Scores at beginning of week.

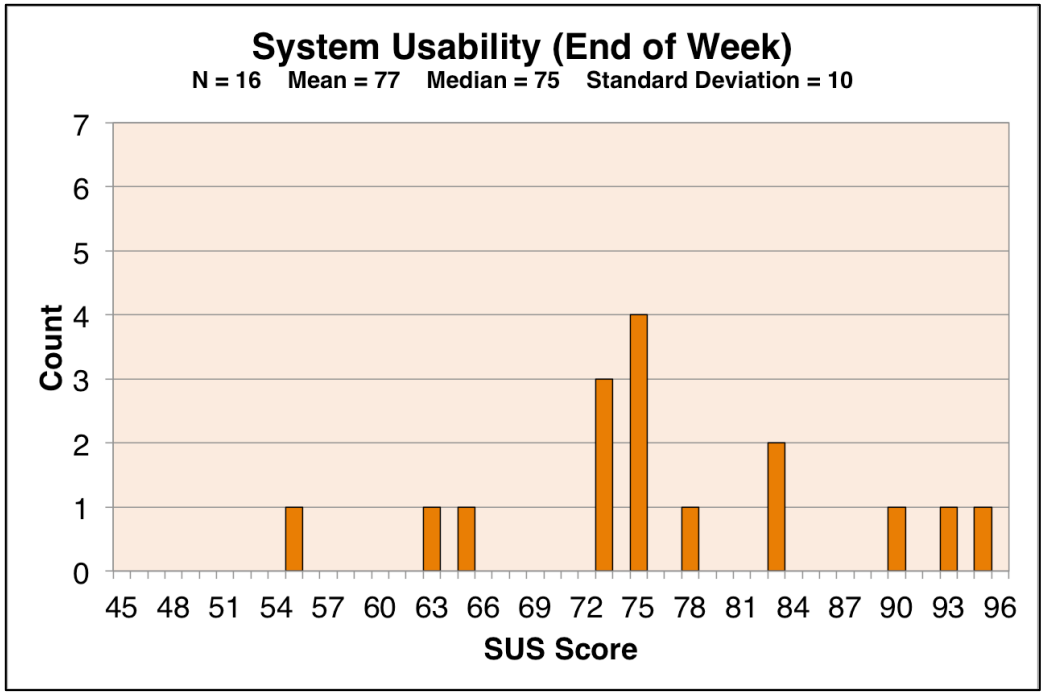

Figure 25. Histogram of System Usability Scores at end of week.

Appendix G: Tips for Displaying Tools in AWIPS-II

A substantial portion of the planning for the HWT-Hydro Experiment was devoted to developing the capability to display the FLASH suite in AWIPS II. AWIPS II consists of two main components: a display interface – CAVE (Common AWIPS Visualization Environment) –and the data server component – EDEX (Environmental Data Exchange System). Development of FLASH display capabilities was undertaken on a standalone (i.e., EDEX and CAVE running on the same computer) installation of AWIPS II version 14.1.1. Attempts were made at using version 14.2.X for HWT-Hydro but later versions of the code proved buggy and, due to time constraints, HWT-Hydro proceeded with the older version.

In the testbed, there were four CAVE workstations running Red Hat Enterprise Linux version 6 and a fifth computer acting primarily as an EDEX server but with CAVE enabled so that experiment staff could diagnose any problems reported by participants. These computers were equipped with 48 Gb of random access memory (RAM) and 16 Intel Xeon processing cores. One of the four CAVE workstations had only 16 Gb of RAM and was noticeably slower than the other three, according to participants. Some even referred to it as nearly unusable. The EDEX server was equipped with a conventional 1 Tb hard drive. Based on monitoring of the EDEX processor and memory usage, the hard drive acted as a speed bottleneck for most of the experiment, and solid-state or hybrid hard drives are recommended for EDEX servers whenever economically possible. (Additionally, forecasters suggested that at least two 22" diagonal monitors per workstation are desirable.) Several modifications to EDEX are possible to increase speed and reduce chances of extreme latency. FLASH tools were brought into AWIPS II as GRIB2 files during the experiment. AWIPS II plugin `com.raytheon.edex.plugin.grib.properties` was modified to read:

```
grib-decode.count.threads=10
```

The `com.raytheon.uf.edex.datadelivery.bandwidth.properties` plugin was modified as follows:

```
bandwidth.dataSetMetaDataPoolSize=6
bandwidth.retrievalPoolSize=12
bandwidth.subscriptionPoolSize=12
```

In `ingestGrib.sh`, the following modifications were made:

```
export INIT_MEM=1024 # in Meg
export MAX_MEM=8196 # in Meg

export JMS_POOL_MIN=4
export JMS_POOL_MAX=24
export METADATA_POOL_MIN=4
export METADATA_POOL_MAX=16
export EDEX_DEBUG_PORT=5007
export EDEX_JMX_PORT=1618
export MGMT_PORT=9603
```

In `request.sh`, the following modifications were made:

```
export INIT_MEM=128 # in Meg
if [ "$EDEX_ARCH" == "64-bit" ]; then
    export MAX_MEM=4096 # in Meg
```

```
else
    export MAX_MEM=1280 # in Meg
fi
export SERIALIZE_POOL_MAX_SIZE=24
export SERIALIZE_STREAM_INIT_SIZE_MB=2
export SERIALIZE_STREAM_MAX_SIZE_MB=8

export JMS_POOL_MIN=16
export JMS_POOL_MAX=32
export EDEX_DEBUG_PORT=5005
export EDEX_JMX_PORT=1616
export MGMT_PORT=9601
```

