2017 Flash Flood and Intense Rainfall Experiment

June 19 - July 21, 2017

Weather Prediction Center

College Park, MD

Findings and Results

Sarah Perfater - *I.M. Systems Group, NOAA/NWS/WPC/HMT*
Benjamin Albright - *Systems Research Group, NOAA/NWS/WPC/HMT*

Table of Contents

September 29, 2017

1. Introduction

The National Centers for Environmental Prediction (NCEP) Weather Prediction Center (WPC) issues Excessive Rainfall Outlooks (ERO) probabilistically identifying regions where rainfall totals may exceed NWS River Forecast Center Flash Flood Guidance (FFG) over Days 1, 2 and 3 (72 hours). Additionally, WPC operates the MetWatch Desk which is responsible for issuing Mesoscale Precipitation Discussions (MPDs): short-term (1-6 hours), event-driven forecasts that highlight regions where heavy rainfall may lead to flash flooding. While the goal of the ERO is to provide information about flooding rain potential several days in advance, MPDs are designed to enhance near-term situational awareness among local NWS offices, the media, and emergency managers.

In an effort to support the advancement of research to WPC and NWS field operations, the Hydrometeorology Testbed at WPC (HMT-WPC) continues to partner with NWS meteorologists, hydrologists, and the development and research communities to conduct the Flash Flood and Intense Rainfall (FFaIR) Experiment.

The 2017 experiment focused on the use of high resolution guidance to improve flash flood forecasts in both the short range (6-12 hours) and at longer time scales (48-72 hours). To simulate the flow of information that occurs from a national center (e.g. WPC) to the local forecast offices, this year's experiment attempted to engage the Science and Operations Officer (SOO) community by hosting a daily GoToMeeting and teleconference to discuss the experimental guidance and forecast activities with the goal of enhancing flash flood situational awareness and collaborate on possible experimental flash flood watch issuance.

2. Science and Operations Goals

The 2017 experiment provided a real-time pseudo-operational environment in which participants from across the weather enterprise could work together to explore the utility of emerging model guidance and tools for improving flash flood forecasts. This year's experiment again emphasized the rapid incorporation of the latest observational and model guidance into the decision making process while also challenging participants to simulate the collaboration that occurs between the national centers and local forecast offices during flash flood events.

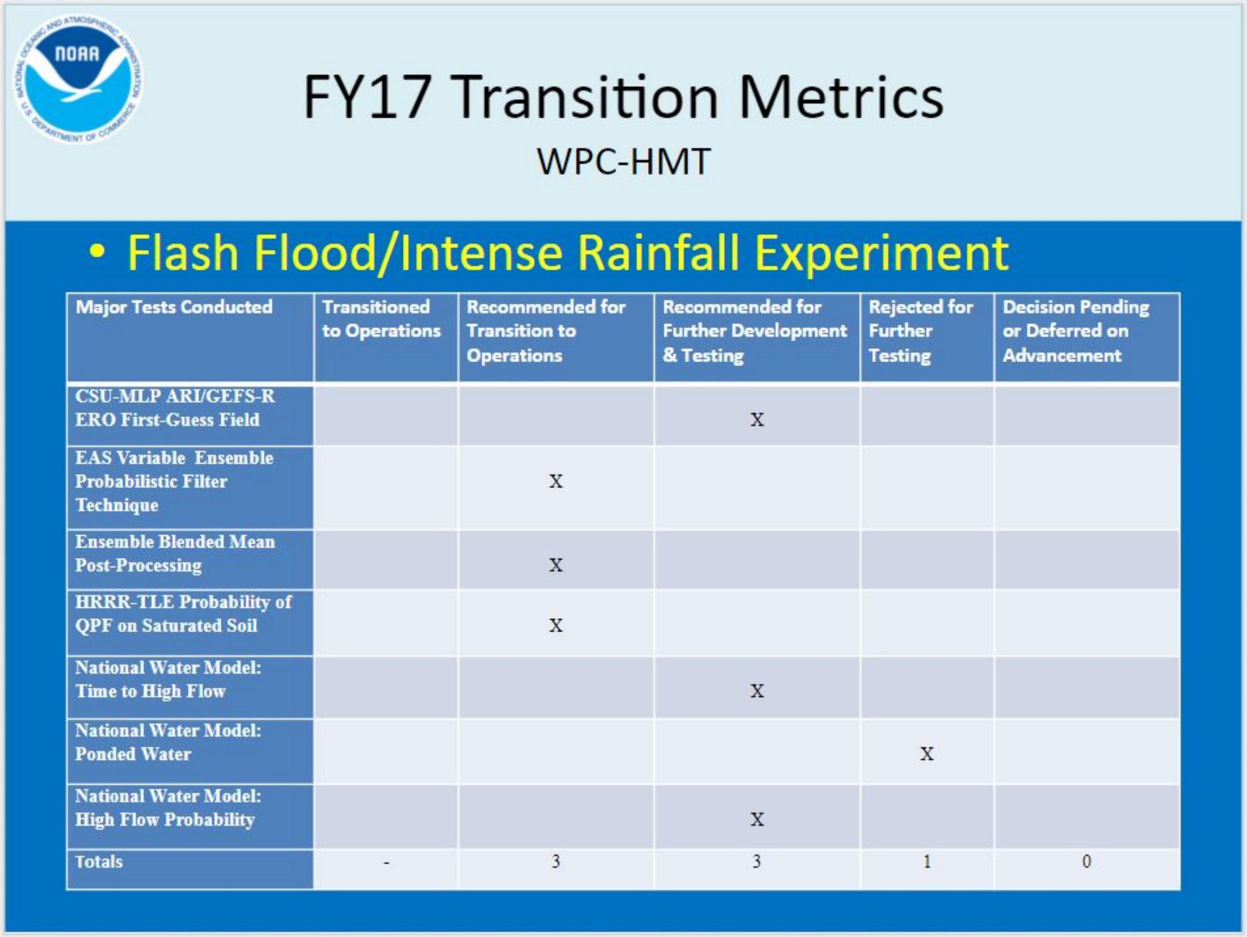
The goals of the 2017 Flash Flood and Intense Rainfall Experiment were to:

- Identify ways to maximize the utility of high resolution convection-allowing models and ensembles for short-term flash flood forecasts.
- Evaluate the utility of high resolution convection-allowing deterministic models for flood forecasts at longer time ranges (Day 2-3).
- Identify effective forms and proper usage of available hydrologic guidance for the assessment of flood risk.
- Explore proposed changes to WPC's operational Excessive Rainfall Outlook by evaluating the utility of an experimental "first guess" recommender for Day 2 and Day 3

respectively.

- Enhance collaboration between the operational forecasting, research, and academic communities on the forecast challenges associated with short-term flash flood forecasting.

Table 1. *Research to Operations Transition Metrics for the 2017 FFaIR Experiment*



FY17 Transition Metrics

WPC-HMT

- Flash Flood/Intense Rainfall Experiment

Major Tests Conducted	Transitioned to Operations	Recommended for Transition to Operations	Recommended for Further Development & Testing	Rejected for Further Testing	Decision Pending or Deferred on Advancement
CSU-MLP ARI/GEFS-R ERO First-Guess Field			X		
EAS Variable Ensemble Probabilistic Filter Technique		X			
Ensemble Blended Mean Post-Processing		X			
HRRR-TLE Probability of QPF on Saturated Soil		X			
National Water Model: Time to High Flow			X		
National Water Model: Ponded Water				X	
National Water Model: High Flow Probability			X		
Totals	-	3	3	1	0

3. Experiment Operations

Forecast Activities

The experiment was conducted for four weeks beginning June 19, 2017 in the WPC-OPC Collaboration Room at the NOAA Center for Weather and Climate Prediction (NCWCP) in College Park, MD:

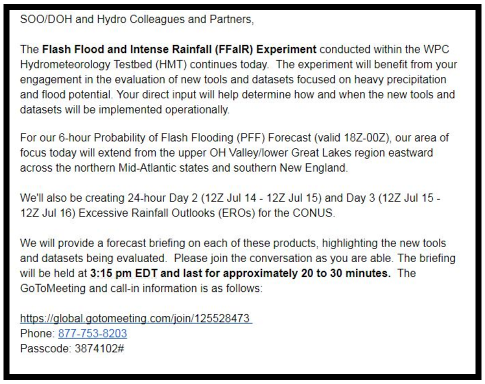
Week 1: June 19 – 23, 2017 (Monday – Friday)
Week 2: June 26 – 30, 2017 (Monday – Friday)
Week 3: July 10 – 14, 2017 (Monday – Friday)
Week 4: July 17 – 21, 2017 (Monday – Friday)

Each morning, participants were paired with a WPC forecaster as part of a collaborative forecast team and were tasked to create a Day 1, short-term, Probability of Flash Flooding (PFF) forecast valid for six hours from 18-00Z using probability contours of 15% (slight), 30% (moderate), and 50% (high) that conveyed the likelihood of flash flooding occurring within 40 km of a point. The forecast was created over a limited domain within the contiguous United States (CONUS). The PFF exercise helps support improvements to the WPC MetWatch Desk operations by using experimental guidance in a shorter, 6 hour timeframe.

During the afternoon, participants utilized longer range, high-resolution, convection-allowing models (CAMs) along with other guidance to create a Day 2 Excessive Rainfall Outlook (ERO), which is defined as the probability of flooding rains occurring within 40 kms of a point. The Day 2 ERO was valid over the whole CONUS for 24 hours beginning the next day (12Z to 12Z, 36-60 hour forecast) and used probability contours of 5% (marginal), 15% (slight), 30% (moderate) and 50% (high) to convey the risk. After issuing the Day 2 ERO, participants had the option to issue an experimental Flood Watch over the same Day 2 period that could be valid between 6 and 24 hours. Participants then issued a Day 3 ERO, defined the same as the Day 2 but valid for 60-84 hours. Lastly, participants utilized the Advanced Weather Interactive Processing System (AWIPS) II to analyze the beginning period of their PFF forecast area by using various real and near-real time models and tools.

SOO/DOH/WFO Daily Briefing

Each day an email message (a sample shown in Figure 1) was distributed to the SOO, Development and Operations Hydrologist (DOH), and associated partner communities inviting all to join the afternoon teleconference call including the limited domain focus of the 6-hour forecast (valid 18-00Z) and other highlights. This call was paired with a GoToMeeting PowerPoint presentation built throughout the day by the participants showing the three experimental forecasts (the PFF and Day 2 and 3 EROs) and a sampling of the experimental guidance that supported those forecasts. The briefings were designed to simulate collaboration between a national center and the WFO field offices when developing and communicating flash flood forecasts and potential flood watches.



SOO/DOH and Hydro Colleagues and Partners,

The **Flash Flood and Intense Rainfall (FFaIR) Experiment** conducted within the WPC Hydrometeorology Testbed (HMT) continues today. The experiment will benefit from your engagement in the evaluation of new tools and datasets focused on heavy precipitation and flood potential. Your direct input will help determine how and when the new tools and datasets will be implemented operationally.

For our 6-hour Probability of Flash Flooding (PFF) Forecast (valid 18Z-00Z), our area of focus today will extend from the upper OH Valley/lower Great Lakes region eastward across the northern Mid-Atlantic states and southern New England.

We'll also be creating 24-hour Day 2 (12Z Jul 14 - 12Z Jul 15) and Day 3 (12Z Jul 15 - 12Z Jul 16) Excessive Rainfall Outlooks (EROs) for the CONUS.

We will provide a forecast briefing on each of these products, highlighting the new tools and datasets being evaluated. Please join the conversation as you are able. The briefing will be held at **3:15 pm EDT and last for approximately 20 to 30 minutes.** The GoToMeeting and call-in information is as follows:

https://global.gotomeeting.com/join/125528473
Phone: 877-753-8203
Passcode: 3874102#

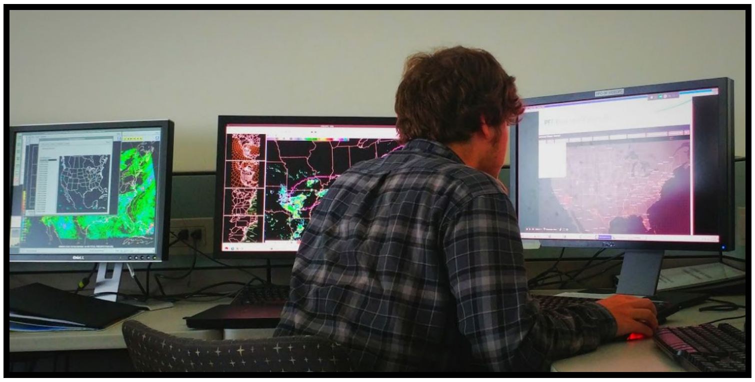
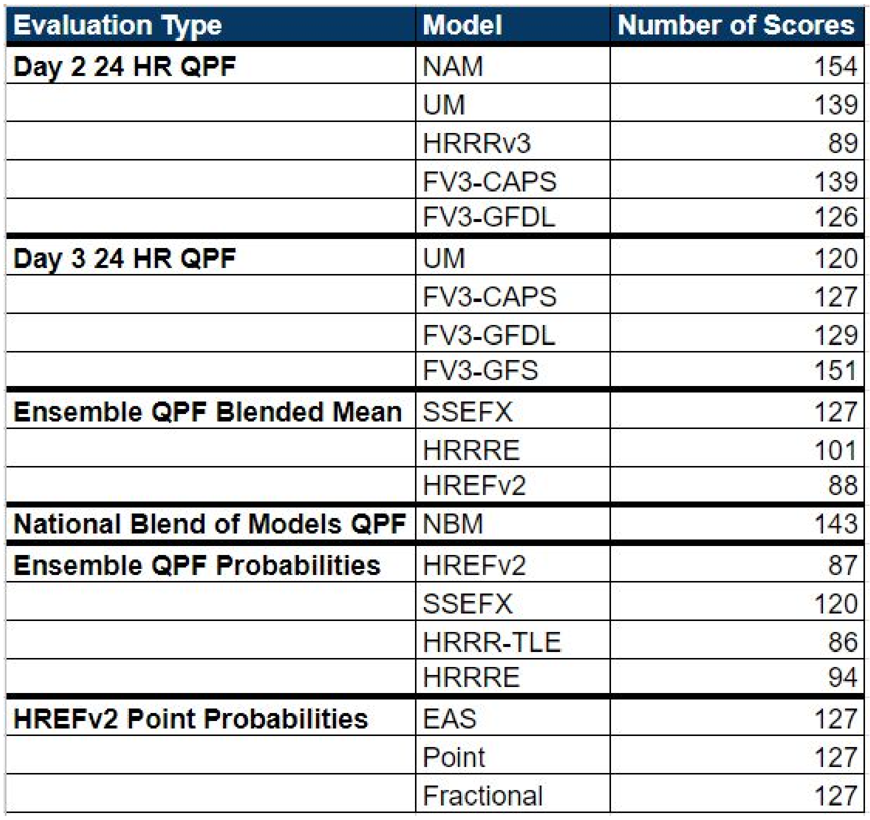


Figure 1. *Left: A sample of the email distributed to the SOO/DOH communities. Right: A FFaIR Experiment participant volunteer delivers the forecast briefing via teleconference call and GoToMeeting.*

Verification

Participants started each day with subjective evaluation which consisted of 11 science questions presented to the participants by the testbed staff. These questions included evaluation of the experimental FFaIR forecasts as well as other experimental models and tools used during the forecasting process. Participants used white boards to rank each experimental guidance, tool, or forecast on a scale from 1 (very poor) to 10 (very good). Individual scores were then recorded and averaged to arrive at one rounded score for the question, however, all of the individual scores were used and considered for all statistics. New this year, individual model and ensemble names were removed from the titles of the evaluation graphics so the participants did not know which model they were assigning scores to each day. This was done to attempt to remove preconceived biases that a participant might hold for any particular model. Table 2 shows the science questions and their associated number of subjective scores. The total number of scores reflects both model availability and the number of participants providing scores.

Table 2. *The number of total science question subjective scores provided by experiment participants for each available.*



Evaluation Type	Model	Number of Scores
Day 2 24 HR QPF	NAM	154
	UM	139
	HRRRv3	89
	FV3-CAPS	139
	FV3-GFDL	126
Day 3 24 HR QPF	UM	120
	FV3-CAPS	127
	FV3-GFDL	129
	FV3-GFS	151
Ensemble QPF Blended Mean	SSEFX	127
	HRRRE	101
	HREFv2	88
National Blend of Models QPF	NBM	143
Ensemble QPF Probabilities	HREFv2	87
	SSEFX	120
	HRRR-TLE	86
	HRRRE	94
HREFv2 Point Probabilities	EAS	127
	Point	127
	Fractional	127

A number of verification resources were used to score the experimental forecasts. Figure 2 displays an example of how the PFF was evaluated. Panel (1) is the PFF valid from 18 UTC June 23 to 00 UTC June 24, 2017 with flash flood local storm reports (LSRs), flood LSRs, mobile Precipitation Identification Near the Ground (mPING) reports, specially screened United States

Geological Survey (USGS) gauge reports, as well as flash flood warning polygons issued by local WFOs during the valid time of the forecast. Panel (2) shows the Multi-Radar, Multi-Sensor (MRMS) quantitative precipitation estimate (QPE) over the forecast period. The MRMS QPE used in the experiment included estimates from radar sources only. Panel (3) displays the practically perfect analysis which creates a neighborhood probabilistic forecast based on the flash flood reports received serving as a representation of what the forecast should have been if the forecaster had prior knowledge of where the reports would be located. Panel (4) shows the areas where the MRMS QPE exceeded the flash flood guidance (FFG).

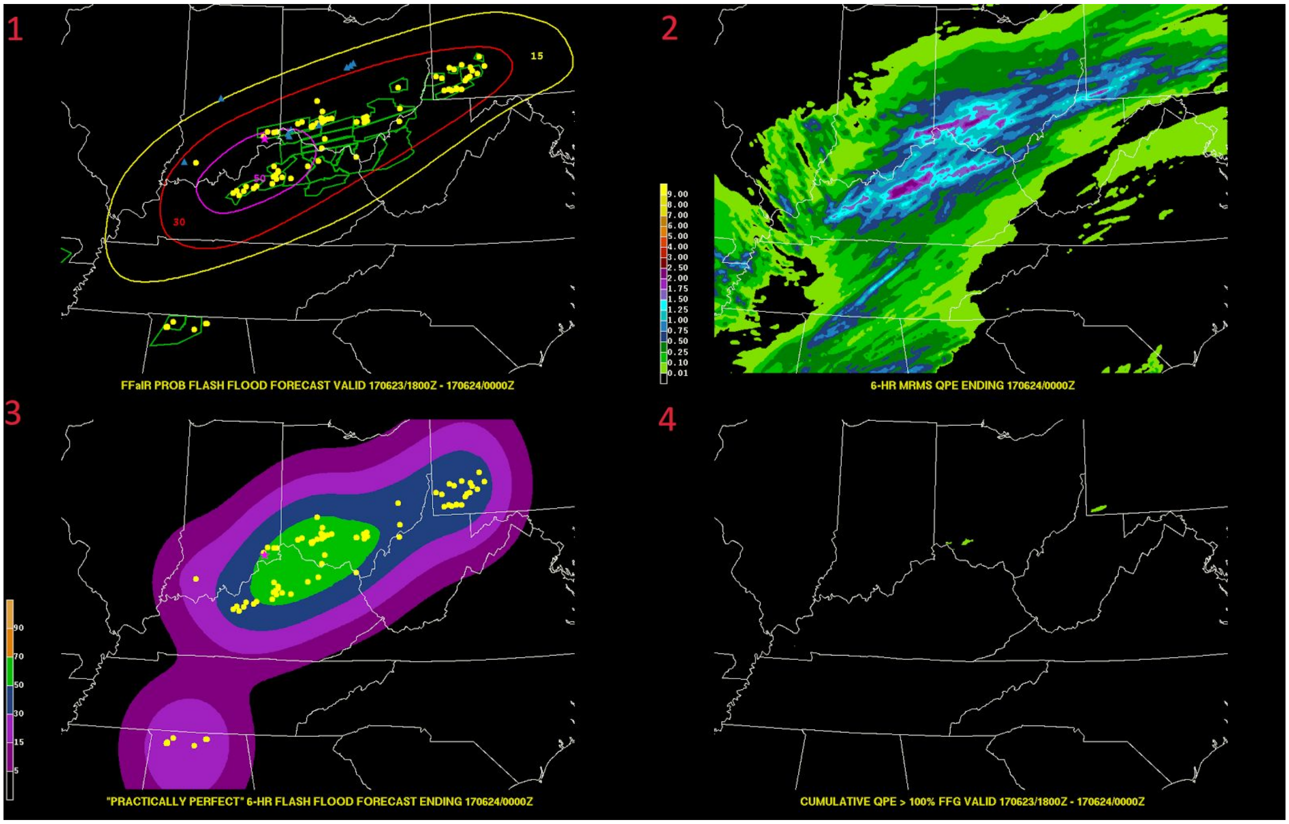

1
15
30
50
FFaIR PROB FLASH FLOOD FORECAST VALID 170623/1800Z - 170624/0000Z
2
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
6-HR MRMS QPE ENDING 170624/0000Z
3
90
70
50
30
15
5
"PRACTICALLY PERFECT" 6-HR FLASH FLOOD FORECAST ENDING 170624/0000Z
4
CUMULATIVE QPE > 100% FFG VALID 170623/1800Z - 170624/0000Z

Figure 2. *Verification for the PFF forecast valid 18 UTC June 23 -- 00 UTC June 24. (1) Display of the forecast contours with flash flood and flood reports and operational flash flood warnings overlaid. (2) MRMS QPE valid over the forecast period. (3) Practically perfect analysis. (4) Areas where MRMS QPE exceeded flash flood guidance.*

WPC operational EROs are verified by areas where precipitation exceeds FFG. For the ERO forecasts issued in FFaIR, a new verification method called the Unified Flood Verification (UFV) system was used in an effort to expand beyond just using FFG to verify the ERO. The UFV system uses a combination of FFG, average recurrence intervals, and observations and reports and applies a 40 km radius to each point that is considered a hit in the three different categories and combines all three onto one map. An example of the UFV is shown in Figure 3 with a Day 2 ERO overlaid, valid 12 UTC June 23 to 12 UTC June 24, 2017. Using FFG, areas where 1, 3, and 6 hour QPE exceeded FFG during the 24 hour period are considered hits. Anywhere the 5 year, 24 hour average recurrence interval (ARI) was exceeded over the 24 hour period is considered a hit. Finally, anywhere that a flash flood LSR, flood LSR, mPING, or USGS gauge report was reported is also considered a hit. This was the first year using the UFV for ERO verification in FFaIR. All model QPF evaluations were verified using the MRMS radar-only QPE.