Finally, `default.sh` read as follows:

```
export INIT_MEM=512 # in Meg
export MAX_MEM=4096 # in Meg
export MAX_PERM_SIZE=128m
export EDEX_JMX_PORT=1616
export EDEX_DEBUG_PORT=5005
export JMS_POOL_MIN=64
export JMS_POOL_MAX=128
export METADATA_POOL_MIN=5
export METADATA_POOL_MAX=50
export DEBUG_PARAM_1=""
export DEBUG_PARAM_2=""
export DEBUG_PARAM_3=""
export DEBUG_PARAM_4=""
export PROFILER_PARAM_1=""
export PROFILER_PARAM_2=""
export PYPIES_MAX_CONN=50

export SERIALIZE_POOL_MAX_SIZE=16
export SERIALIZE_STREAM_INIT_SIZE_MB=2
export SERIALIZE_STREAM_MAX_SIZE_MB=6

export LOG4J_CONF=log4j.xml
export MGMT_PORT=9600
```

All CAVE workstations had their `cave.ini` files modified to avoid memory issues after several out-of-memory crashes in the first day of the experiment. In that file, the following lines were changed:

```
-Dthrift.stream.maxsize=200
-XX:MaxPermSize=256m
-Xmx4096m
```

These modifications permit AWIPS II to display high-resolution (1-km grid cell) grids that extend across the entire Lower 48. Although AWIPS II may still run more slowly than desired, these modifications are believed necessary to successfully load national FLASH and MRMS-Hydro grids.

In general, FLASH data can be stored in GeoTIFF format. Experiment staff wrote a

utility to automatically convert these GeoTIFF files into GRIB2 format, with headers corresponding to the properties described hereafter. The GRIB2 table properties file resides in the following location:

```
awips2/edex/data/utility/edex_static/base/grib/tables/
161/1/4.2.0.16.table
```

Inside this table should be a list of tools, line-by-line, using the following format: `grib_varID:grib_varID:product_menu_name:units:parameterID`, where `grib_varID` is an integer between 192 and 254. The `product_menu_name` should be identical to the string displayed for the tool's menu entry in CAVE. `Units` can take many formats, including "year", "%", "in", and "m^3*s^-1".

Next, `/awips2/edex/data/utility/edex_static/base/grib/models/` was modified to add `gribModels_FLASH.xml` which contains the following XML code:

```
<gribModelSet>
    <model>
        <title>FLASH</title>
        <name>FLASH</name>
        <center>161</center>
        <subcenter>1</subcenter>
        <process>
            <id>100</id>
        </process>
    </model>
</gribModelSet>
```

Then in `/awips2/edex/data/utility/edex_static/base/distribution/`, `grib.xml` was edited to add `<regex>FLASH</regex>` after the pre-existing entries. Modifications to Warngen templates (see Appendix D) were desired, and the files to do so exist in this location:

```
awips2/edex/data/utility/common_static/site/[WFO_code]/
warngen/
```

Finally, inside `/awips2/edex/data/utility/cave_static/user/[username]/`, the `styleRules` directory contains a file named `gridImageryStyleRules.xml`, which should be modified to add the parameters from the GRIB2 table described above. This file also contains information about the units to be displayed in CAVE (which should correspond to those in the GRIB2 table), the colormap to be used in CAVE, the type of scale used in the display (e.g., linear or logarithmic), the minimum and maximum values corresponding to the beginning and end of the colormap, the values to be explicitly displayed on the colormap legend in CAVE, and whether smoothing will be turned on by default for the tool.

CAVE menus are controlled by the `menus` directory. For HWT-Hydro, a subdirectory `FLASH` was created and inside it, `index.xml`, which contained the following text:

```
<menuContributionFile>
    <include
    installTo="menu:org.eclipse.ui.main.menu?after=n
    cephydro" filename="menus/FLASH/flash.xml"/>
```

```
</menuContributionFile>
```

Then that `flash.xml` file in the `menus` directory contains the actual organization and text that will appear within CAVE. Colormaps are defined inside the `colormaps` directory. Each colormap is stored a .cmap file with the following format:

```
<color r="____" g="____" b="____" a="____" />
```

where each line is a color in the RGB system, such that "r" is between zero and one and represents a normalized value of the red component of a color in the RGB system (and "g" and "b" correspond to green and blue, respectively). "a" stands for alpha and represents, on a zero to one scale, the degree of transparency to be used in CAVE for that color, where zero corresponds to full transparency and one corresponds to no transparency. CAVE automatically interpolates between lines in the .cmap file, so the more lines, the smoother the color transitions. CAVE limits .cmap files to 8,192 lines. Experiment staff created a Python script to automate colormap creation.

Finally, in the `bundles` directory an XML file must be created that corresponds to the bundle name used in the XML file establishing the CAVE menus. This bundle file sets the default CAVE values for density, magnification, brightness, contrast, transparency (unless otherwise specified in the colormap), and blinking.

Data enters EDEX from an LDM server. Depending on the tools desired, the LDM configuration files should be modified according to that program's instructions, at all times remembering that filenames and other properties must remain consistent with the properties fed to CAVE and EDEX.