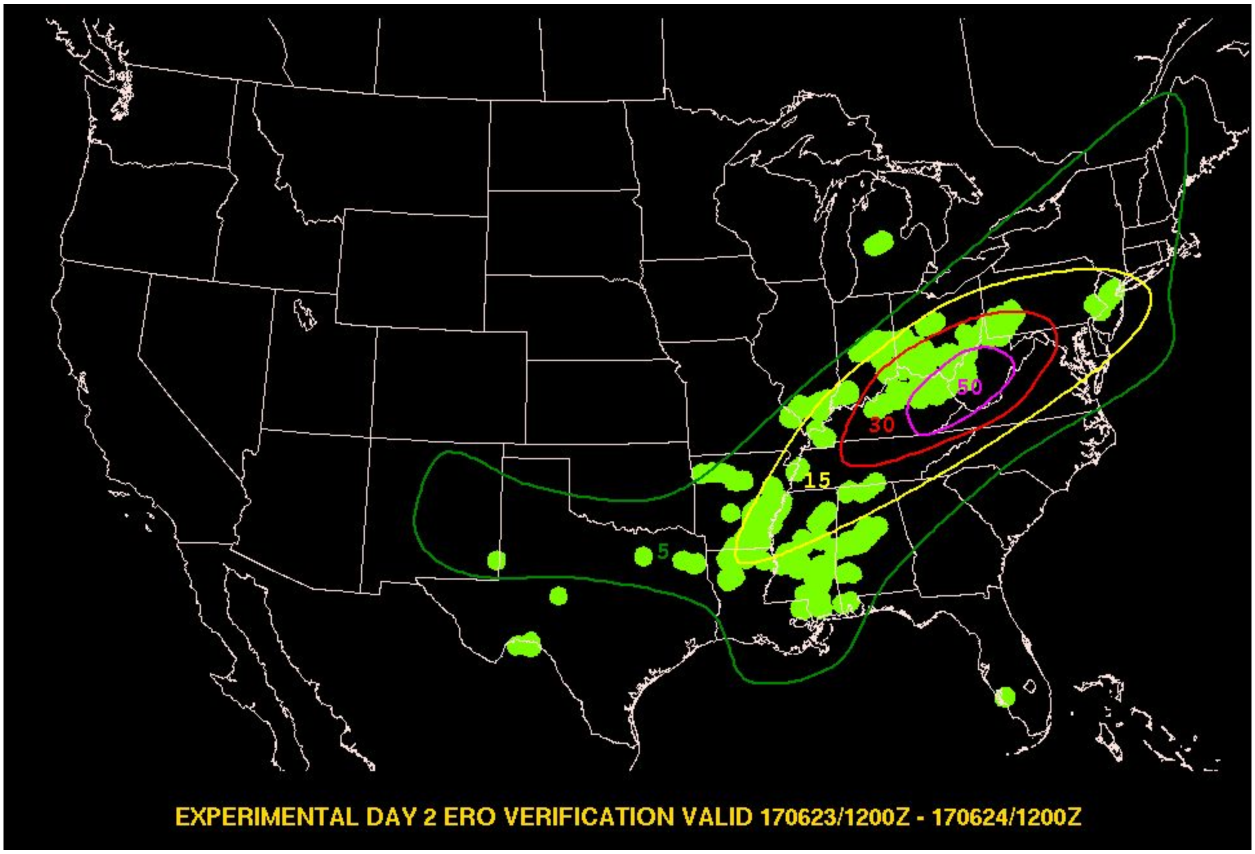
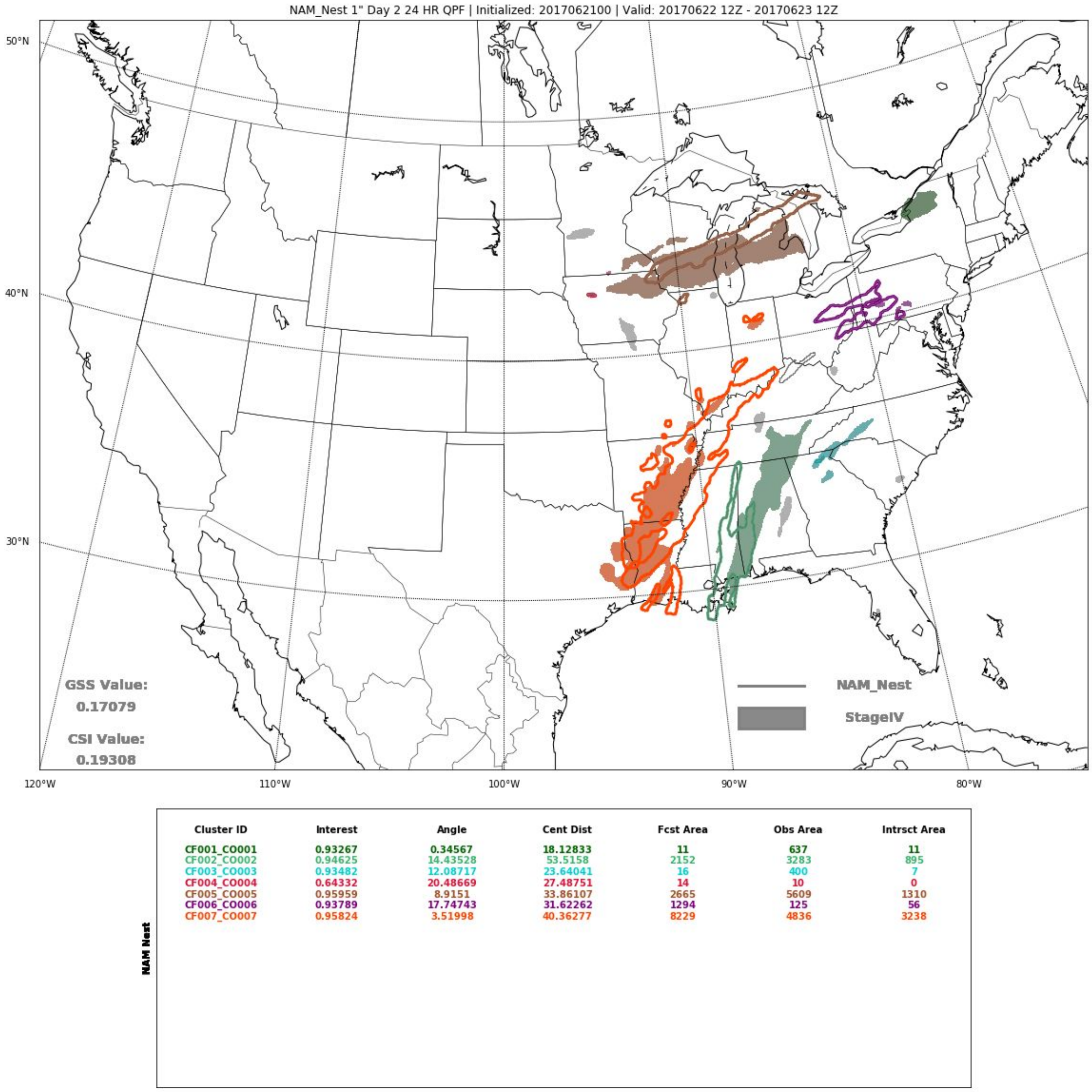

50
30
15
5
EXPERIMENTAL DAY 2 ERO VERIFICATION VALID 170623/1200Z - 170624/1200Z

Figure 3. *The Unified Flood Verification system (green dots) which includes 1-, 3-, 6-hour QPE exceeding FFG, 5 year 24 ARI exceedance, and flash flood and flood LSRs, mPING reports, and USGS gauge reports, where all hits have a 40 km neighborhood radius filter applied. The Day 2 ERO valid 12 UTC June 23 to 12 UTC June 24, 2017 is overlaid.*

After the subjective verification, the Method for Object-Based Diagnostic Evaluation (MODE) was used to compare various forecasted QPF thresholds from several models to Stage IV QPE (see Appendix C for WPC MODE settings). MODE outputs various statistics comparing the forecasted objects (model QPF) to the observed objects (Stage IV QPE) including centroid distance, angle, and intersection area. The Gilbert Skill Score (GSS) and critical success index (CSI), commonly referred to as Equitable Threat Score and Threat score respectively, were also computed over the whole domain for several models. All model QPF and QPE were re-gridded to a common 5km grid with a CONUS mask applied. An example of the MODE verification for 60 hour forecast from the 3 km NAM Nest of 24 hour QPF at the 1 inch threshold valid at 12 UTC June 23, 2017 is shown in Figure 4. The overall performance of select models were tracked on a daily basis as well as cumulatively throughout the entire experiment using Roebber Performance Diagrams (Roebber, 2009), pictured in Figure 5. A Roebber Performance Diagram provides a way to visualize a number of measures of forecast quality including probability of detection, false alarm ratio, bias, and critical success index in a single diagram.



Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.93267	0.34567	18.12833	11	637	11
CF002_CO002	0.94625	14.43528	53.5158	2152	3283	895
CF003_CO003	0.93482	12.08717	23.64041	16	400	7
CF004_CO004	0.64332	20.48669	27.48751	14	10	0
CF005_CO005	0.95959	8.9151	33.86107	2665	5609	1310
CF006_CO006	0.93789	17.74743	31.62262	1294	125	56
CF007_CO007	0.95824	3.51998	40.36277	8229	4836	3238

Figure 4. *MODE analysis for the 60 hour 3 km NAM Nest forecast for 24 hour QPF at the 1 inch threshold valid from 12 UTC June 22 to 12 UTC June 23, 2017.*

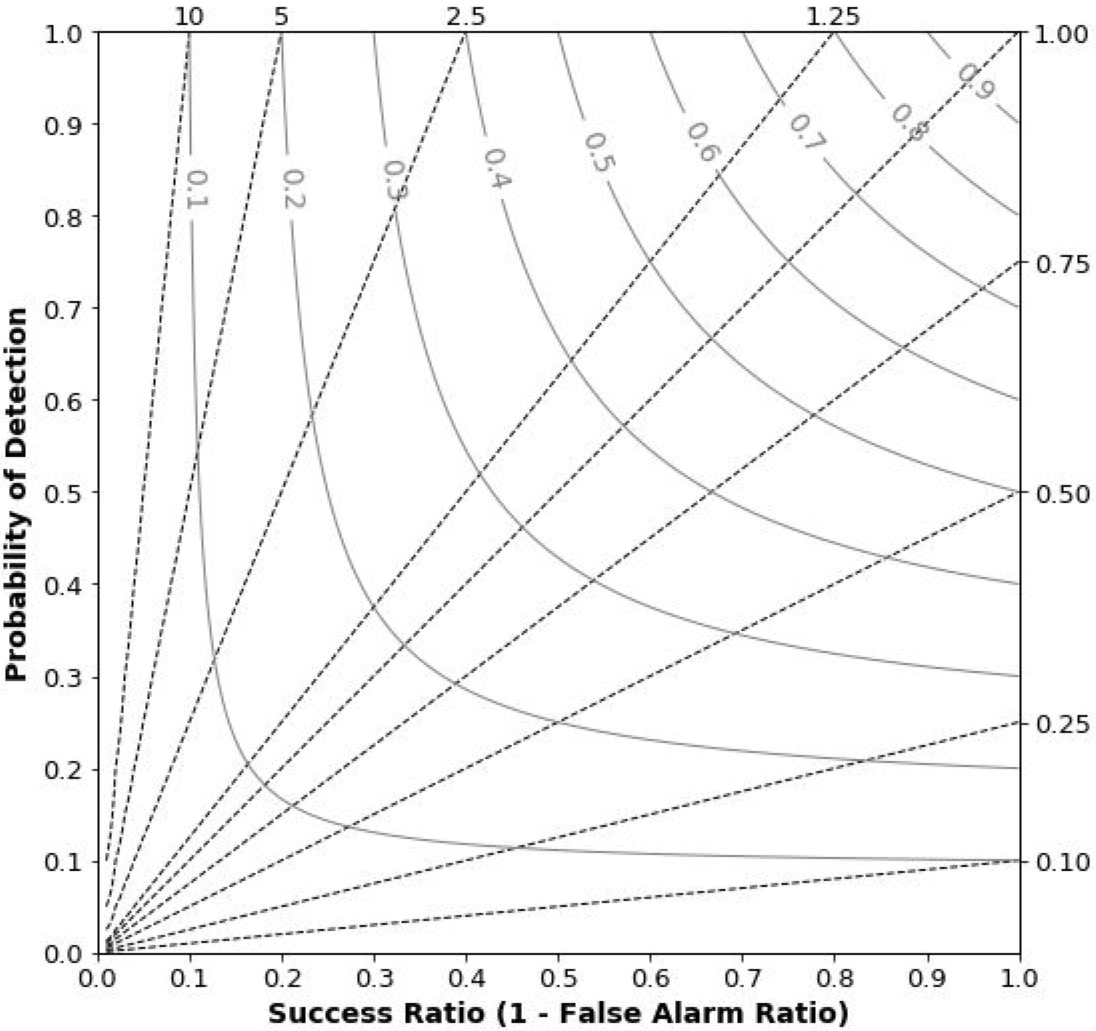


Figure 5. *Example of a Roebber Performance Diagram. Y-axis probability of detection, x-axis success ratio (1 - false alarm ratio), dashed diagonal lines represent the bias, and curved solid lines represent critical success index.*

Featured Guidance and Tools for Experimental Forecasts

In addition to the full multi-center suite of operational deterministic and ensemble guidance, the 2017 FFaIR Experiment featured several experimental ensemble systems including the experimental Storm-Scale Ensemble Forecast (SSEFX) from the University of Oklahoma (OU) and Center for Analysis and Prediction of Storms (CAPS), the High-Resolution Rapid Refresh Ensemble (HRRRE) and Time-Lagged HRRR (HRRR-TLE) from Earth Systems Research Laboratory (ESRL), and the Experimental High Resolution Ensemble Forecast (HREFv2) provided by Environmental Modeling Center (EMC). The experiment also featured several deterministic high-resolution guidance systems including the High Resolution Rapid Refresh (HRRRv3) provided by ESRL, the 3 km nested North American Mesoscale Model, version 4 (NAM) provided by EMC (now the operational NAM Nest), and the Unified Model (UM) provided by the Met Office. At longer ranges, the experiment featured probabilities derived from the

National Water Model (NWM) provided by the Office of Water Prediction (OWP), the GFS run of the FV3 dynamical core provided by EMC (FV3-GFS), the OU/CAPS run of the FV3 dynamical core (FV3-CAPS), as well as a version of the FV3 provided by the Geophysical Fluid Dynamics Laboratory (GFDL, FV3-GFDL). Table 3 summarizes the model data that was the focus of the experiment. Other scientific tools that were evaluated include a "first guess" field for the Day 2 and Day 3 ERO developed at Colorado State University (CSU) which utilizes reforecast data, ARIs, and machine learning to produce the probability of the QPE exceeding a 1 or 10 year ARI within a 40 km radius. Some real time experimental tools that were provided include the Flooded Locations and Simulated Hydrographs (FLASH) data, GOES-16 total and layered precipitable water, as well as the Cooperative Institute for Research in the Atmosphere (CIRA) layered precipitable water. More detailed information including more detailed descriptions and configurations of the models included in Table 3 can be found in Appendix B.

Table 3. *Featured 2017 FFaIR deterministic and ensemble model guidance (Experimental guidance is in the darker shade)*

Provider	Model	Resolution	Forecast Hours	Notes
EMC	NAM (parent)	12 km (parent)	84 (parent)	Operational 12 km NAM parent model.
EMC	NAM (nest)	3 km	60 hours (nest/00, 06, 12, 18 hours)	Features an hourly forecast and assimilation cycle for its 3 km CONUS nest. Uses hybrid 3DEnVar and incorporates radar reflectivity into its assimilation system via a complex cloud analysis approach.
RFCs	Flash Flood Guidance	5 km	01, 03, 06, 12 and 24 hour values	CONUS mosaic grid created by compiling individual RFC-domain grids
EMC	HRRR	3 km	18	High resolution, hourly updated, convection allowing nest of the Rapid Refresh (RAP) model
EMC/NSSL	NMMB ARW WRF-NSSL	4 km	48 36 (WRF-NSSL)	High resolution, convection allowing CONUS models
OWP	National Water Model (NWM)	250 m 1 km	18 hours 10 days 30 days	Hourly, uncoupled analysis and forecast system that provides streamflow for 2.7 million river reaches and other hydrologic information on 1km and 250m grids.
NSSL/HDSC/NERFC /CSU	Precipitation Recurrence Data (Atlas 14)	5 km	6 and 24 hr (2, 5, 10, 25 and 100 year intervals)	Precipitation frequency estimates based on historical observations.

ESRL/GSD/EMC	HRRRv3	3 km	Hourly out to 36 hours every 3 hours Hourly out to 18 hours every hour 48 hour cycle at 12Z	Experimental version of the HRRR, hourly updating, convection allowing
ESRL/GSD	HRRR Time-lagged Ensemble (HRRR-TLE)	3 km	24	Neighborhood ensembling approach calculated over a 3km grid of time-lagged HRRRv3 deterministic members. Probabilities at a point refer to the chance of exceeding a given threshold somewhere with a 40-km radius around that point.
ESRL/GSD	HRRR Ensemble (HRRRE)	3 km	36 hours at 00Z	9 members,full CONUS domain, stochastic
EMC	HREFv2	3 km	36	Experimental version of HREF with 8 members which produces ensemble mean precip in three different forms, and precipitation probability of exceedance of QPF, FFG, and RIs.
OU/CAPS	WRF-ARW SSEFX	3 km	60	11-member (9 ARW+1 FV3+1 control) ensemble forecast) 00Z cycle
OU/CAPS	FV3-CAPS	3 km	84	Uses the FV3 dynamical core and Thompson microphysics
Met Office	UM	2.5 km	84 for experiment	Convection allowing; sub-CONUS domain, 70 vertical levels up to 40km; UKMet boundary conditions/initialization
EMC	FV3-GFS	13 km	84 for experiment	3D hydrostatic dynamical core; vertically Lagrangian; GFS analyses initialization/physics
GFDL	FV3-GFDL	3 km	84 for experiment	Using the FV3 dynamical core and GFDL microphysics
MDL	NBMv3	2.5 km	Hourly out 36 hrs 3-hrly to Day 8 6-hrly Days 8-10	Runs every hour with 15 different deterministic and ensemble systems

Synoptic Overview and Highlights of Daily Impacts Throughout the Experiment

FFaIR forecasters dealt with a wide variety of weather and flooding challenges over the course of the four week experiment ranging from a landfalling tropical system to monsoon activity in the Southwest United States. Figure 6 shows the 500 hPa mean geopotential height over the United States during the first half of FFaIR (June 19 - June 30) in panel 1 and second half (July 10 - 21) in panel 3. During the first two weeks (panel 1), a 500 hPa ridge dominated the western and southwestern United States with a broad trough in the central and eastern U.S. centered close to the Mississippi River Valley. The western extent of a Bermuda high extended into Florida, creating a gap between the two areas of high pressure. This gap is where, during week 1, Tropical Storm Cindy formed and moved through making landfall near the Louisiana and Texas border, after which it lifted north through the Mississippi and Ohio River Valleys. The 500 hPa trough, which is evident by the anomalously low heights shown in Figure 6 panel 2, helped focus the remnant moisture from Tropical Storm Cindy as it moved farther north. The beginning of week 2 was relatively less active, dominated by scattered, diurnal type thunderstorms in the early part of the week. However, late in week 2, an organized Mesoscale Convective System (MCS) affected northwestern Missouri with 10-11 inches of rain and numerous flash flooding reports. More organized flash flooding also occurred in New England out ahead of a low pressure system that was situated in eastern Canada late in the week. Figure 7 panel 1 displays 1000-500 hPa total column precipitable water anomalies over the first half of FFaIR. The very high anomalies along the Gulf Coast can be linked to the moisture associated with Tropical Storm Cindy. With the exception of the Southeast, most other areas of the country during the first half of FFaIR had near normal or anomalously low 1000-500 hPa total column precipitable water.

In Figure 6, panel 3 and panel 4 show 500 hPa mean geopotential height and 500 hPa geopotential height anomalies, respectively during the second half of the FFaIR Experiment. A broad 500 hPa ridge was firmly in place over Great Plains and Rocky Mountain regions of the US. A trough was located in the eastern Pacific Ocean off the coast of Washington state and a second, weaker, trough was located over the Mid Atlantic and New England regions. The strong ridge in the center of the country led to positive 500 hPa heights anomalies over much of the country during this time and was also a forcing mechanism for much of the weather during the second half of the experiment. Numerous MCS events in week 3 followed the northern periphery of the ridge and traveled from the northern Mississippi River Valley down through the base of the trough across Ohio. This led to repeated rounds of very heavy rainfall over the same areas. Toward the latter half of week 3 the southwest monsoon flow increased on the western portion of the ridge bringing ample moisture and scattered very heavy rain to the Southwest almost daily. Week 4 followed the same pattern with daily monsoon moisture in the Southwest and storm systems tracking along the northern periphery of the ridge, often originating in Minnesota or Wisconsin and travelling southeastward into the Ohio River Valley region. The 1000-500 hPa total column precipitable water anomalies in Figure 7, panel 2 show anomalously high precipitable water over the Southwest United States as well as above average anomalies that stretch from Iowa eastward into Ohio, the Mid Atlantic region and New England.

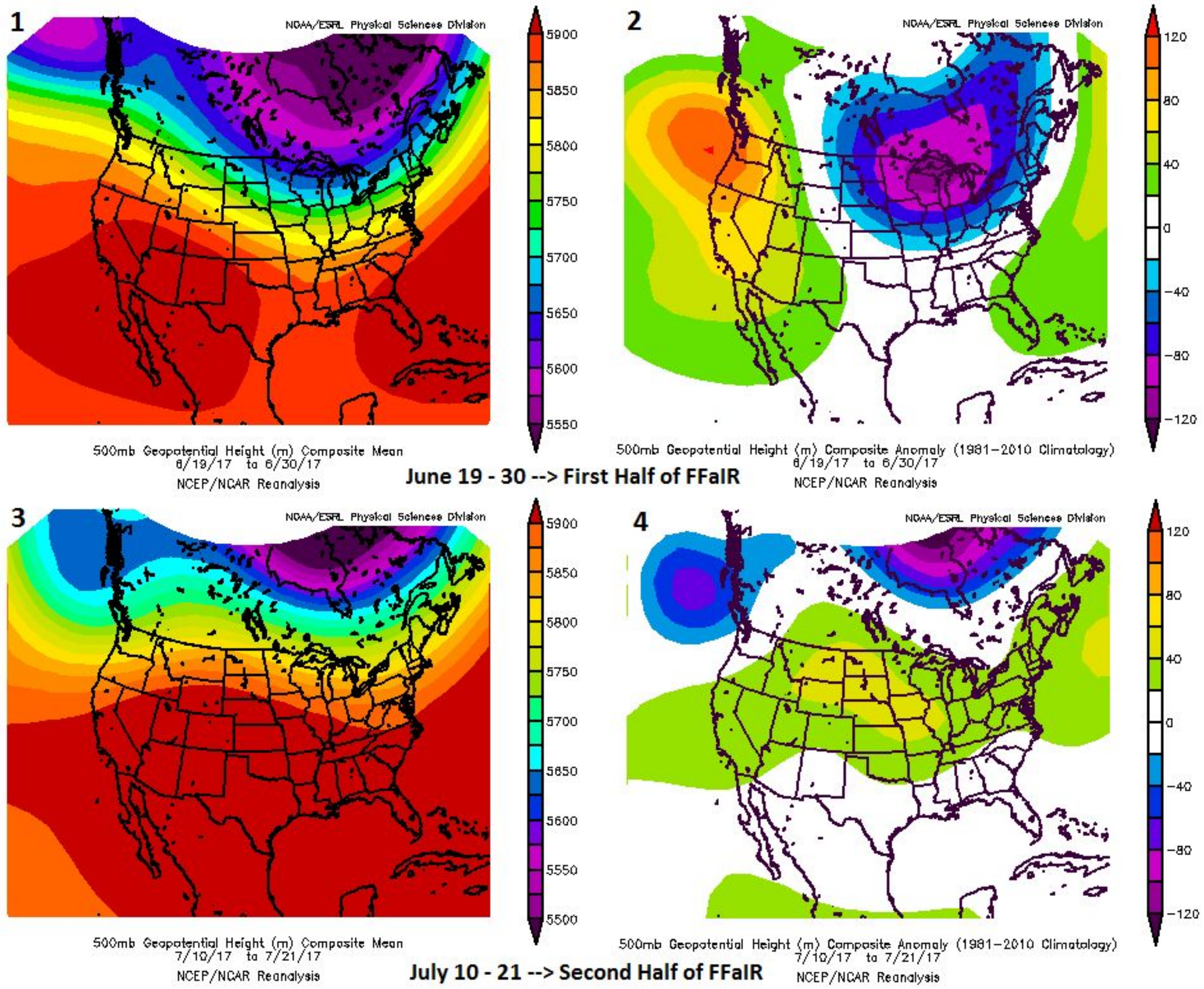

1
NOAA/ESRL Physical Sciences Division
5900
5850
5800
5750
5700
5650
5600
5550
500mb Geopotential Height (m) Composite Mean
6/19/17 to 6/30/17
NCEP/NCAR Reanalysis
2
NOAA/ESRL Physical Sciences Division
120
80
40
0
−40
−80
−120
500mb Geopotential Height (m) Composite Anomaly (1981−2010 Climatology)
6/19/17 to 6/30/17
NCEP/NCAR Reanalysis
June 19 - 30 --> First Half of FFaIR
3
NOAA/ESRL Physical Sciences Division
5900
5850
5800
5750
5700
5650
5600
5550
5500
500mb Geopotential Height (m) Composite Mean
7/10/17 to 7/21/17
NCEP/NCAR Reanalysis
4
NOAA/ESRL Physical Sciences Division
120
80
40
0
−40
−80
−120
500mb Geopotential Height (m) Composite Anomaly (1981−2010 Climatology)
7/10/17 to 7/21/17
NCEP/NCAR Reanalysis
July 10 - 21 --> Second Half of FFaIR

Figure 6. *(1) 500 hPa mean geopotential height and (2) 500 hPa geopotential height composite anomalies for the first half of FFaIR covering June 19 - June 30, 2017. (3) 500 hPa mean geopotential height and (4) 500 hPa geopotential height composite anomalies for the second half of FFaIR covering July 10 - 21, 2017. Images generated from the NCEP/NCAR Reanalysis provided by NOAA/ESRL/Physical Sciences Division (http://www.esrl.noaa.gov/psd/data/composites/day/).*

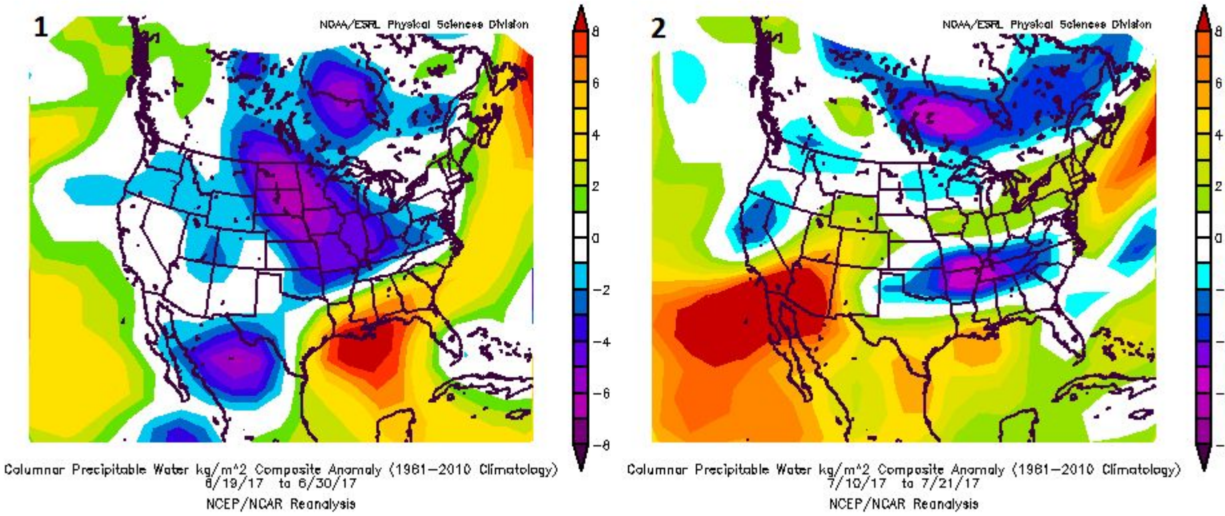
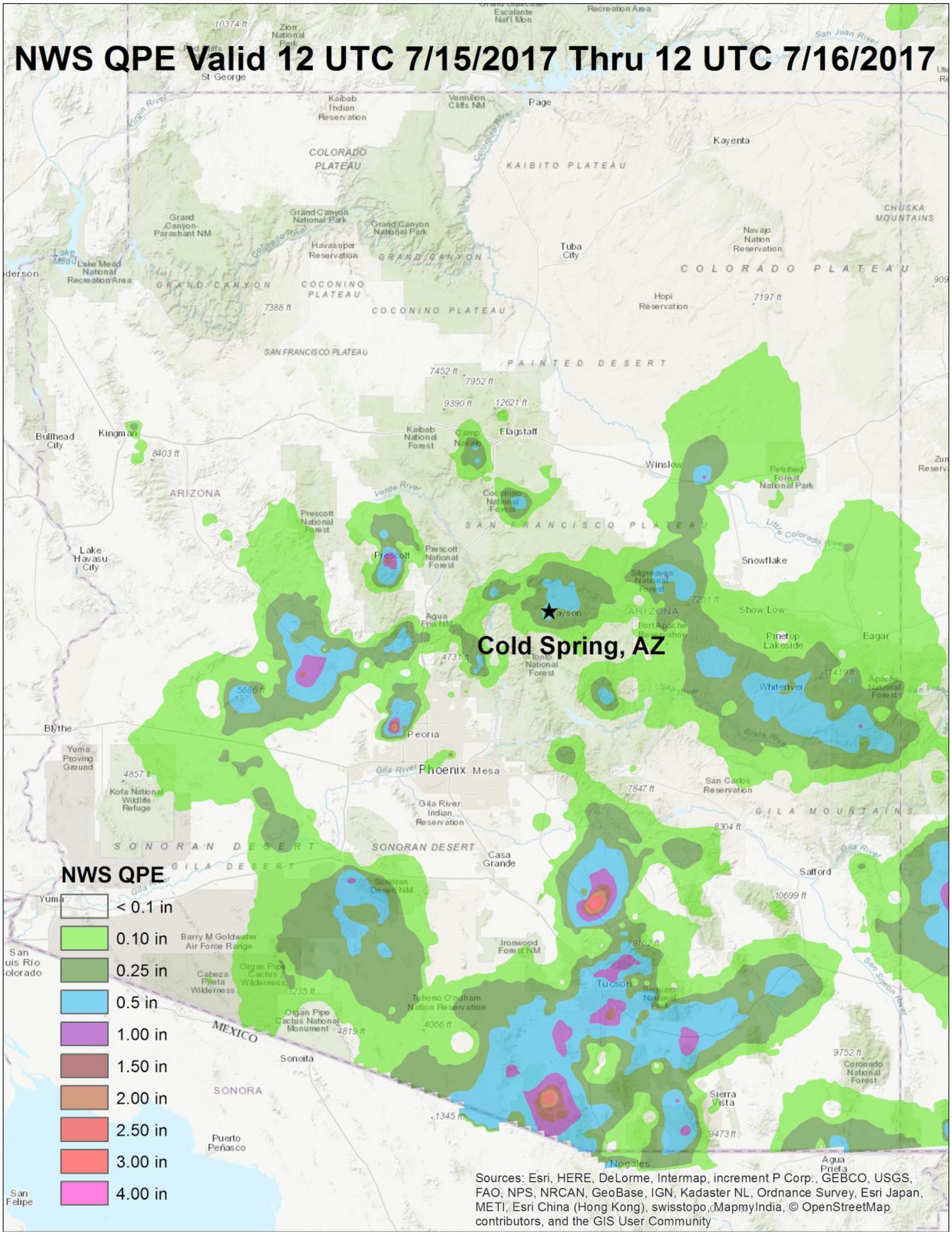

1
NOAA/ESRL Physical Sciences Division
8
6
4
2
0
-2
-4
-6
-8
Columnar Precipitable Water kg/m^2 Composite Anomaly (1961-2010 Climatology)
6/19/17 to 6/30/17
NCEP/NCAR Reanalysis
2
NOAA/ESRL Physical Sciences Division
8
6
4
2
0
-2
-4
-6
-8
Columnar Precipitable Water kg/m^2 Composite Anomaly (1961-2010 Climatology)
7/10/17 to 7/21/17
NCEP/NCAR Reanalysis

FIgure 7. *1000 - 500 hPa precipitable water composite anomalies for (1) the first half of FFaIR (June 19 - June 30, 2017) and (2) the second half of FFaIR (July 10 - 21, 2017). Images generated from the NCEP/NCAR Reanalysis provided by NOAA/ESRL/Physical Sciences Division (http://www.esrl.noaa.gov/psd/data/composites/day/).*

The highest impact event in terms of loss of life during the 2017 FFaIR Experiment occurred during the weekend between weeks 3 and 4 on Saturday, July 15 when 10 people died due to a flash flood that occurred at a swimming hole in Cold Springs, AZ. Although it was during the weekend, the FFaIR Day 3 ERO issued Thursday and the Day 2 ERO issued Friday would have covered the time period. Monsoon moisture led to thunderstorms where 1.5 inches of rain fell within an hour about eight miles upstream of the canyon where the swimming hole was located causing 6 foot high floodwaters to rush through. Figure 8 provides a closer view of the affected area. Table 4 below summarizes the regions highlighted in each forecast over the course of the experiment and any significant events that occurred due to flooding.



Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

Figure 8. *NWS QPE amounts valid from 1200 UTC 7/15/2017 through 1200 UTC 7/16/2017. Cold Spring, AZ represented by black star in the above figure. This is near the area in which there were 10 fatalities due to flash flooding.*

Table 4. *Experimental 6-HR PFF (18-00 UTC) and Day 2 and 3 ERO (12-12 UTC) forecasts issued during the 2017 FFaIR Experiment along with notable impacts.*

Forecast Valid End Date	Valid Time (UTC) (18 - 00 PFF) (12-12 Day 2/3 ERO)	Forecast Area	Notes
6/20/2017	18 - 00	Northeast	Numerous flash flood LSRs in New Hampshire and flash flood warnings throughout the New England area.
6/21/2017	18 - 00	Southeast	Flash flood LSR and swift water rescue reports in and around the city of Atlanta, GA.
	12 - 12	Southeast and Southwest	
6/22/2017	18 - 00	Gulf Coast	Outer rainbands and moisture associated with Tropical Storm Cindy.
	12 - 12	Southeast/Gulf Coast and central US with focus over Iowa	
6/23/2017	18 - 00	Upper Great Lakes region, NW PA	**1 dead** in Indiana County, PA. Railcars washed off tracks by flood waters.
	12 - 12	Upper Great Lakes region, Southeast	Tropical Storm Cindy in the Southeast.
6/24/2017	18 - 00	Ohio River Valley	Numerous reports of flash flooding through Kentucky, southern Ohio, and southwest PA associated with remnants of Cindy and a frontal boundary.
	12 - 12	Central Gulf Coast states, Ohio River Valley, Mid-Atlantic	
6/25/2017	12 - 12	Southeast, New Mexico	
6/26/2017	12 -12	Southeast, New Mexico	
6/27/2017	18 - 00	Florida	Isolated flood LSRs around the Jacksonville, FL area.
6/28/2017	18 - 00	No Forecast Issued	
	12 - 12	Northern central Plains	
6/29/2017	18 - 00	Iowa/Illinois/Wisconsin	

	12 - 12	Central Plains/Upper Great Lakes, western Gulf Coast	
6/30/2017	18 - 00	No Forecast Issued	
	12 - 12	Central Plains, Southeast, Upstate New York	8-11 inches of rain in NW Missouri. Numerous reports of water rescues.
7/1/2017	18 - 00	Indiana/Ohio, New England	Road closures due to high water in both Northeast Ohio as well as Vermont.
	12 - 12	South-central Plains, Southeast, and the Mid Atlantic/New England	
7/2/2017	12 - 12	Eastern New Mexico, Southeast, and New England	Several roads washed out in Vermont. In New Hampshire, high flood waters stranded and displaced campers.
7/3/2017	12 - 12	Central Plains	
7/11/2017	18 - 00	Ohio	
7/12/2017	18 - 00	Ohio River Valley	Numerous roads flooded just south of Indianapolis, IN.
	12 - 12	North Central Plains into Ohio River Valley, Louisiana, Southwest	Numerous reports of flooding in Milwaukee, where manhole covers burst due to amount of water and in North Chicago.
7/13/2017	18 - 00	Western Great Lakes, New England	Reports of street flooding and cars stuck in high water in and around the Boston, MA area.
	12 - 12	Great Lakes/Ohio, Southwest	
7/14/2017	18 - 00	Eastern Ohio, Mid Atlantic, southern New England	Street flooding in southern Rhode Island. Numerous reports of flooding in Ohio, including a high water rescue in Hebron, OH.
	12 - 12	Southwest, Central Plains, Ohio River Valley, Mid Atlantic, New England	
7/15/2017	18 - 00	Arkansas, Kentucky, West Virginia, Delaware, South	Homes evacuated, cars stranded in high water in Vestal, NY. Two cars and bus stranded in high water near

		Central New York	Newark, DE. No injuries.
	12 - 12	Southwest into the northern Southeast, Mid Atlantic, and central New England	
7/16/2017	12 - 12	Southeast, Southwest/Colorado	**10 killed** in flash flood at swimming hole in Cold Springs, AZ.
7/17/2017	12 - 12	Southeast, Southwest	
7/18/2017	18 - 00	Southern South Carolina/Georgia and northern Florida	
7/19/2017	18 - 00	Minnesota, Wisconsin, Upper Michigan	
	12 - 12	Northern central Plains, Southwest	
7/20/2017	18 - 00	Southwest	Minor localized flooding around Las Vegas, NV.
	12 - 12	Southwest, northern Mississippi River Valley	Widespread flooding in southeastern MN and NW Wisconsin. Mudslides covered highways and 100s of people evacuated from homes in Arcadia, WI.
7/21/2017	18 - 00	Southwest	About 50 homes affected by flooding in Salem, UT.
	12 - 12	Southwest, Rocky Mountains, north central Plains	
7/22/2017	18 - 00	Minnesota, Wisconsin, Iowa, Illinois	
	12 - 12	Southwest, northern Mississippi River Valley, Ohio River Valley	Numerous areas of flooding reported in NE Iowa. Iowa Highway 93 closed in several places due to high water.
7/23/2017	12 - 12	Southwest, Ohio River Valley	Two homes reportedly swept away in Augusta, KY. Street flooding in Ohio.
7/24/2017	12 - 12	Southwest, Ohio River Valley, Mid Atlantic, southern New England, western Gulf Coast	17 hikers stranded by flash flooding in Arizona. Roads flooded and cars stranded in Middletown, PA where over 4 inches of rain fell in one hour.

4. Atmospheric Guidance Results

Days 2 and 3 Deterministic Guidance Performance

Several deterministic models were featured during the 2017 FFaIR experiment both for forecast guidance and evaluation. A primary science goal was to determine the utility of longer-range, high-resolution convective-allowing models for improving the prediction of heavy rains that may lead to flash flooding.

The participants were presented with a display of Day 2 24-hour QPFs (Figure 9) and Day 3 24-hour QPFs (Figure 10) along with the MRMS QPE for that same time period (see Verification section for details). Subjective scores and comments were collected for available guidance each day during the experiment.

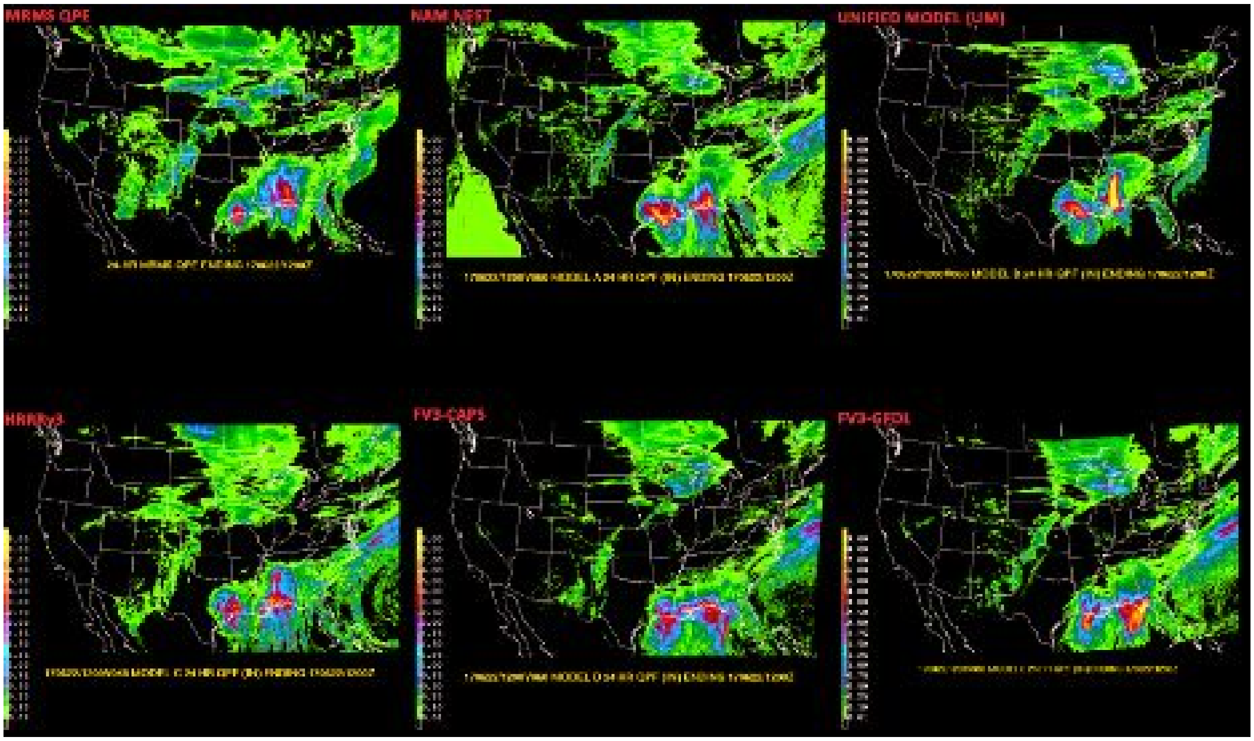


Figure 9. *24-hour QPF forecasts subjectively evaluated for Day 2 using 24-hour NSSL QPE. Day 2 evaluation included the NAM Nest, UM, HRRRv3, FV3-CAPS and FV3-GFDL.*

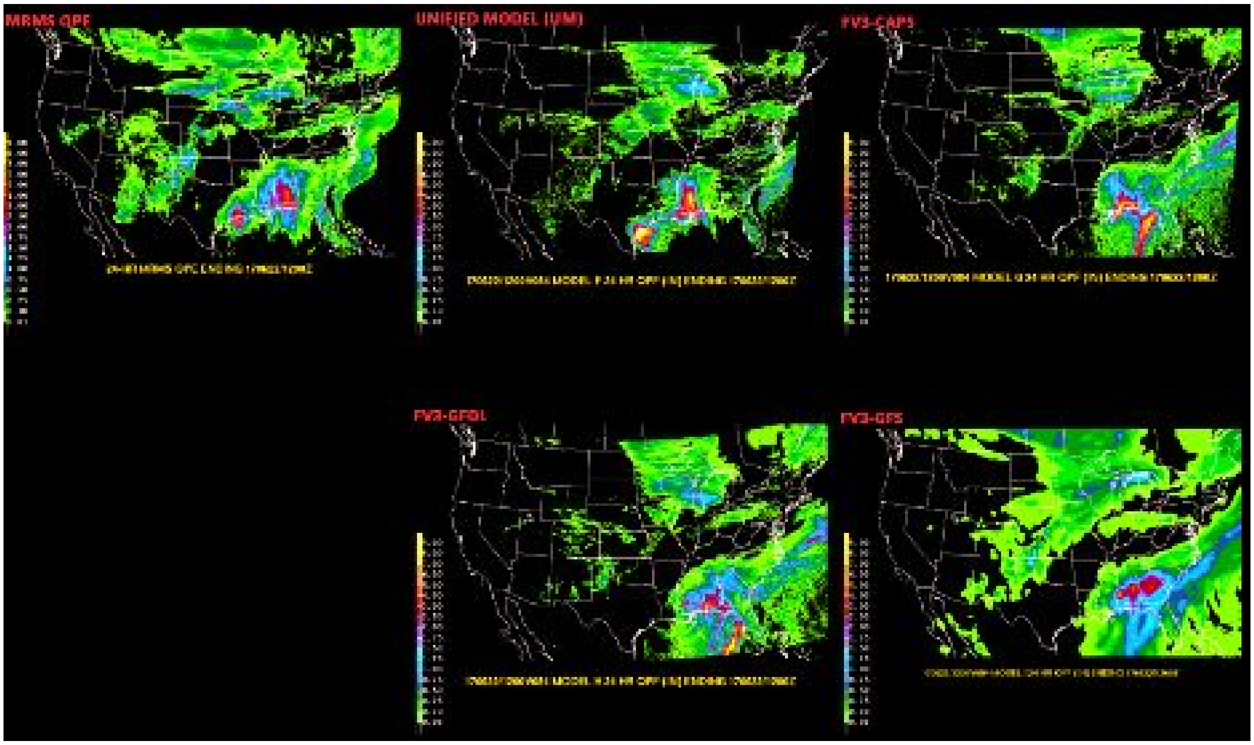


Figure 10. *24-hour QPF forecasts subjectively evaluated for Day 3 using 24-hour NSSL QPE. Day 3 evaluation included the UM, FV3-CAPS, FV3-GFDL, FV3-GFS.*

<u>Day 2 Results</u>

Figure 11 is the box plot of the subjective evaluation collected for Day 2 deterministic 24-hour QPF guidance. The NAM Nest and the HRRRv3 received the highest average scores from participants, but as seen in the diagram, the NAM Nest scores had much greater spread than the HRRRv3. The FV3-CAPS scored the lowest.

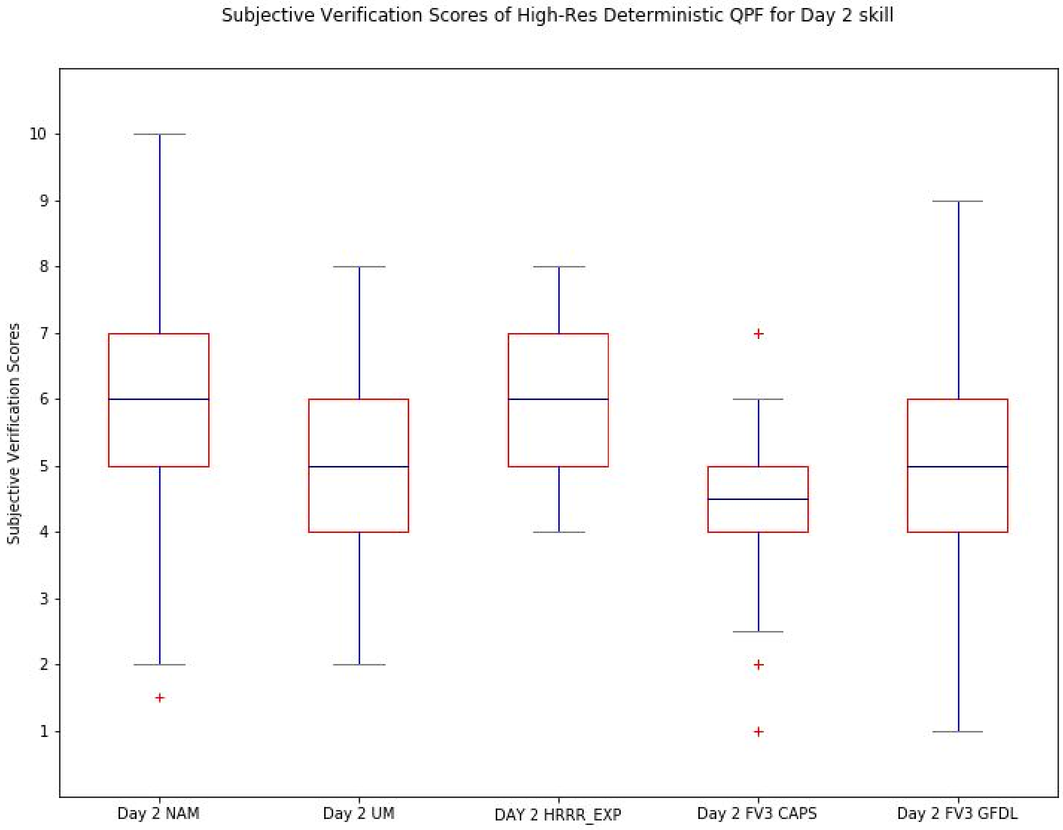


Figure 11. *Box plot of the subjective scores for the Day 2 24-hour QPF from the NAM Nest, UM, HRRRv3, FV3-CAPS, and FV3-GFDL.*

The **NAM Nest** received an average score of 6.03 out of 10 over a total of 152 collected scores with a standard deviation of 1.40. Overall, participants liked the NAM Nest Day 2 guidance expressing that despite typical CAM predictability issues with details, especially for more marginal events and convective outbreaks, the forecast spatial extent and magnitude captured the general shape and provided a reasonable representation of the rainfall. Noteworthy is how compared to previous years, "over done" and "magnitude too high" were rarely noted. In fact, a large number of experiment days noted areas of the CONUS where the NAM Nest had too low magnitude as compared to verification.

The Day 2 **Unified Model (UM)** emerged as the preferred guidance for capturing the spatial pattern of the precipitation and useful magnitudes early in the experiment, most likely due to the predominance of more synoptically-forced events such as Tropical Storm Cindy. But it quickly became less favorable as it suffered from an overall low bias, noisy QPF fields that were scattered in nature, phase errors, and magnitudes that were too high where rain actually occurred (especially along the Gulf Coast). The spatial and magnitude inconsistencies decreased forecaster confidence in the UM guidance. The UM achieved an average score of 5.17 out of 10 with a standard deviation of 1.50 for its Day 2 QPF guidance.

With an average score of 6.03 out of 10 and a standard deviation of 1.04, the Day 2 **HRRRv3** was one of the higher-scoring deterministic models, alongside the NAM Nest. However, its average was based only on a total of 89 scores (available 13 forecast days out of 20 possible) whereas the NAM Nest had a total of 159 scores (see Table 2 in the Verification section for more details). Additionally, the HRRRv3 was evaluated at 48 hours as opposed to the full 60 hours available in the rest of the Day 2 guidance, which may have impacted its scores. The HRRRv3 was often praised for its capture of the general precipitation pattern over the CONUS. Only on a few days did the HRRRv3 struggle to advance the synoptic pattern effectively. Conversely, participants often noted a low bias of precipitation magnitude for most events. As with most of the CAMs, the HRRRv3 also consistently produced too much precipitation amid the Gulf Coast states during the one tropical event during the 2017 FFaIR.

The **FV3-GFDL** for Day 2 acquired daily comments noting a low bias of precipitation over the CONUS and a misleading high bias for strongly-forced synoptic events such as tropical systems or well-defined frontal boundaries. The FV3-GFDL struggled to organize precipitation for both synoptic (mesoscale) and more marginal convective events resulting in general scatter, and displacement errors in precipitation maxima were often noted. The participants struggled to find utility in the guidance without consulting other CAM precipitation to identify any hints at a pattern. The FV3-GFDL achieved an average score of 4.85, with a standard deviation of 1.69.

The Day 2 **FV3-CAPS** achieved an average score of 4.41 with a standard deviation of 1.24. The FV3-CAPS produced daily concerns about not generating enough precipitation nor having any identifying mesoscale nor convective patterns. On days when the guidance did capture the general precipitation pattern, the result was misleading in areal coverage and amounts and therefore difficult to trust. The lack of spatial extent, presence of precipitation areas that were too narrow, extreme low bias, and random QPF where no present discernible forcing present were noted by participants in their daily comments. The FV3-CAPS was rarely a model that provided utility in forecast activities due to these errors which decreased forecaster confidence in the guidance.

Day 3 Results

Figure 12 is the box plot of the subjective evaluation scores collected for Day 3 deterministic 24-hour QPF guidance. The Unified Model (UM) received the highest overall subjective score for Day 3 and the FV3-CAPS the lowest. The FV3-GFDL had the greatest spread in scoring Day 3 QPF over the experiment.

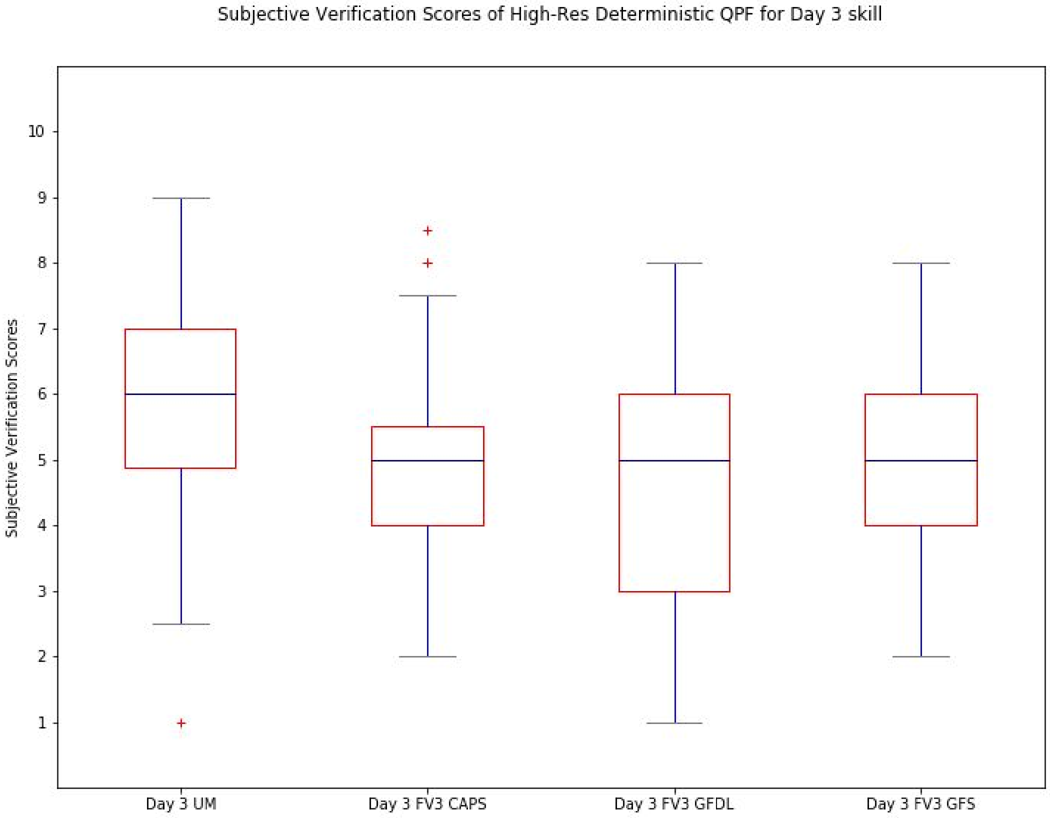


Figure 12. *Box plot of the subjective scores for the Day 3 24-hour QPF from the UM, FV3-CAPS, FV3-GFDL and FV3-GFS.*

With an average score of 5.64 out of 10 and standard deviation of 1.52, the **UM** performed slightly better on Day 3 than on Day 2. Participants felt the precipitation from the UM on Day 3 was often a bit more organized and captured the general pattern. Spatial extent was more useful than on Day 2, although magnitude still struggled with a significant low bias over the entire domain and localized convection associated with strong forcing mechanisms was often missed.

For Day 3, the **FV3-GFDL** achieved an average score of 4.65 out of 10 with a standard deviation of 1.77. Similar to its performance on Day 2, the FV3-GFDL continued to struggle with producing enough precipitation to glean meaningful patterns to contribute to the forecast process on Day 3. The low bias and disorganized, scattered QPF were noted as often occurring over the whole CONUS. For large synoptic-scale events and well-initialized convection, the FV3-GFDL produced a recognizable pattern that was useful to the forecast process. But the run to run inconsistencies in the QPF pattern and displacement errors did not allow the forecasters to have confidence in the guidance day-to-day (Figure 13).

The **FV3-CAPS** precipitation fields attempted to capture the general pattern of QPF and responded best with strong synoptic forcing, as was the case with a string of low pressure systems riding along a stationary boundary on July 22, 2017. But it often struggled to generate organized precipitation over the CONUS when the pattern weakened, as shown just a day later in Figure 13. A significant low bias in both areal extent and magnitude with disorganized and scattered, cells were often the resulting guidance. Many events were completely missed and for those which the FV3-CAPS derived structure were displaced, too localized and overdone in magnitude. Participants struggled to have confidence in this guidance on Day 3, scoring it an average of 4.6 out of 10 with a standard deviation of 1.44.

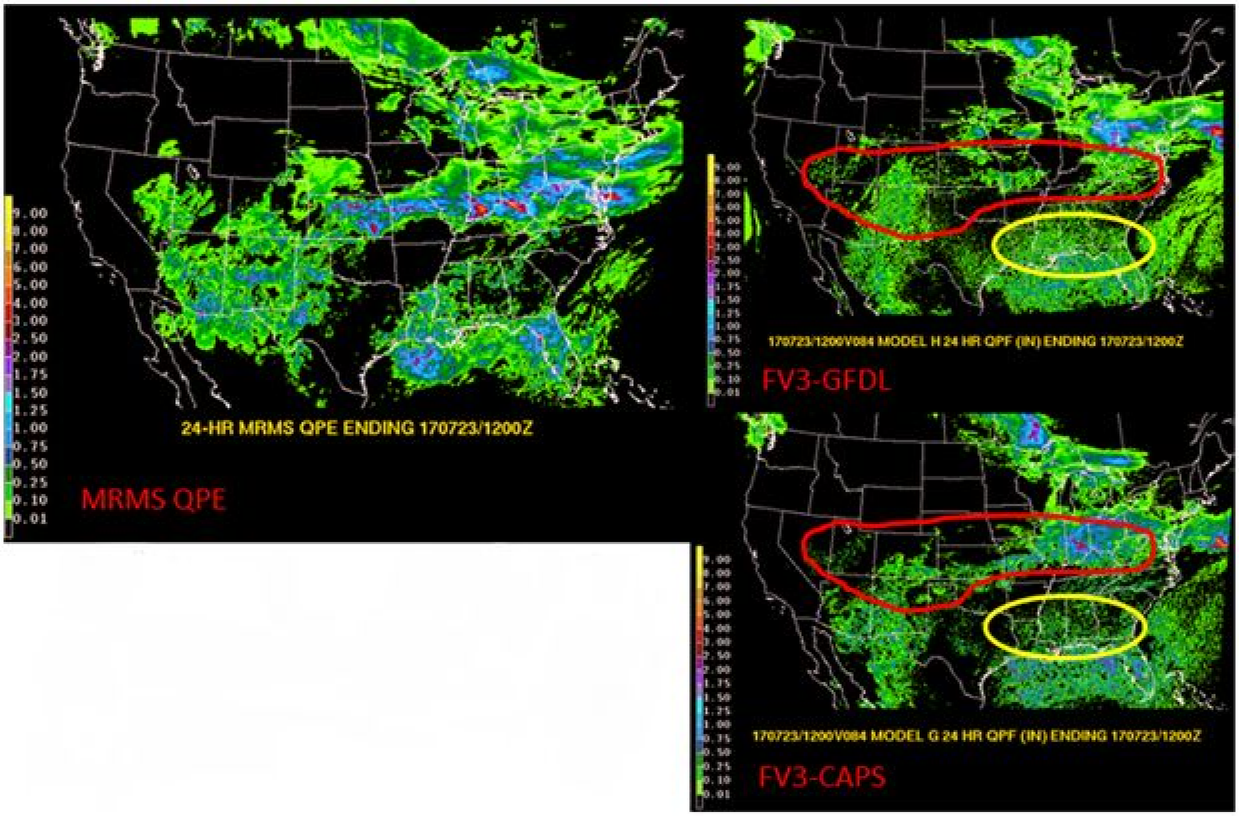

24-HR MRMS QPE ENDING 170723/1200Z
MRMS QPE
170723/1200V084 MODEL H 24 HR QPF (IN) ENDING 170723/1200Z
FV3-GFDL
170723/1200V084 MODEL G 24 HR QPF (IN) ENDING 170723/1200Z
FV3-CAPS
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01

Figure 13. *An example of the MRMS QPE verification along with the 24-hour QPF showing a low bias (red polygons) apparent in both the FV3-GFDL and FV3-CAPS as well as disorganized scatter (yellow polygons) along the Gulf Coast. Day 3 forecast valid 12Z July 23, 2017.*

Achieving an average score of 5.13 out of 10 with a standard deviation of 1.33, the Day 3 **FV3-GFS** at its 13-km resolution received many positive comments during the FFaIR experiment, mostly for accuracy of spatial extent and capturing the location of where heavier precipitation was expected to occur. Due to its lower resolution, participants did remark often on the low bias of the maximum precipitation, the tendency to "broad-brush" areal coverage, and the struggle to capture convection. Displacement errors of heavier precipitation were noted several times during the experiment, and were most often to the north.

Conclusions

Objective verification was performed on the 0.5” and 1” QPFs for Day 2 and Day 3 on all of the deterministic models available in the 2017 FFaIR Experiment using MODE (see the Verification section and Appendix C for MODE settings). The critical success index (CSI), which combines the probability of detection (POD) and the false alarm ratio (FAR), and frequency bias were calculated and graphed onto performance diagrams (Figure 14-17).

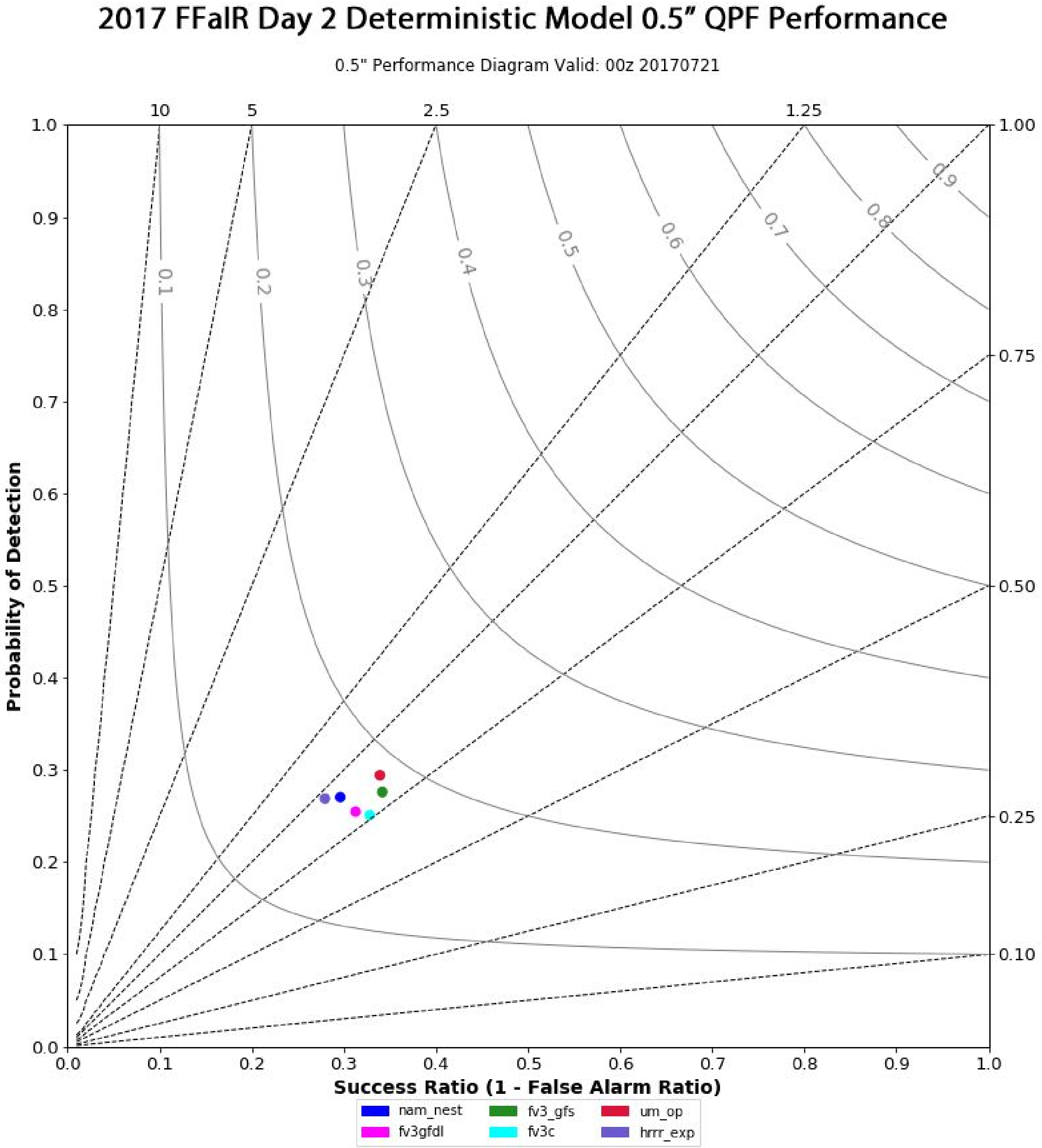


***Figure 14.** Day 2 0.5” Deterministic QPF Performance Diagram including the NAM Nest, Unified Model (UM), FV3-GFDL, FV3-CAPS, and FV3-GFS.*

As shown in the objective analysis of the Day 2 0.5" threshold, the HRRRv3 had the best bias score of 0.97 and the UM had the best CSI at 0.19. All of the models fell between 0.17 and 0.2 for the CSI mean and 0.75 and 0.97 for the bias means, indicating under-forecasting of the QPF at this threshold for Day 2.

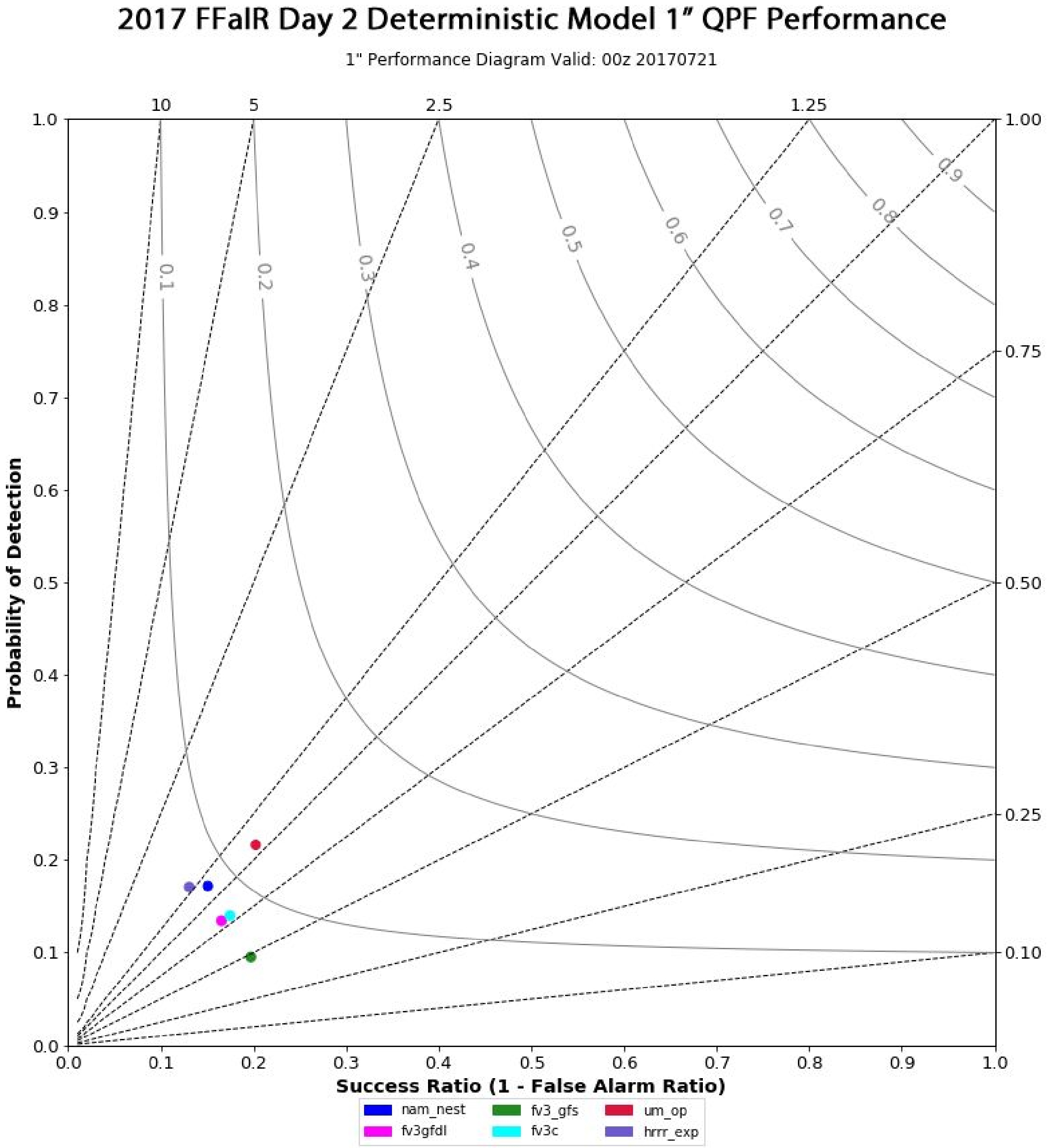


Figure 15. *Day 2 1" Deterministic QPF Performance Diagram including the NAM Nest, Unified Model (UM), FV3-GFDL, FV3-CAPS, and FV3-GFS.*

At the 1” QPF threshold for Day 2, the models tended to fall off in skill with the UM as the exception. The UM had a CSI of 0.12 and a bias of 1.1, making it the best performer. The FV3-GFS scored the lowest in the performance evaluation with a CSI of 0.06 and bias of 0.45.

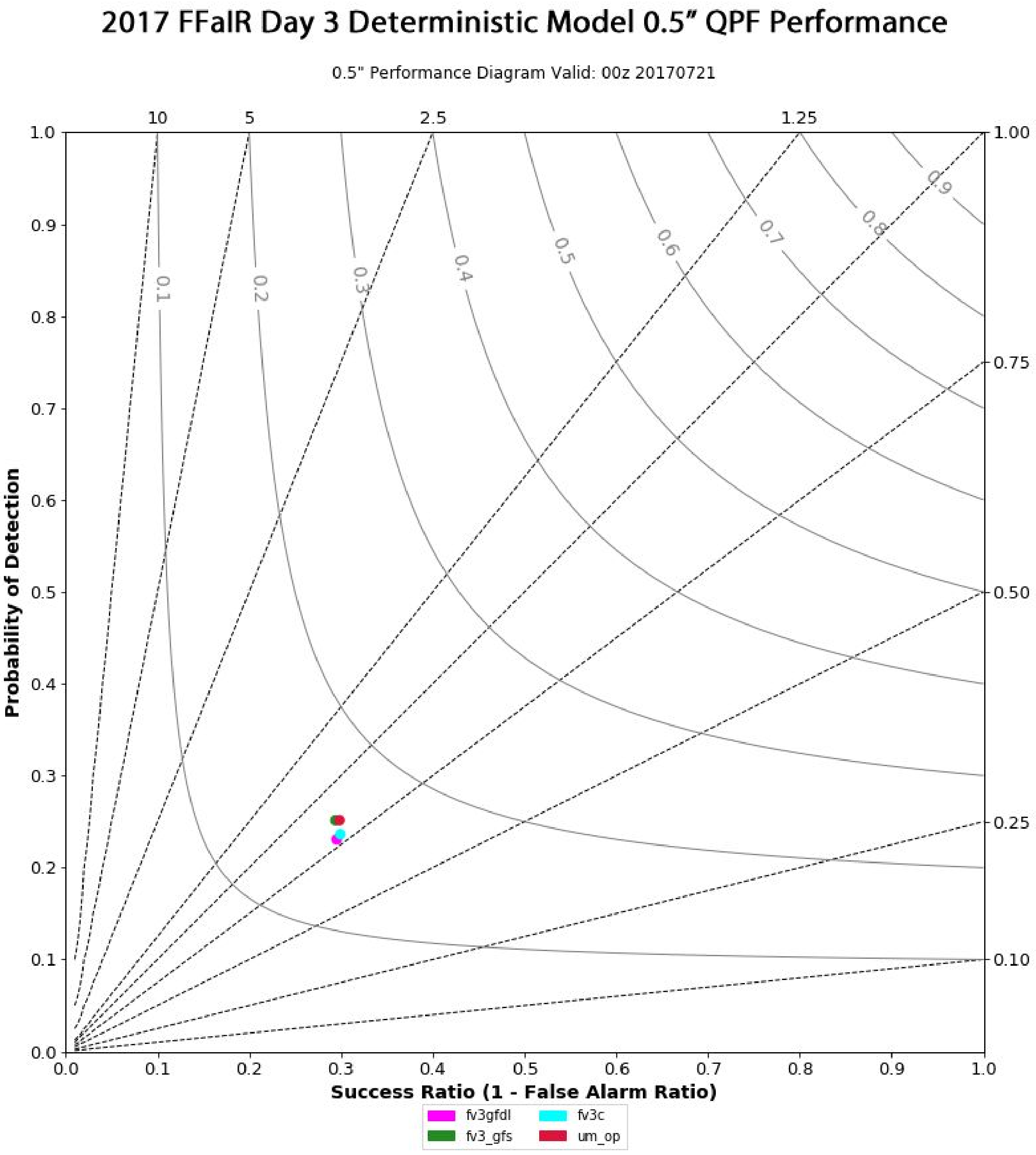


Figure 16. *Day 3 0.5” Deterministic QPF Performance Diagram including the Model (UM), FV3-GFDL, FV3-CAPS, and FV3-GFS.*

As shown in the objective analysis of the Day 3 0.5" threshold, the UM again outperformed the other deterministic models with a CSI of 0.16 and bias of 0.87. All of the models fell between the CSI mean of 0.14 and 0.16, and bias mean of 0.76 and 0.87 which is indicative of slightly more under-forecasting than Day 2. This low bias is also reflected often in the subjective comments. Both the FV3-CAPS and FV3-GFDL tended to perform similar to each other.

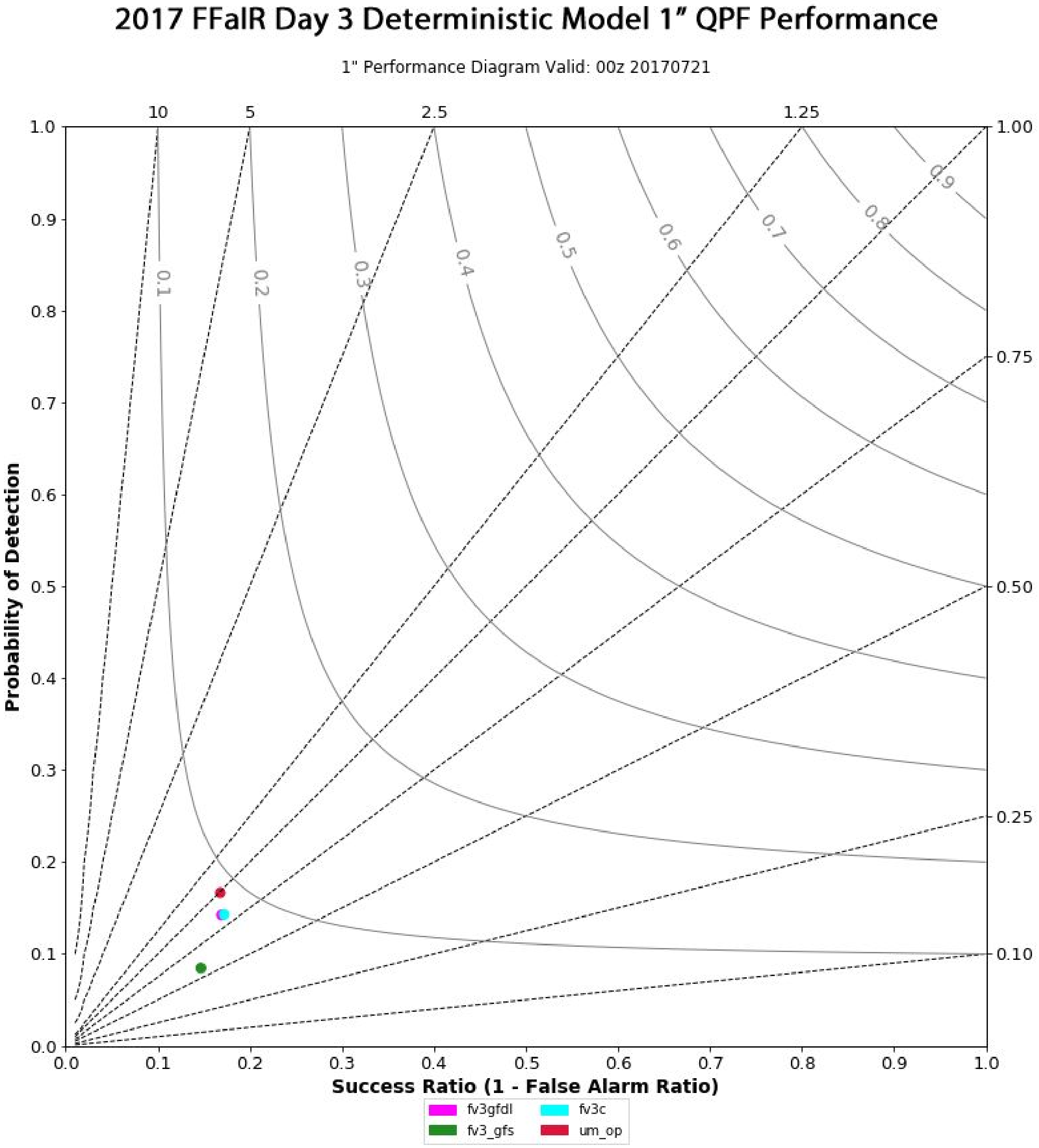


Figure 17. *Day 3 1" Deterministic QPF Performance Diagram including the Model (UM), FV3-GFDL, FV3-CAPS, and FV3-GFS.*

At the 1” QPF threshold for Day 3, the models again fell off greatly in skill from the 0.5” threshold. Most extreme was the FV3-GFS with a CSI of 0.05 and a bias of 0.55. The UM again objectively performed best at this threshold with a CSI of 0.09 and bias of 1.01. At this threshold, the FV3-CAPS and FV3-GFDL again performed similar to each other.

Taking all of the MODE performance averages over all of the available models, it can be determined that overall, the models are more skillful on Day 2 than Day 3. The average of all models on Day 2 for the 0.5” QPF threshold is 0.18 for the CSI mean and 0.83 for bias, and at the 1” threshold 0.09 for the CSI mean and 0.93. The average of all models on Day 3 for the 0.5” QPF threshold is 0.16 for the CSI mean and 0.80 for bias, and at the 1” threshold 0.08 for the CSI mean and 0..81 for bias.

A comparison of the subjective and objective analysis for the Day 2 and 3 forecasts reveals some disagreement as to which model was the best. Objective verification showed the UM achieved the highest scores for the Day 2 forecast at both 0.5 and 1.0 inch thresholds, however, it finished 3rd out of 5 in the subjective evaluation of the guidance. The differences in scores among the models are likely due to the common basis of subjective ratings on areas of highest threat for flooding rains, while the objective statistics reflect the entire CONUS.

<u>Ensemble Guidance Performance and The National Blend of Models</u>

Testing the Blended Mean Approach from CAM Ensembles

During the 2016 FFaIR Experiment, both the probability matched mean (PMM) and conventional ensemble mean (CM) for QPF were compared between two CAM ensembles. General feedback overall showed that participants favored the PMM over the CM, however there were several cases where the PMM proved to be too heavy and thus misleading. In response to the mixed feedback, the 2017 FFaIR Experiment introduced a blended mean (BM) in which the PMM and the CM are averaged together, each with equal weighting (50%).

The BM for QPF was evaluated from three different CAM ensembles: the HREFv2, the HRRRE, and the SSEFX. Each was an 18-24 hour forecast from the 00Z cycle of six hour QPF valid from 18-00 UTC over a limited domain, typically the same domain participants had chosen for the PFF forecast. Participants provided a subjective ranking from 1 (very poor) to 10 (very good) as well as commented on, in particular, how well the BM magnitude matched the QPE verification. It should be noted that for the SSEFX, instead of using the traditional PMM in the BM calculation, a localized probability matched mean was used (LPM). More information on the LPM can be found within the detailed ensemble description of the SSEFX in Appendix B. Figure 18 shows an example of how the evaluation was presented to the participants.

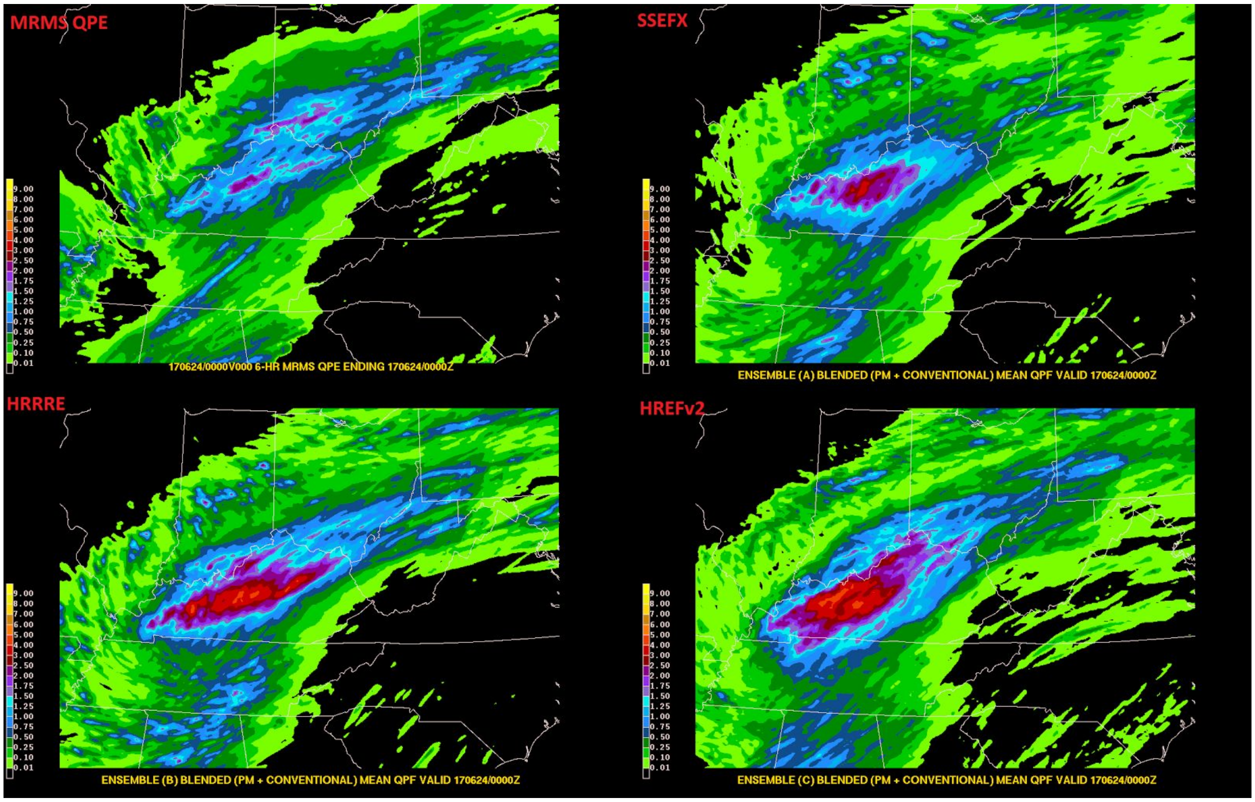


Figure 18. *Four panel display showing MRMS QPE (top left), BM from SSEFX (top right), BM from HRRRE (bottom left) and BM from the HREFv2 (bottom right) valid 18 June 23 to 00 UTC June 24, 2017.*

Figure 19 displays box plots showing statistics of the subjective verification scores for the blended mean evaluation from each ensemble. Each of the three models scored similarly with the SSEFX having the highest average subjective score of 6.73, with a median of 7 and standard deviation of 1.32. The HRRRE's average subjective score was 6.57 with a median of 7 and a standard deviation of 1.49, and the HREFv2's average subjective score was 6.02 with with a median of 6 and a standard deviation of 1.53. It should be noted that the SSEFX was the only model of the three to use a local probability matched mean rather than the traditional PMM in the blended mean. More testing will be required to determine whether the LPM method improves on the PMM and results in higher scores. Also, due to data availability issues during the experiment, the HREFv2 had 39 less overall scores recorded than the SSEFX and 13 less overall scores recorded than the HRRRE. The HRRRE had 16 less overall scores recorded than the SSEFX. All three of these subjective scores from this year's three ensembles are higher than the overall subjective ratings for the individual PMM and CM from the two ensembles tested during the 2016 FFaIR Experiment.

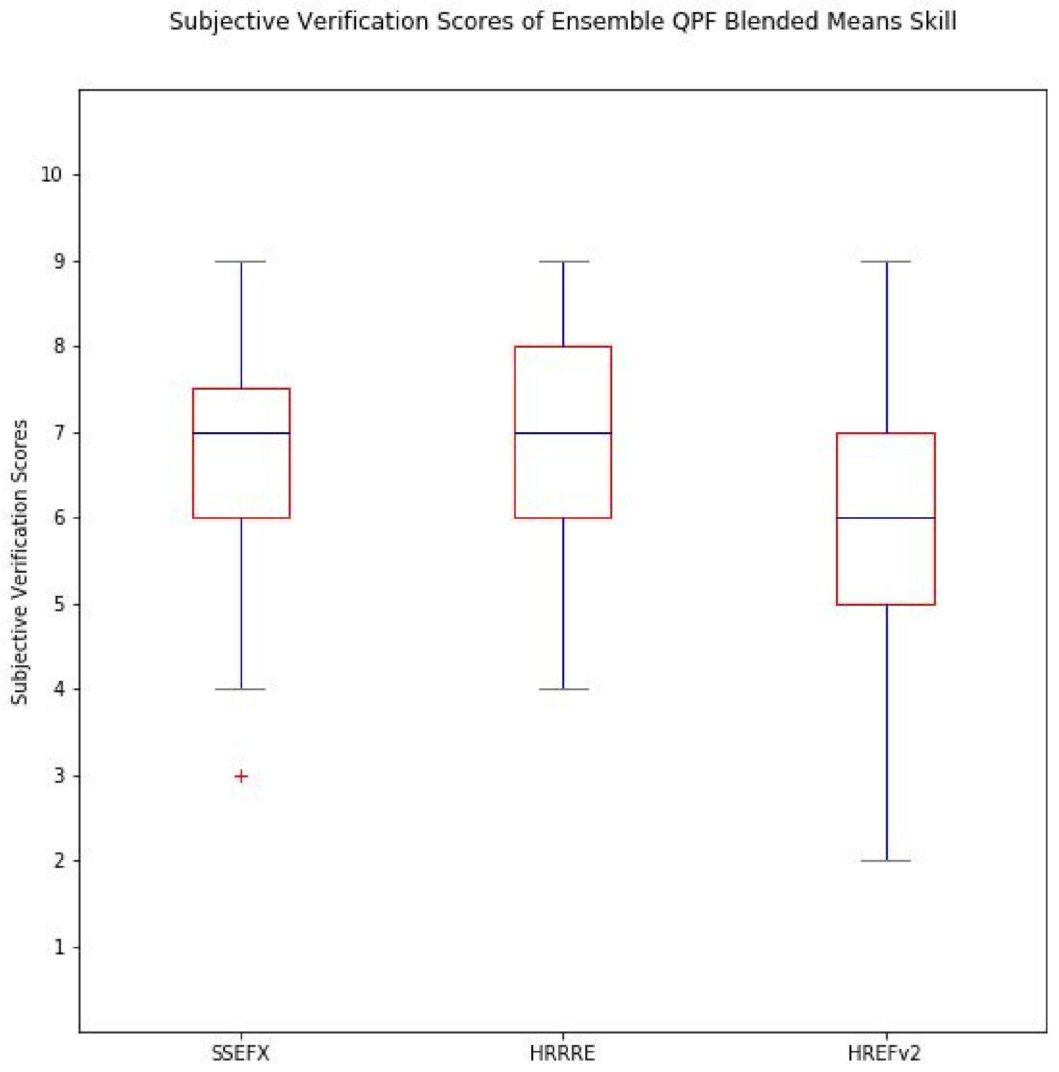


Figure 19. *Box plot of the subjective scores for the SSEFX, HRRRE, and HREFv2 6 hour blended mean QPF over the course of the experiment. Red plus symbols denote outliers.*

Figure 20 shows a Roebber Performance Diagram for all of the six hour ensemble blended means cumulatively over the entire experiment at 0.5 inch threshold. WPC QPF for the same Day 1 18-00 UTC period is also displayed. The HRRRE and HREFv2 had very similar CSI values of 0.15 and 0.14, respectively, however the HRRRE had a better frequency bias of 1.04 when compared to the HREFv2, which was 0.81. The SSEFX had a better bias (0.96) than the HREFv2 but a worse CSI value (0.12). At the one inch threshold, shown in Figure 21, all three models had a CSI value below the 0.10 threshold. The frequency bias at the one inch threshold for the HRRRE was 1.56 and the SSEFX was 1.04. The HREFv2, however, had a frequency bias below one at 0.74 for the one inch threshold.

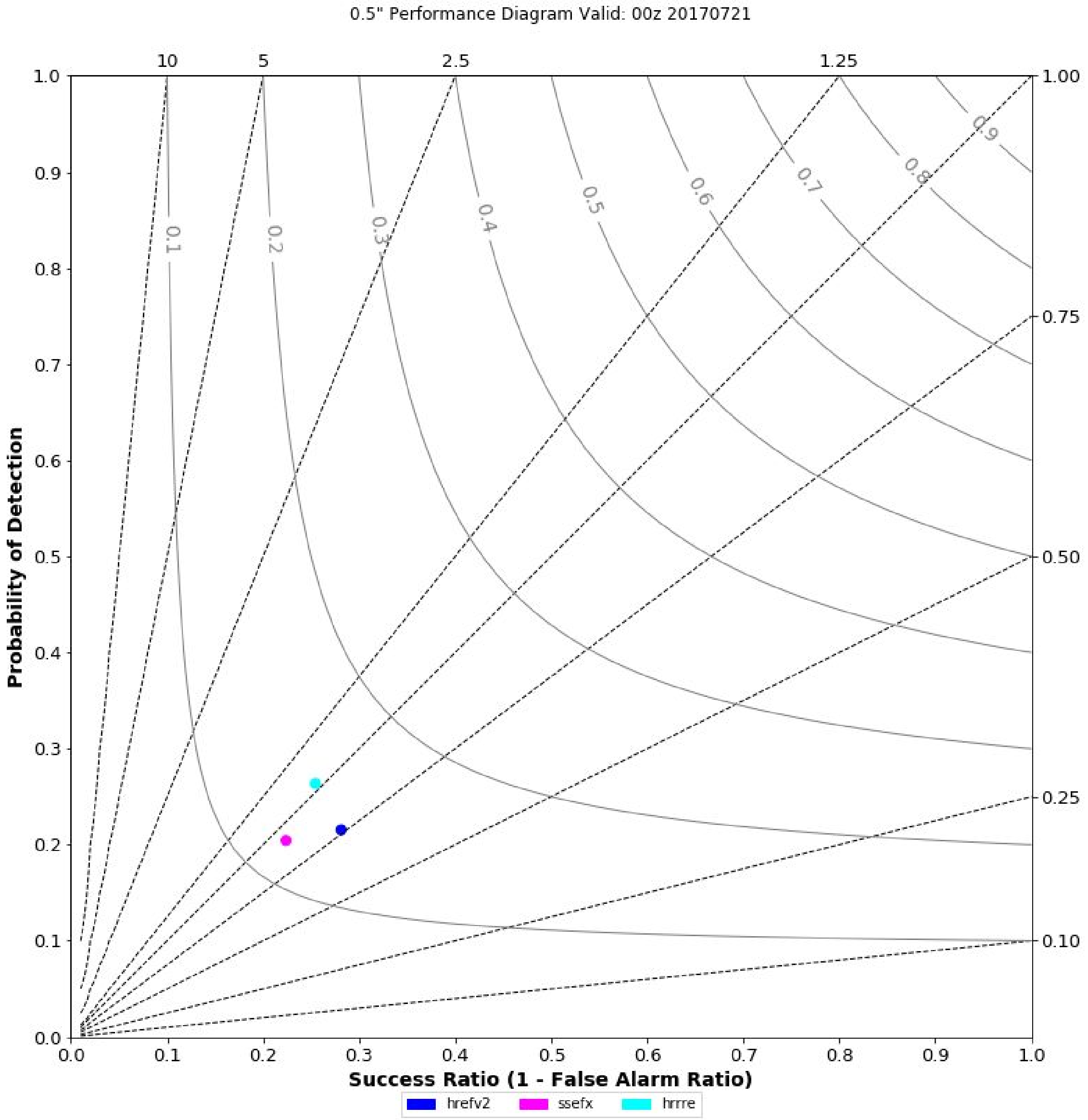


Figure 20. *Roebber Performance Diagram showing skill of the 6 hour blended mean forecast for 0.5 inch QPF from the HREFv2 (dark blue), SSEFX (magenta), and HRRRE (cyan).*

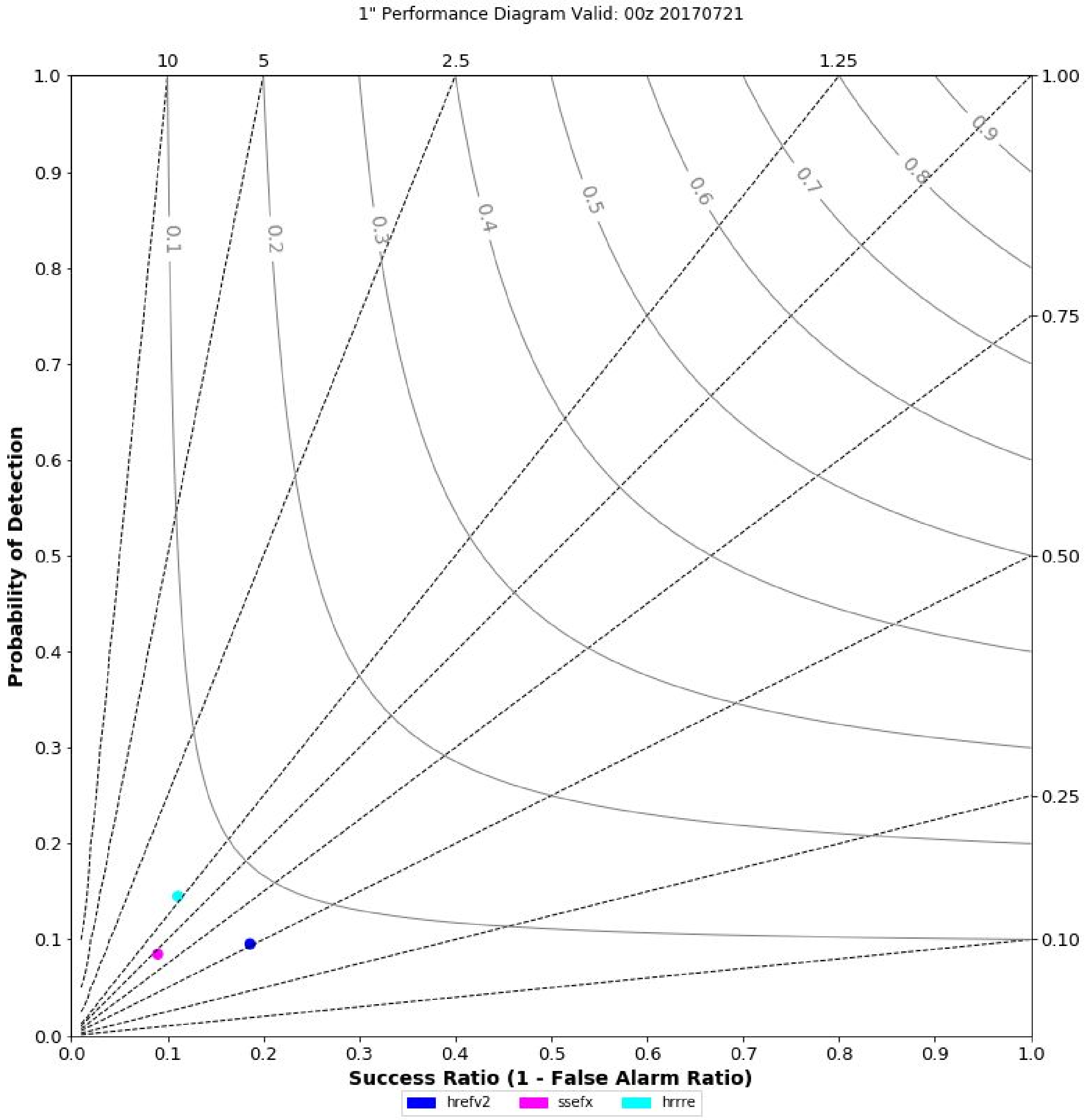


Figure 21. *Roebber Performance Diagram showing skill of the 6 hour blended mean forecast for 1 inch QPF from the HREFv2 (dark blue), SSEFX (magenta), and HRRRE (cyan).*

As the subjective scores reflect, feedback for the ensemble BMs was generally positive. Only five of the days during the experiment did the average subjective score for the majority (2 out of 3) models equal a 5 or below. On those days, the most common problem was usually that the QPF was significantly displaced from the observed QPE for the forecast to have any positive value. Aside from those days, participants were impressed with location and spatial coverage of the QPF for the majority of the forecasts from all three ensembles. Magnitudes varied from being overdone or underdone on most days but usually not enough for the forecast to lose its value. Future work in testing the blended mean include more direct comparisons of the components in an experimental setting, experimenting with weighting the PMM and CM differently, and examining the impact of the LPM more closely.

Ensemble QPF Probability Results and Comments

In addition to the blended means, participants also evaluated the six hour, 40 km neighborhood probability of one-half inch and one inch of QPF occurring within the 18-00 UTC time period over a limited domain from the HREFv2, SSEFX, HRRR-TLE, and HRRRE. All probability forecasts were 18-24 hour forecasts initialized from a 00 UTC cycle. The probabilities were compared with six hour MRMS QPE and participants provided a subjective ranking from 1 (very poor) to 10 (very good) as well as comments focused on the representativeness of the probabilities (whether they felt the values were too high or low based on the magnitude of the event) as well as the overall coverage. Figure 22 shows an example of how the six hour probability of one inch of QPF evaluation was shown to participants.

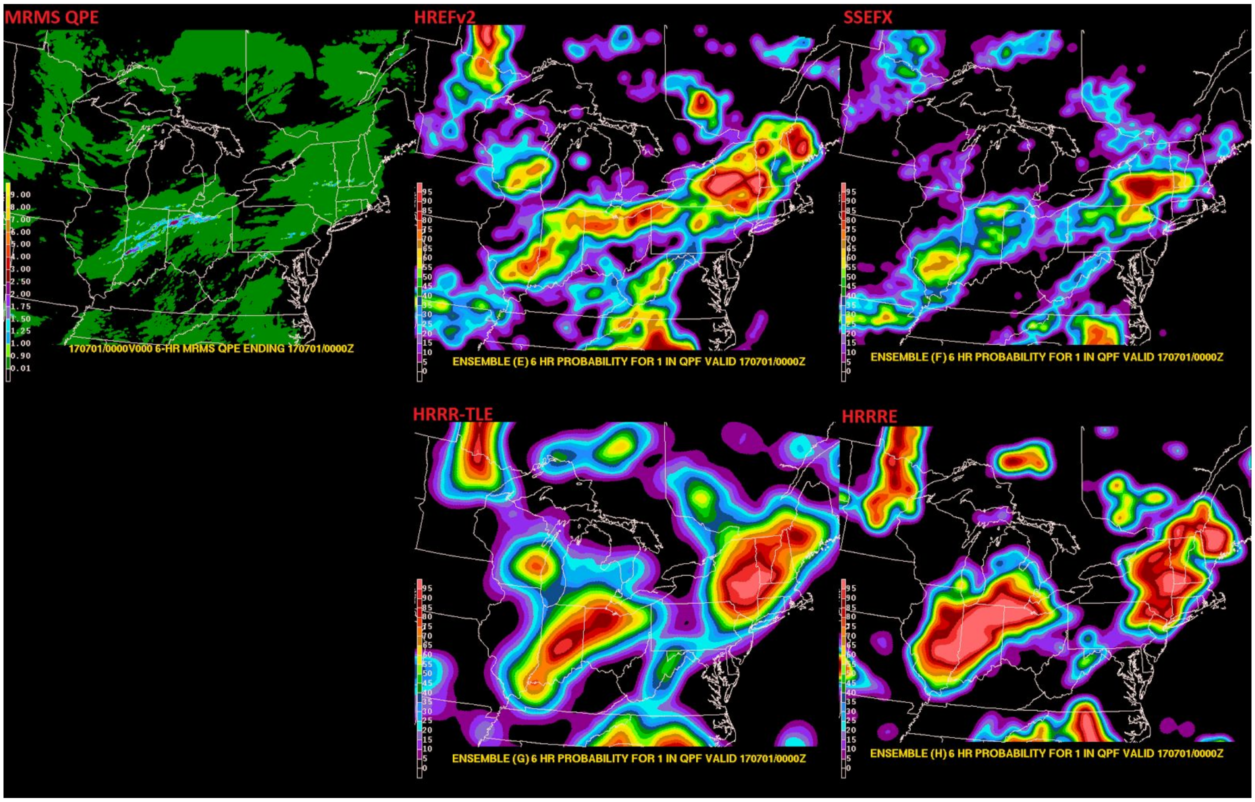

MRMS QPE
HREFv2
SSEFX
HRRR-TLE
HRRRE
170701/0000V000 6-HR MRMS QPE ENDING 170701/0000Z
ENSEMBLE (E) 6 HR PROBABILITY FOR 1 IN QPF VALID 170701/0000Z
ENSEMBLE (F) 6 HR PROBABILITY FOR 1 IN QPF VALID 170701/0000Z
ENSEMBLE (G) 6 HR PROBABILITY FOR 1 IN QPF VALID 170701/0000Z
ENSEMBLE (H) 6 HR PROBABILITY FOR 1 IN QPF VALID 170701/0000Z

Figure 22. *Five-panel display showing 6 HR MRMS QPE (top left) with one inch and above in blue and anything below an inch in green, HREFv2 (top middle), SSEFX (top right), HRRR-TLE(bottom middle), and the HRRRE (bottom right). All display a 40 km neighborhood probability of 1 inch QPF in 6 hours.*

Due to data availability issues during the experiment, the HREFv2, HRRR-TLE, and HRRRE had between 87 and 94 subjective scores, whereas the SSEFX was scored subjectively 120 times. During the first two weeks of the experiment, the HRRRE, SSEFX, and HREFv2 were using different post processing techniques and this caused the probabilities from the HRRRE to look a lot different than the other three. The HRRRE had no Gaussian smoother, the SSEFX was using a 40 km Gaussian smoother, and the HREFv2 used a 25 km Gaussian smoother. Because the HRRR-TLE is a time-lagged ensemble, the probabilities are post processed differently. The HRRR-TLE probabilities are the fraction of bias-corrected QPF grid points in a 100-km radius that have a threshold exceedance somewhere within 40 km, averaged across the three members. The results behave similarly to running a ~50-100km Gaussian filter on a grid of 40km neighborhood probabilities. Figure 23 shows an example of how the different post processing Gaussian smoothers can affect the visualization of the probabilities using the HRRRE as an example. Starting week 3, the post processing was normalized between those three ensembles; a 25 km Gaussian smoother was applied to the HRRRE, SSEFX, and HREFv2 during the post-processing of the probabilities. The HRRR-TLE post processing remained the same. In addition to the smoothing differences, there were also noticeably different atmospheric regimes during the first and second half of FFaIR. The first week of the experiment was dominated by Tropical Storm Cindy and a well defined synoptic frontal zone. The second week was not as active during the 18-00 UTC forecasts periods on which the ensembles were scored. Weeks three and four were mainly focused on monsoonal moisture as well as smaller, mesoscale convective systems. Due to these differences, subjective scores will be split between the first and second half of the experiment for this evaluation.

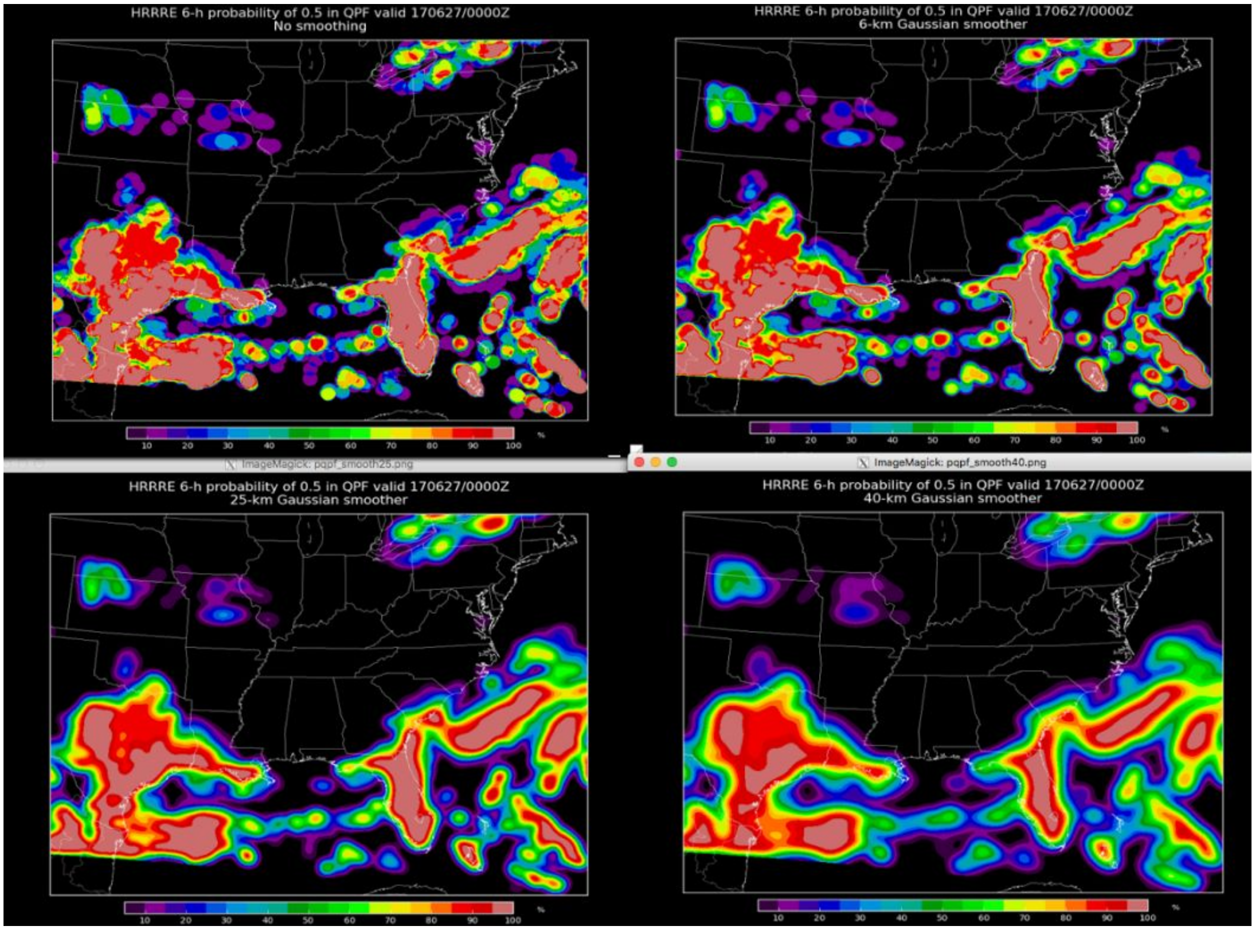

HRRRE 6-h probability of 0.5 in QPF valid 170627/0000Z
No smoothing
HRRRE 6-h probability of 0.5 in QPF valid 170627/0000Z
6-km Gaussian smoother
10 20 30 40 50 60 70 80 90 100 %
ImageMagick: pqpf_smooth25.png
ImageMagick: pqpf_smooth40.png
HRRRE 6-h probability of 0.5 in QPF valid 170627/0000Z
25-km Gaussian smoother
HRRRE 6-h probability of 0.5 in QPF valid 170627/0000Z
40-km Gaussian smoother

Figure 23. *Different post processing Gaussian smoothing applied to 6 hour probability of 0.5 inch QPF from the HRRRE. No smoothing (top left), 6 km Gaussian smoothing (top right) 25 km Gaussian smoothing (bottom left), and 40 km Gaussian smoothing (bottom right).*

Figure 24 shows the average subjective scores each model mean received each week over the course of the experiment. For all ensembles, the subjective scores for the half inch and the one inch threshold will be averaged as there were no significant differences between the two for any of the four ensembles.

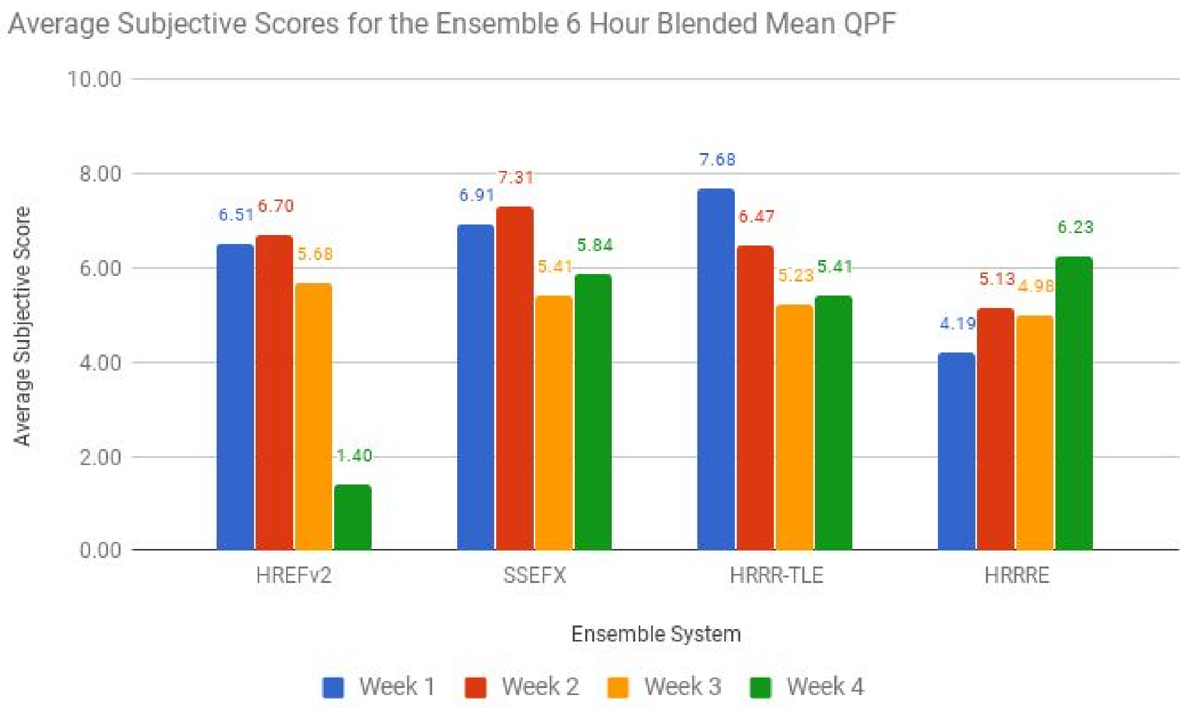


Figure 24. *Average subjective scores for the 6 hour blended mean QPF from the HREFv2, SSEFX, HRRR-TLE, and HRRRE for week 1 (blue), week 2 (red), week 3 (yellow), and week 4 (green).*

The HREFv2 had an average subjective score of 6.60 for the first half (weeks 1-2) and 5.09 for the second half (weeks 3-4) out of a possible 10. The HREFv2 was not available for all but one day of week 4, so the average of 1.40 represents just five participant scores for one case. The HREFv2 used a 25 km Gaussian smoother for both halves of the experiment, so the drop in scores in the second half of FFaIR could be attributed to a more unpredictable weather pattern and also fewer scores overall scores due to data availability issues.

The SSEFX had an average subjective score of 7.15 during the first half of the experiment and a 5.66 in the second half out of a possible 10. During weeks three and four, the SSEFX went from a 40 km to a 25 km Gaussian smoother. It is difficult to draw conclusions as to the cause of the decrease in scores for this ensemble without further testing as both the smoothing radius changed and weeks three and four presented more challenging weather patterns to model as mentioned previously.

The HRRR-TLE had an average subjective score of 7.03 during the first half of the experiment and a 5.31 in the second half out of a possible 10. Despite no post-processing changes being made to the HRRR-TLE during the experiment, a 1.72 point drop was still observed in subjective scores.

The last ensemble, the HRRRE, had an average subjective score of 4.65 during the first half of the experiment and a 5.66 in the second half out of a possible 10. No Gaussian smoother was applied to the post processing during the first half of the experiment and a 25 km smoother was applied during the second half. The subjective scores increased a full point and this was the only ensemble to increase from the first half to the second half. Having no smoothing in the post processing for these types of probabilities led to very sharp cut-offs in the probability field, as can be seen in the top left of Figure 23. Participants commented that this type of display is really difficult to use for a forecast. This exercise shows that even if an ensemble system is performing properly, the visualization of the output can significantly affect how the forecast is perceived by forecasters.

Outside of the post processing issues already detailed, general feedback and comments were positive, much like for the blended mean QPF from these same ensembles. Participants noted many times throughout the experiment that the SSEFX probabilities never increased above the 85-90% threshold. This was noticeable during many events where the other ensembles all would have probabilities of 95% or above. Subjectively, participants felt the probabilities were often too high for most of the cases across all the ensembles. Recognizing that a 40 km neighborhood probability approach will not show small details in convection, most participants viewed the overall spatial coverage of the probabilities from the ensembles positively for most cases.

HREFv2 Point Probability Schemes Results and Comments

Two different types of fractional coverage approaches for the generation of point probabilities were evaluated and scored subjectively each day alongside traditional point probabilities each day during the experiment. The first method expands the traditional point probability calculation to cover a 100-km radius around a given point, known as the “fractional probability scheme” in the experiment. The second method was based upon ensemble agreement scale (EAS) similarity criteria outlined in Dey et al. (2016) and varies the neighborhood radius size between 10 km and 100 km according to member to member similarity criteria. This method was referred to as the “EAS probability scheme” in the experiment. Both these schemes and the traditional point probabilities were derived from the HREFv2 ensemble system. More details on the two probability methods can be found in Appendix B under the HREFv2 description.

Each day the three probability schemes for one-half inch of QPF over six hours were displayed to the participants over the same limited domain and time range, 18-00 UTC, as the PFF forecast. Six hour MRMS QPE was used for verification and HREFv2 ensemble mean QPF (after week 2, when available) was also plotted as a reference. Participants would then subjectively rank each scheme on a scale of 1 (very poor) to 10 (very good) based on how well the probabilistic values represented what happened and comment on the utility of each probabilistic scheme. Figure 25 shows an example of how this question was presented to the participants during subjective evaluation.

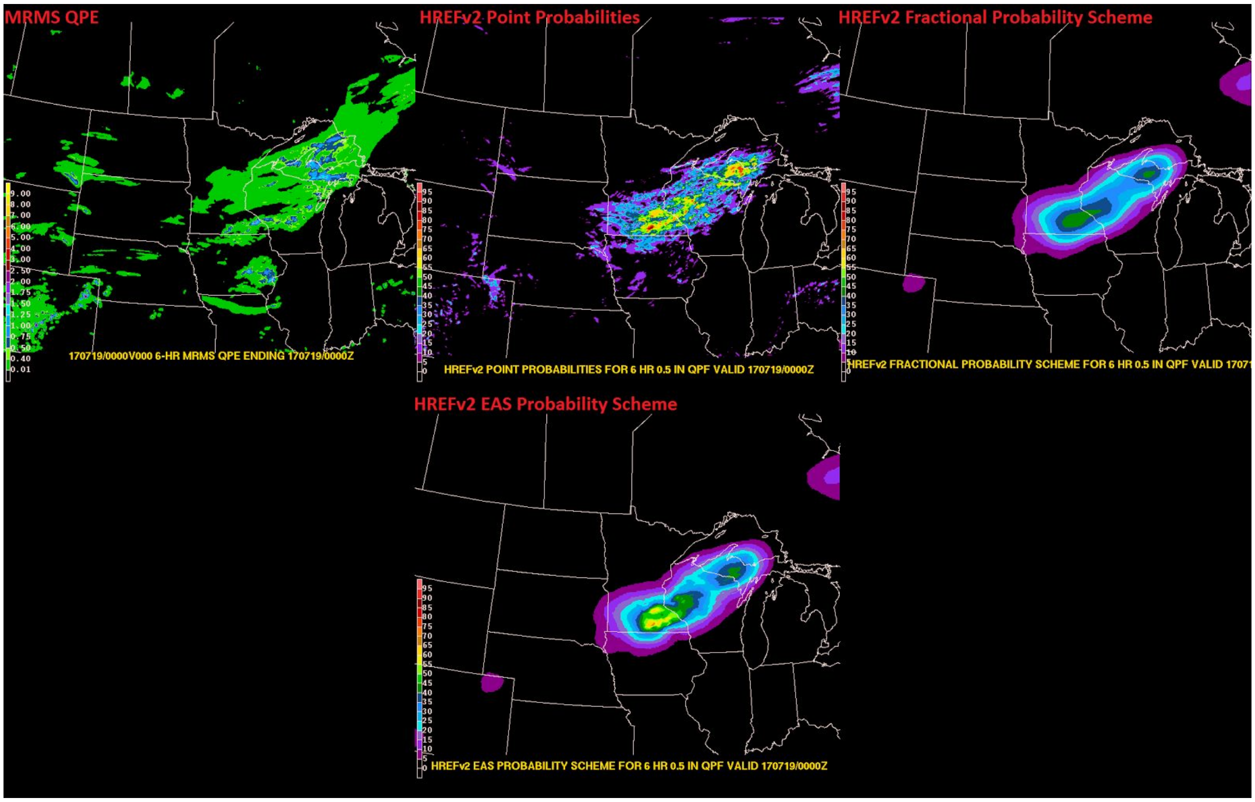

MRMS QPE
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.40
0.01
170719/0000V000 6-HR MRMS QPE ENDING 170719/0000Z
HREFv2 Point Probabilities
95 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0
HREFv2 POINT PROBABILITIES FOR 6 HR 0.5 IN QPF VALID 170719/0000Z
HREFv2 Fractional Probability Scheme
95 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0
HREFv2 FRACTIONAL PROBABILITY SCHEME FOR 6 HR 0.5 IN QPF VALID 1707
HREFv2 EAS Probability Scheme
95 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0
HREFv2 EAS PROBABILITY SCHEME FOR 6 HR 0.5 IN QPF VALID 170719/0000Z

Figure 25. *A four panel image showing six hour MRMS QPE filtered for half inch and above (blue) and all amounts below (green) in the top right, HREFv2 traditional point probabilities top middle, HREFv2 fractional probability scheme top right, and the HREFv2 EAS probability scheme bottom middle.*

Figure 26 shows a box plot of all the subjective scores for the experiment for all three point probability methods. The traditional point probabilities subjectively had the highest average score of 6.44 out of 10 with a standard deviation of 1.83. Out of the two experimental schemes, the EAS method has the highest average score, a 6.31 out of 10 with a standard deviation of 1.69. Finally, the fractional scheme with a fixed 100 km radius had the lowest average subjective score at 5.42 out of 10 with a standard deviation of 1.31.

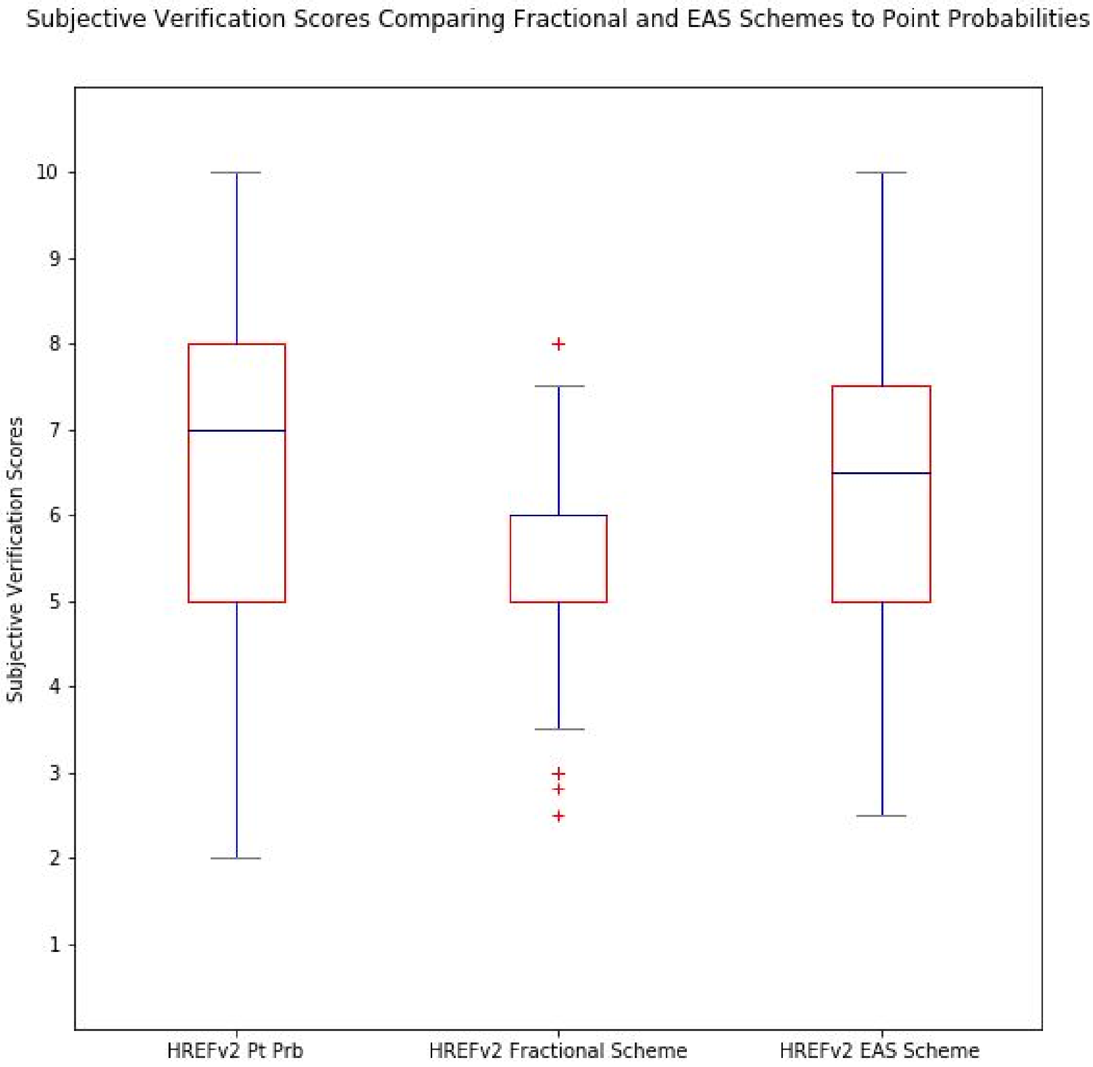


Figure 26. *Box plot of the subjective verification scores of the three HREFv2 point probability schemes including the traditional point probabilities, fractional scheme, and EAS scheme over the entire experiment.*

Evaluating the three different point probability schemes generated a lot of discussion each day. Because the probabilities were ultimately reliant on the underlying forecast from the HREFv2, the comments focused more on how the three different point probability methods visually represented the probabilistic field and the utility they provided to a forecaster in the forecasting process. The traditional point probabilities that had no filters applied were generally favored due to the amount of detail provided within the probability field. The detail was usually preferred by the forecaster participants compared to the model developers that participated. They were particularly useful in situations that showed smaller scale convection, such as diurnally driven or sea-breeze based convection that occurred often in the Southeast. In these cases, the traditional point probabilities typically had some signal whereas the other two schemes may have had no signal. The fractional coverage scheme with a fixed radius of 100 km was generally the least favored because participants felt that it smoothed the probability field too much and the probabilities were often much too low in most cases. The EAS probability scheme garnered positive feedback from participants as the variable filter, based on model agreement, allowed for more detail in certain situations that was often appreciated by the forecasters. Both filtered fields were preferred by forecasters as a tool to use when forecasting QPF amounts as opposed to areas of flash flooding, as the traditional point probability field was described as too noisy for drawing QPF contours. A suggestion from feedback received is to reduce the lower bound radius of the EAS scheme below 10 km in order to produce a compromise between the extreme detail of the traditional point probabilities and the more smoothed fields of the filtered options. Figure 27 shows two examples of the fractional coverage and EAS probability schemes, one (panels 1 and 2) where there is very little difference between the two and a second (panels 3 and 4) where the EAS provides more detail compared to the fractional, fixed 100 km radius.

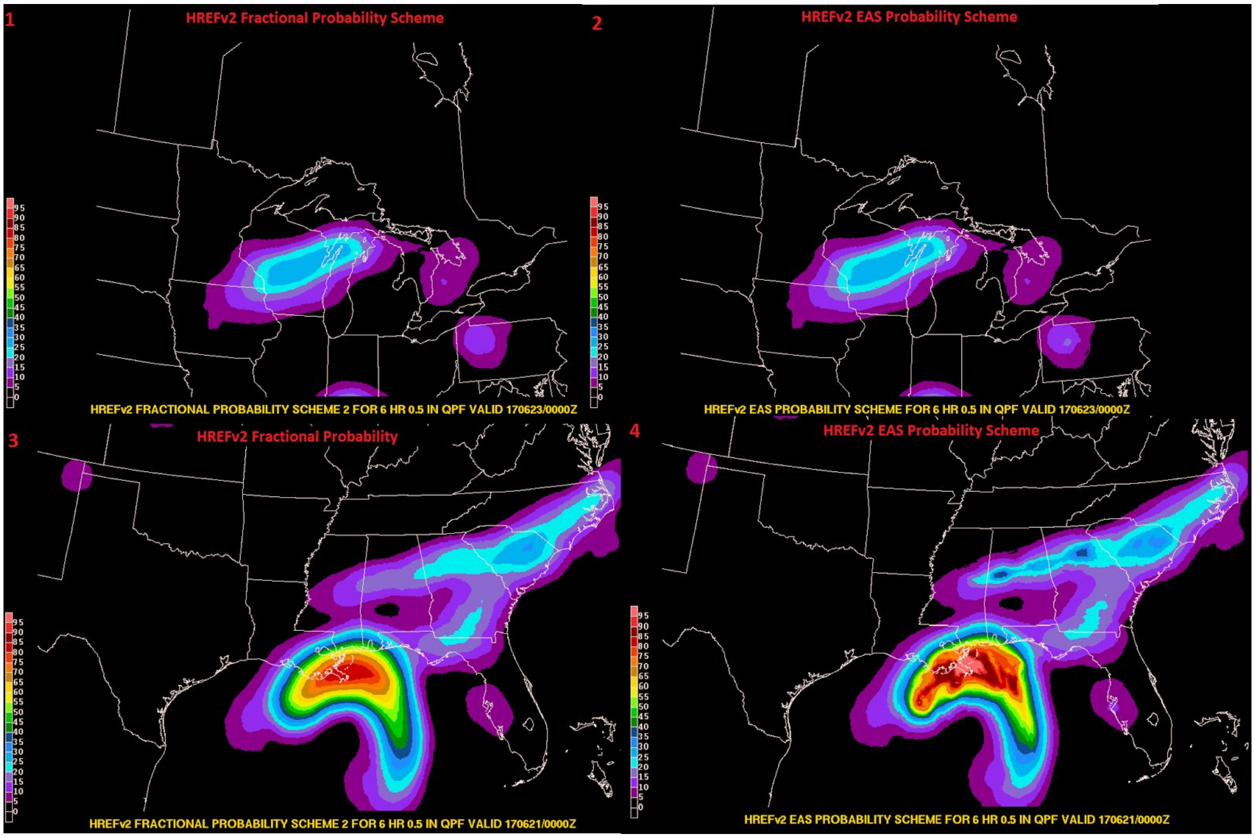

1
HREFv2 Fractional Probability Scheme
HREFv2 FRACTIONAL PROBABILITY SCHEME 2 FOR 6 HR 0.5 IN QPF VALID 170623/0000Z
2
HREFv2 EAS Probability Scheme
HREFv2 EAS PROBABILITY SCHEME FOR 6 HR 0.5 IN QPF VALID 170623/0000Z
3
HREFv2 Fractional Probability
HREFv2 FRACTIONAL PROBABILITY SCHEME 2 FOR 6 HR 0.5 IN QPF VALID 170621/0000Z
4
HREFv2 EAS Probability Scheme
HREFv2 EAS PROBABILITY SCHEME FOR 6 HR 0.5 IN QPF VALID 170621/0000Z
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0

Figure 27. *HREFv2 probability for a half inch of precipitation over 6 hours using the fractional probability scheme (1) and EAS probability scheme (2) valid at 00 UTC on June 23, 2017 and the fractional probability scheme (3) and EAS probability scheme (4) valid at 00 UTC on June 21, 2017.*

<u>National Blend of Models, Version 3</u>

The National Blend of Models version 3 (NBMv3) was evaluated during the 2017 FFaIR Experiment. At the time of the experiment, the parallel version of NBMv3 was used as it was not yet operational. More information on the NBMv3 can be found in Appendix B. Hours 18-24 of the 6-hour QPF forecast from the 00Z cycle of the NBMv3 was subjectively evaluated on a 1 (very poor) to 10 (very good) scale. Participants were encouraged to comment on the overall magnitude and spatial extent of the QPF. Figure 28 shows an example of the evaluation.

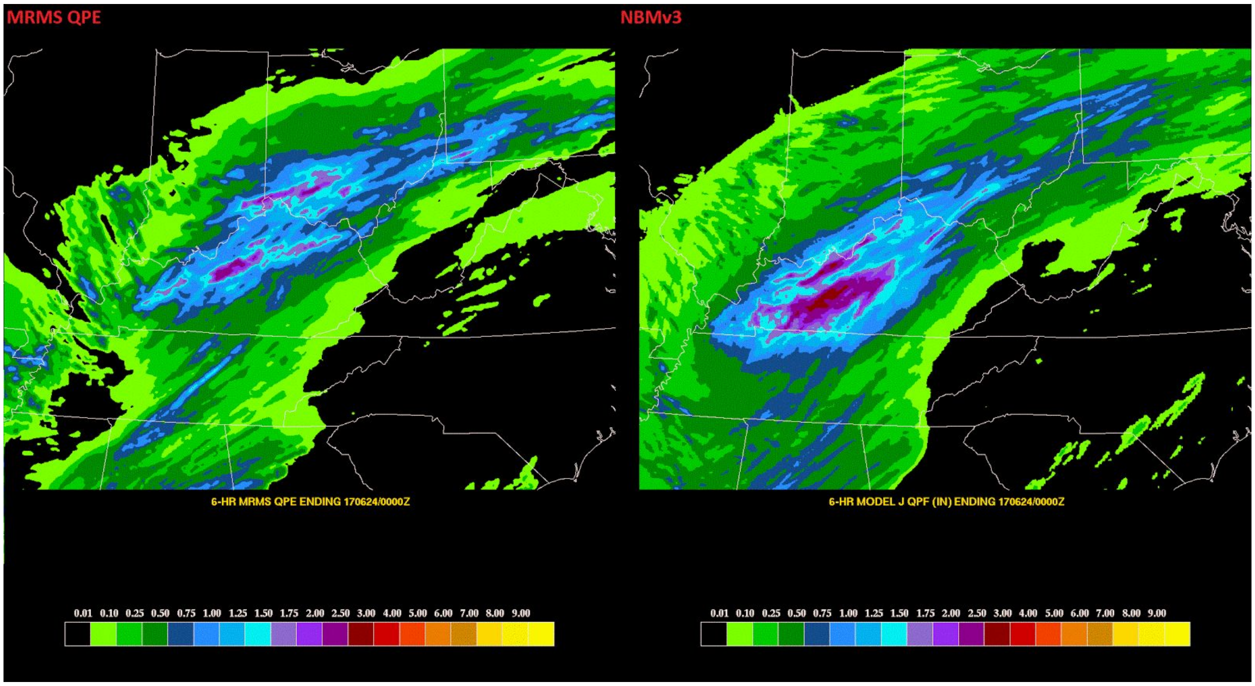


Figure 28. *Two panel display showing 6-hour MRMS QPE on the left and 6-hour NBMv3 QPF on the right valid at 00 UTC on June 24, 2017.*

Figure 29 shows a box plot of all the subjective scores for the 6 hour QPF from the NBMv3 over the course of the experiment. The NBMv3 had an average subjective score of 5.41 out of 10, a median of 5, and a standard deviation of 1.69. Figure 30 shows a distribution of how many of each individual score was assigned to the NBMv3 during each of the four weeks.

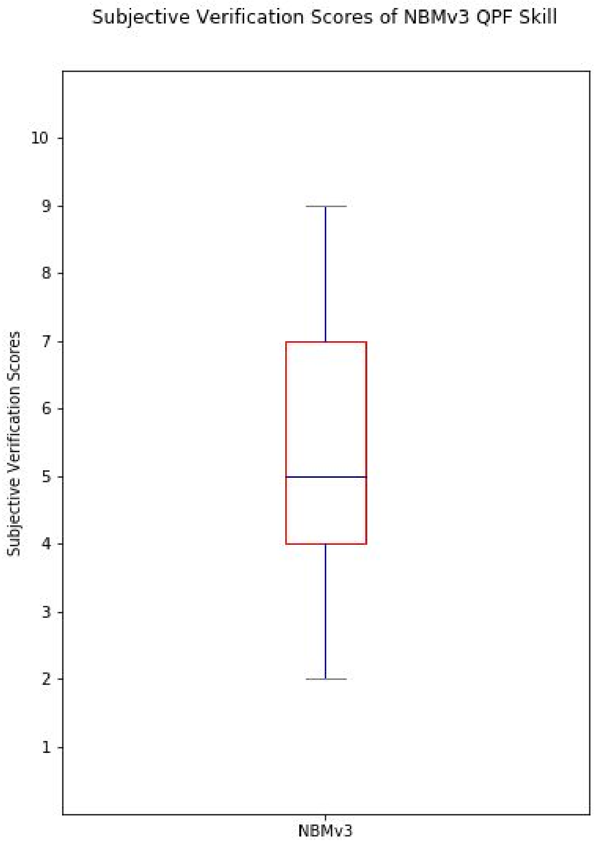


Figure 29. *Subjective scores for the 6 hour QPF from the NBMv3 over the 2017 Experiment.*

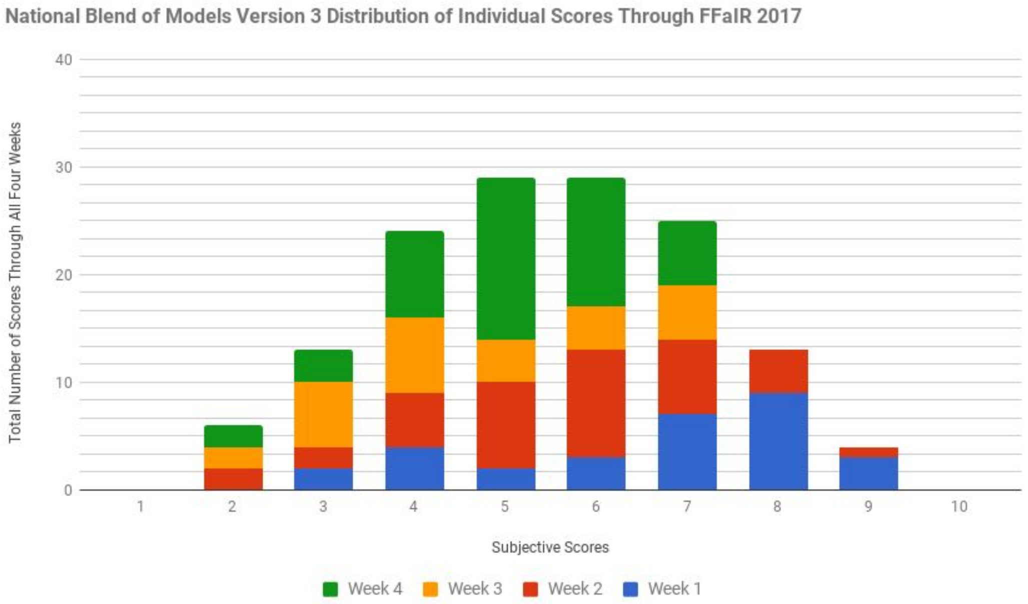


Figure 30. *The individual subjective score distribution over all four weeks of FFaIR for the NBMv3. Week 1 scores are blue, week 2 scores red, week 3 scores are yellow, and week 4 scores are green.*

Figure 31 shows a Roebber Performance Diagram for the skill of the half inch, 6 hour QPF forecast from the NBMv3 over the course of the experiment. The CSI value for the half inch threshold was 0.08 and there was a low frequency bias of approximately 0.45. At the one inch value shown in Figure 32, the CSI value (0.03) and frequency bias (0.24) are much lower for the NBMv3.

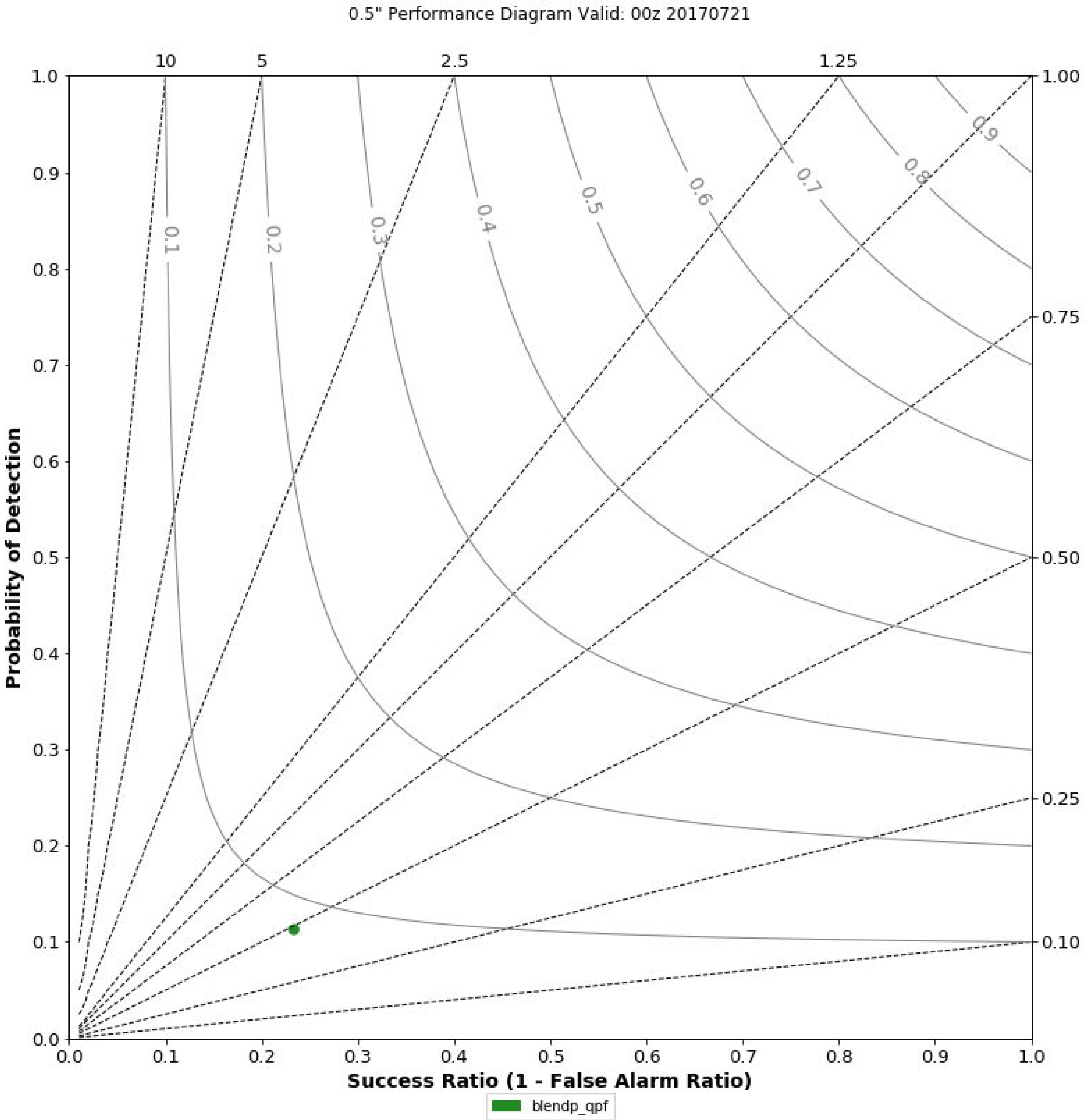


Figure 31. *Roebber Performance Diagram showing the skill of the half inch, 6 hour QPF forecast from the NBMv3 over the whole experiment.*

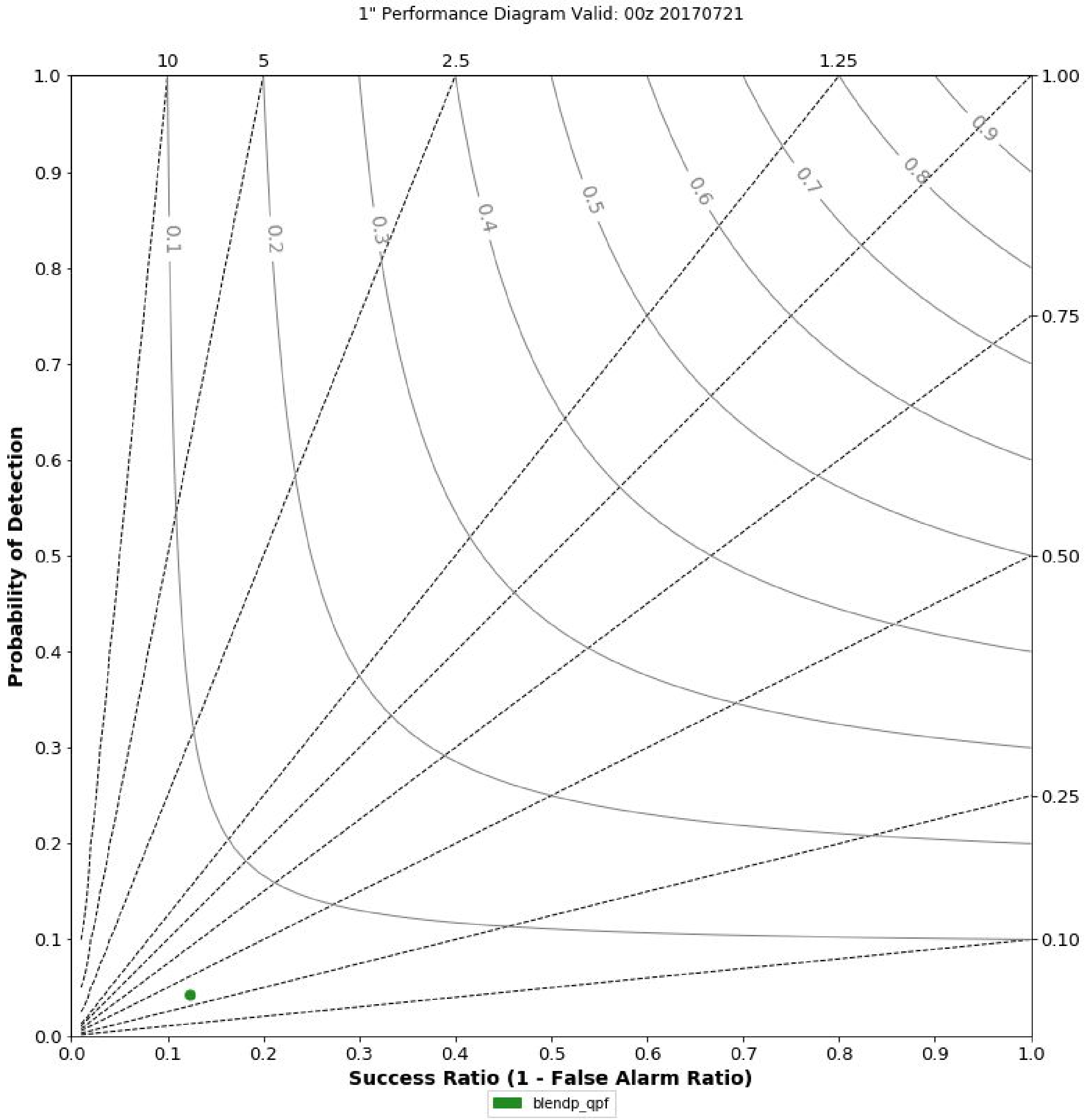


Figure 32. *Roebber Performance Diagram showing the skill of the one inch, 6 hour QPF forecast from the NBMv3 over the whole experiment.*

According to the participants, the most common problem with the NBMv3 QPF was a low bias for the majority of the cases. It did do quite well with both magnitude and spatial extent during the first week of the experiment handling the moisture of Tropical Storm Cindy. However, during weeks 3 and 4, especially when the cases were more mesoscale in nature such as Southwest monsoon moisture or MCS activity, participants often noted the QPF was too low when compared to observations. This difference can be seen in Figure 30 where scores of 7, 8, and 9 were more often given in weeks 1 and 2 compared to weeks 3 and 4. However, the overall shape and spatial coverage was often quite good and gave an indication that there would be precipitation in areas that received rain. Figure 33 shows an example where the NBMv3 was considered underdone.

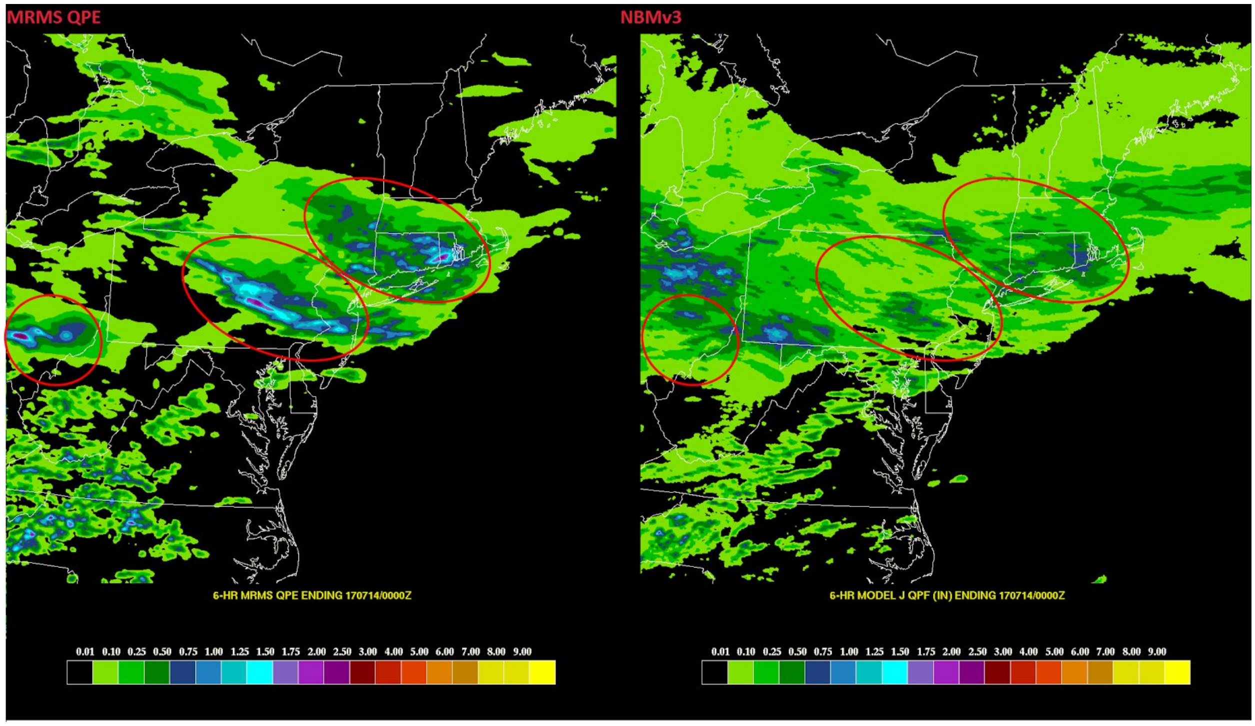


Figure 33. *6-hour MRMS QPE (left) compared to 6-hour NBMv3 QPF (right) both valid at 00 UTC on July 14, 2017. Circles highlight areas where the NBMv3 (right) had lower QPF values than what was observed in the MRMS QPE (left).*

5. Hydrologic Guidance Results

Experimental National Water Model Products

Several experimental products derived from the National Water Model were available to participants to use during the forecast activities. At the end of each week, the participants were asked to provide written feedback about the products collectively.

1. *High Flow Potential*

The High Flow Potential product depicts areas where current streamflow anomalies are in the Moderate, High, and Very High stages. This product forces the underlying water model using observational rainfall data sets and is updated every hour. The streamflow anomalies are calculated from climatological annual average flow values provided by NHDPlus dataset. The idea is to draw attention to areas over which there is high flow/flow above bankfull conditions for situational awareness.

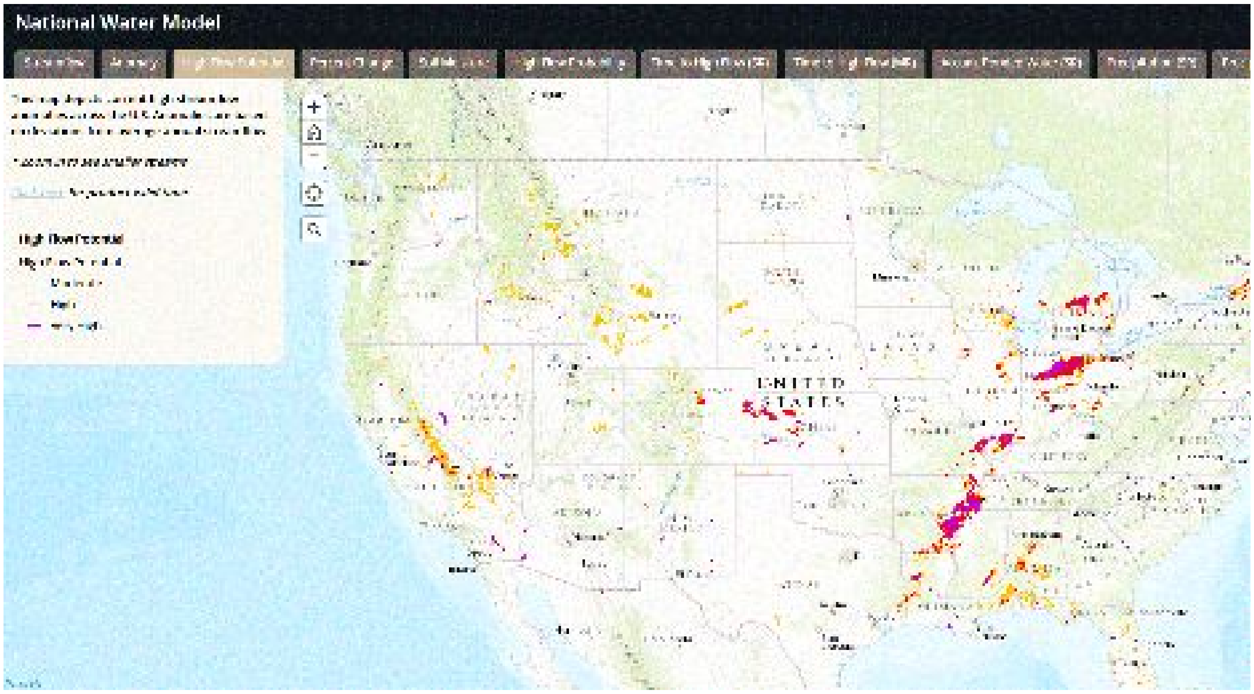


Figure 34. *Experimental National Water Model High Flow Potential product valid 11Z on June 21, 2017.*

Many participants felt that the High Flow Potential product could be used as another situational awareness tool, but that overall it did not provide any information that was new or not obvious from other guidance. This high-flow stream information would be more valuable if it were combined with current radar estimates and indications of areas currently under flash flood warnings.

2. *High Flow Probability*

The High Flow Probability product from the National Water Model is informed by an ensemble of time-lagged operational HRRR forecasts of quantitative precipitation and updates every hour. The first attempt at a probabilistic National Water Model product, it depicts the probability of streams exceeding high flow within the next 6-8 hours.

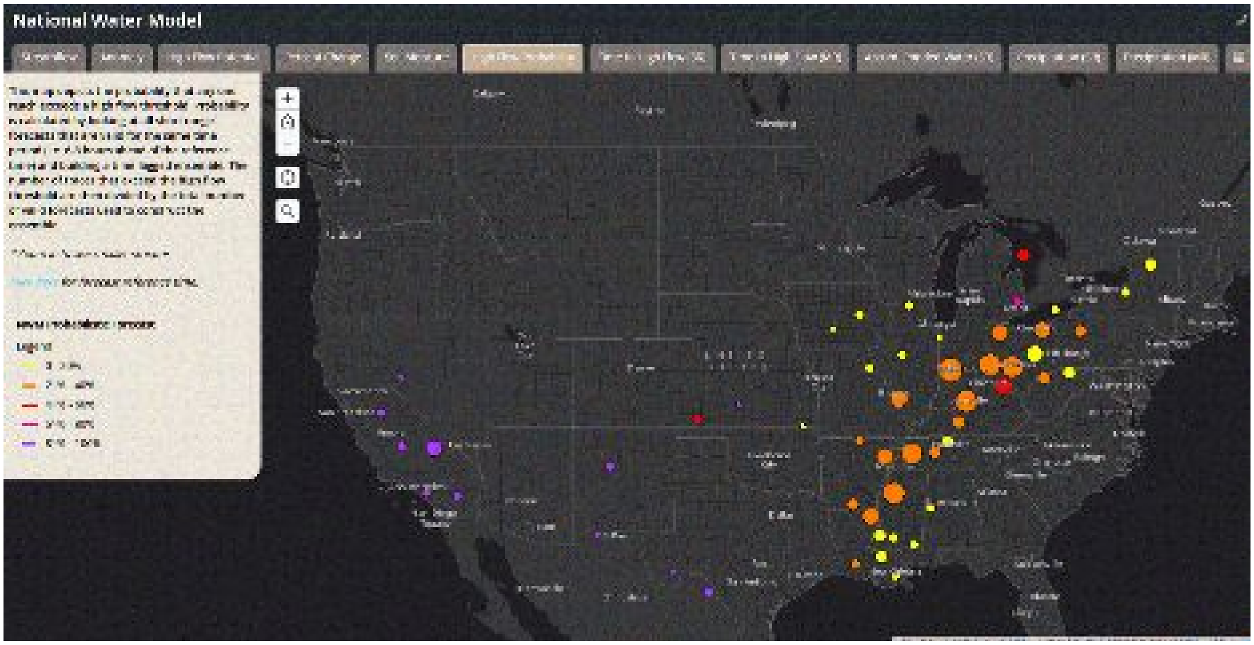


Figure 35. *Experimental National Water Model High Flow Probability at the outmost zoom level displays clusters of river segments with probabilities > 0.*

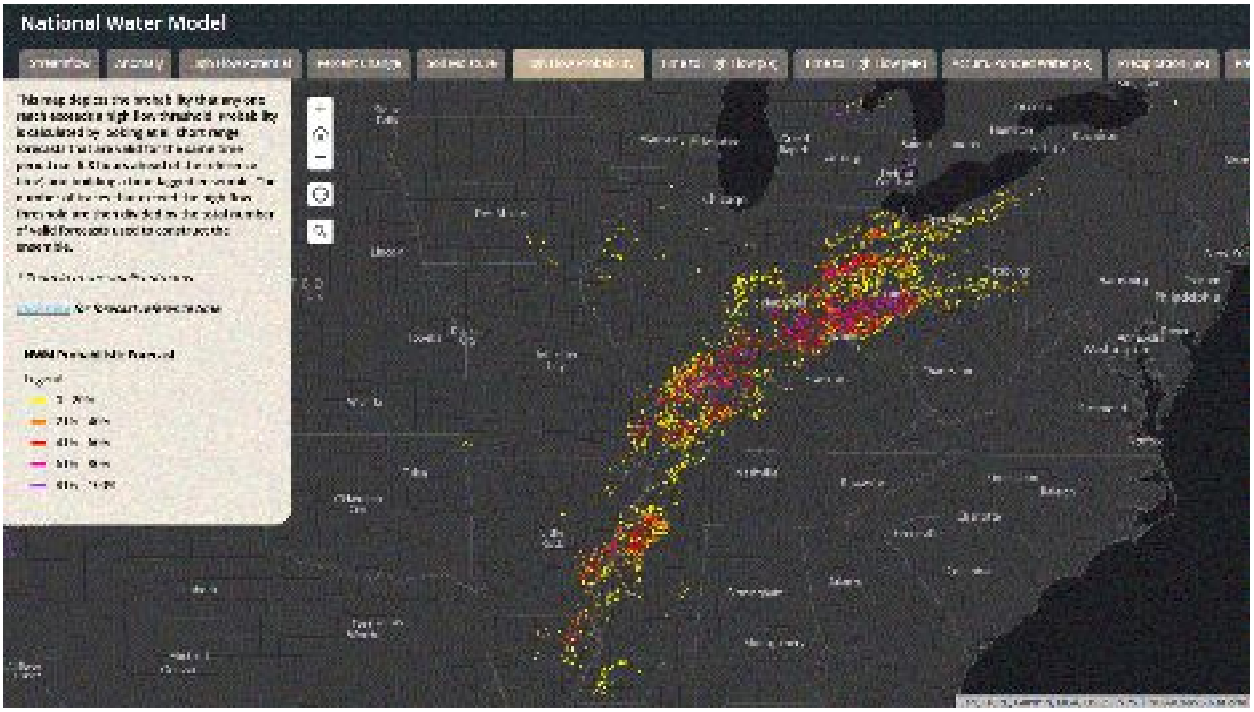
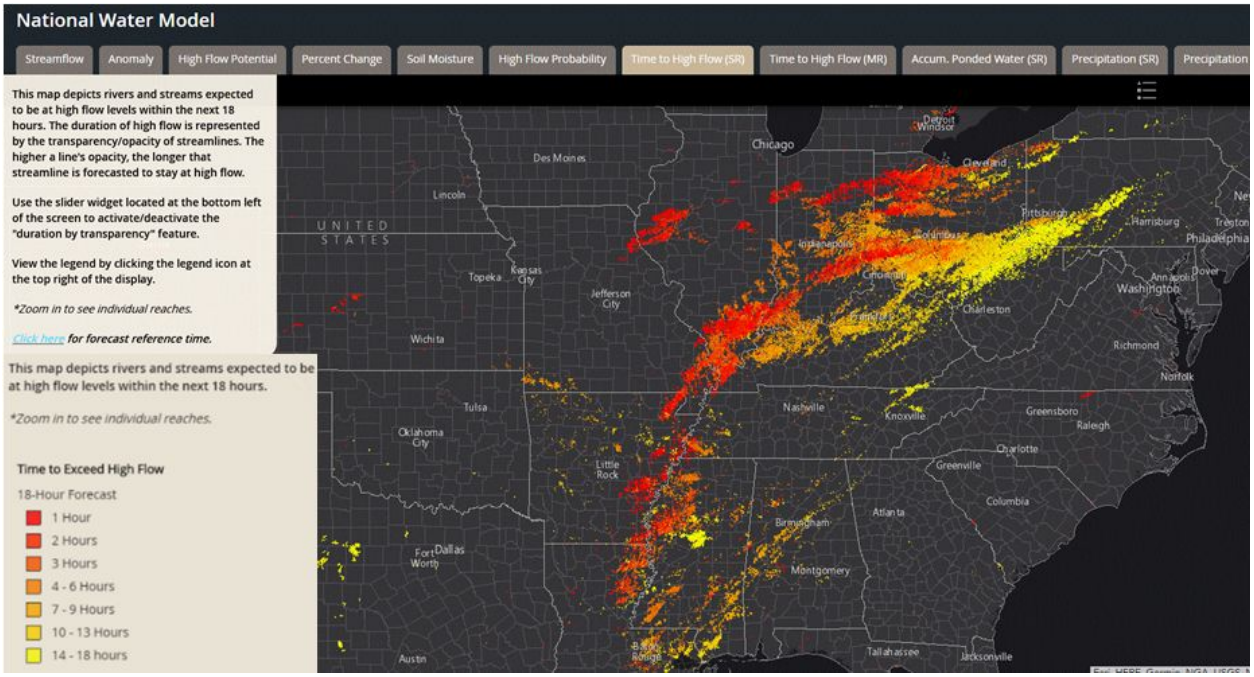


Figure 36. *Experimental National Water Model High Flow Probability gives high resolution detail of the streams that are probabilistically forecast to exceed high flow, color coded in increments of 20%, valid 11Z on June 21, 2017.*

Participants overall agreed that probabilistic output from the National Water Model would produce the most valuable forecast information from its guidance. However, this next-6-to-8-hour High Flow Probability product presented inconsistencies among the other products offered. For example, the probabilities of high flow tended to be low in areas where the High Flow Potential and deterministic short-range Time to High Flow (next section) showed a strong response for high flow. Although more spread in the guidance is preferred, additional concerns were being limited by the precipitation forecast from one model ensemble. There is much anticipation from the hydrology community to see more probabilistic guidance emerge from the National Water Model.

3. *Time to High Flow*

The experimental Time to High Flow product is available for both the short range (18-hour forecast informed by the HRRR precipitation forecast and updated hourly) and the medium range (72-hour forecast informed by the GFS precipitation forecast and updated 6-hourly). This product depicts the rivers and streams that are expected to be at high flow levels within the forecast period. The color scale indicates the time step or increment during which high flow is predicted to occur within that forecast period.


National Water Model

This map depicts rivers and streams expected to be at high flow levels within the next 18 hours. The duration of high flow is represented by the transparency/opacity of streamlines. The higher a line's opacity, the longer that streamline is forecasted to stay at high flow.

Use the slider widget located at the bottom left of the screen to activate/deactivate the "duration by transparency" feature.

View the legend by clicking the legend icon at the top right of the display.

*Zoom in to see individual reaches.

Click here for forecast reference time.

This map depicts rivers and streams expected to be at high flow levels within the next 18 hours.

*Zoom in to see individual reaches.

Time to Exceed High Flow

18-Hour Forecast

- 🟥 1 Hour
- 🟥 2 Hours
- 🟧 3 Hours
- 🟧 4 - 6 Hours
- 🟧 7 - 9 Hours
- 🟨 10 - 13 Hours
- 🟨 14 - 18 hours

Figure 37. *Experimental National Water Model Time to Exceed High Flow colors streams and rivers that are forecast to exceed high flow within the forecast period. This short range forecast is valid 11Z June 21 - 06Z June 22, 2017.*

The Time to High Flow product, both for short and medium range, was the most widely understood and applicable National Water Model guidance provided in the FFaIR Experiment. Forecasters felt this product was helpful by depicting how long it takes for streams to respond and how much they will be impacted by the rainfall. This knowledge provides details about potential flooding impacts and timing which is useful in messaging and decision support.

Concerns were expressed about the display of the information in the product. Although the color scale is intended to represent the time within the forecast period that the flow response is to occur, once colored for a response in the early part of the time period, the colors do not change or propagate in time to communicate whether or not rivers may remain at high flow throughout the the entire forecast time period. Conversely, if high flow is occurring in the beginning of the forecast period, it is not clear how long that high flow will persist throughout the period. It is recommended that a different time representation be used for this product, such as time steps or animation.

Additionally, forecasters wanted to see the precipitation forecast overlaid with the stream response, and an indication of time to bankfull rather than just high flow.

4. *Ponded Water*

The experimental Ponded Water, a 6-hour product available for the short range (18-hour forecast), is informed by the HRRR precipitation forecast, and updated hourly. Ponded water depicts the depth of rainfall accumulated over 6 hours (i.e. 1-6, 7-12, 13-18 hours ahead) which is not being absorbed into the soil nor evaporating. This represents depth in inches over a 250-meter grid cell and not actual depth of ponding at any given location. The Ponded Water product could be interpreted as a proxy to water inundation or a raw model representation of potential runoff.

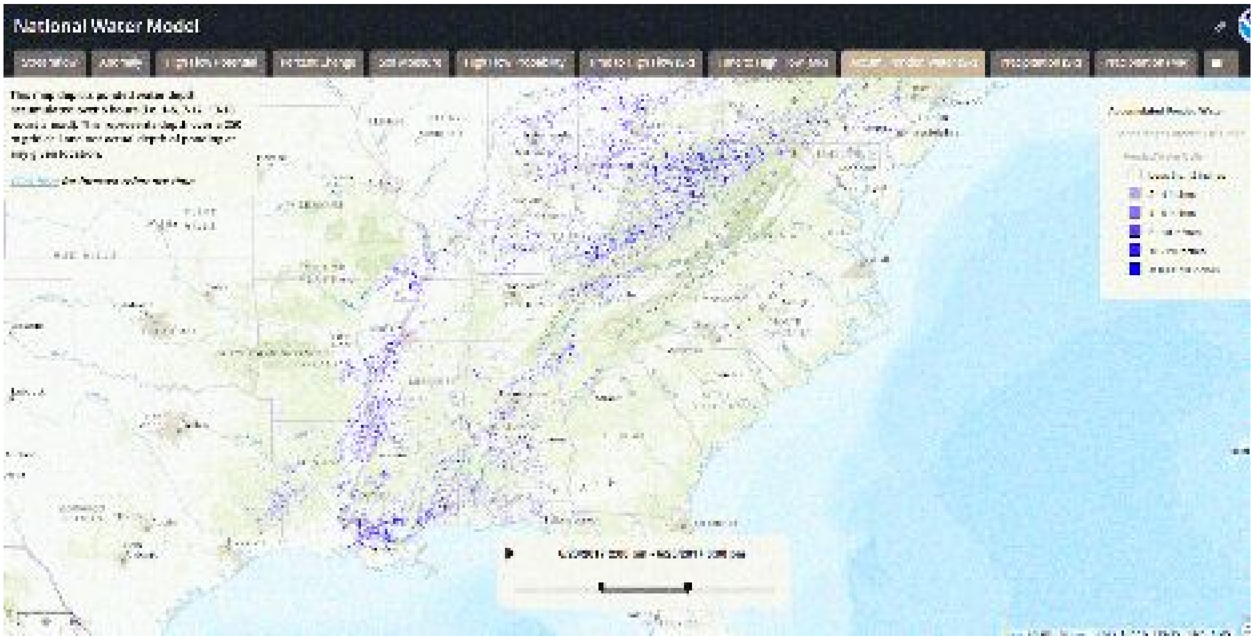


Figure 38. *Experimental National Water Model Ponded Water depicting areas over which water is piling up due to heavy rainfall, valid 18Z 21 June - 00Z 22 June, 2017.*

The Ponded Water product intends to indicate areas over which the HRRR model precipitation may accumulate and lead to overland flooding. The participants struggled with identifying ways to discern meaning from the data and apply it to the forecast process. In its current state, it can be another display for situational awareness but does not provide any new or different information than other available tools. However, the data has the potential to evolve into a more meaningful product such as inundation.

The depiction of grid cells was often described as "peppery," and not representative of actual ground response to the rainfall. Additionally, the various shades of blue on the depth scale made it difficult to distinguish between the different depths. And depth over a 250 m grid box within the model as a scale for ponding rainfall was very difficult for the forecasters to understand in a way that would make the product useful guidance for flash flood prediction. Many participants desired a probabilistic product that would increase data spread, areal coverage and utility. However, others countered that a probabilistic product with a neighborhood filter would force a tradeoff with the river and stream detail so valuable from the National Water Model itself.

<u>HRRR Time-Lagged Ensemble QPF on Saturated Soil Product</u>

The HRRR Time-Lagged Ensemble (HRRR-TLE) QPF on saturated soil product (Figure 39, left) was available for scoring on only 7 days (79 total participant scores) during the 2017 FFaIR Experiment, resulting in an average of 5.23 out of 10.

Participants liked the idea of the probabilistic product showing where heavy rains falling on saturated soil may run off and cause flooding and agreed that the product had potential for situational awareness. But participants needed to develop confidence in this tool and often desired to see the parameters that comprised the product (the QPF and the soil saturation) to determine whether or not the resulting probabilities made sense.

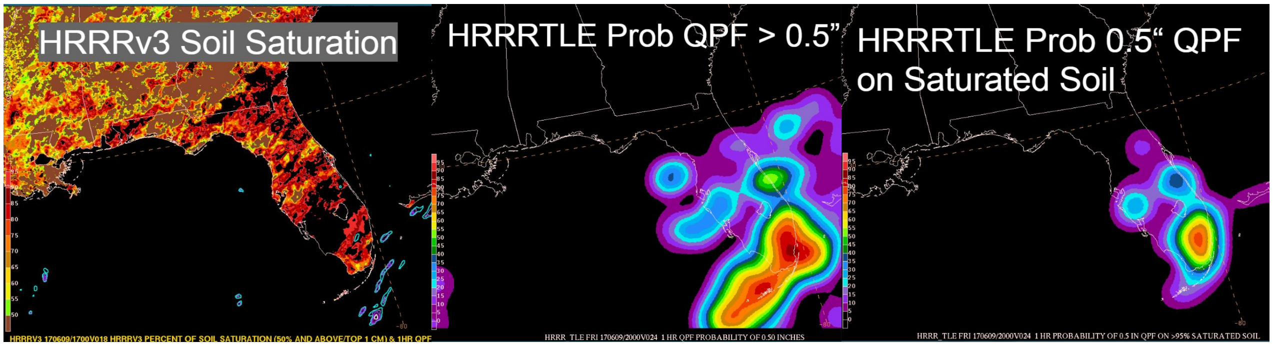

HRRRv3 Soil Saturation
HRRRTLE Prob QPF > 0.5"
HRRRTLE Prob 0.5" QPF on Saturated Soil
HRRRV3 170609/1700V018 HRRRV3 PERCENT OF SOIL SATURATION (50% AND ABOVE/TOP 1 CM) & 1HR QPF
HRRR_TLE FRI 170609/2000V024 1 HR QPF PROBABILITY OF 0.50 INCHES
HRRR_TLE FRI 170609/2000V024 1 HR PROBABILITY OF 0.5 IN QPF ON >95% SATURATED SOIL

Figure 39. *An example of the components that comprise the HRRR-TLE probability of QPF on saturated soil product (right) including the HRRRv3 percent soil saturation at the top 1 cm (left) and HRRR-TLE probability of precipitation (center).*

The probability of QPF on saturated soil was very often successful in not producing a probability where the QPF was occurring but soil was not saturated. However, there were several instances when the probabilities were displaced from the flooding events or presented a probabilistic high bias. This reduced confidence in the resulting forecasts. The consensus was that the participants needed more events and to have this product set in context with other guidance to develop applicability and messaging for potential flash flooding.

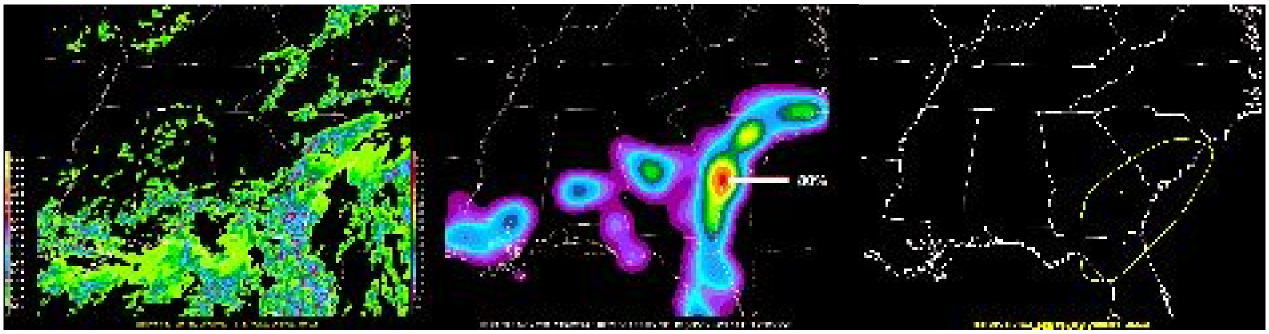


Figure 40. *An example of the total 6-hour HRRRv3 QPF (left), 1-hour HRRR-TLE probability of 1 inch of QPF on Saturated Soil within that 6-hour period (center), and the resulting 6-hour experimental probability of flash flooding (PFF) forecast (right) created by the FFaIR participants valid 00Z July 18, 2017.*

Traditional QPF forecasters tend to be more comfortable with numerical weather prediction (NWP) than with hydrologic models, so this representation was easy for them to consume and apply to the forecast process. However, a caveat of this product is that the land surface models (LSMs) do not route water or have the fine scale mapping of variable soil conditions found in hydrologic models. A hybrid product that allows forecasters to see all of the elements together (QPF, soil state, exceedance probabilities) would be ideal for most efficient flood prediction.

Colorado State University Machine Learning Probabilities First Guess Field

A new experimental tool from researchers at Colorado State University (CSU), referred to here as the CSU-Machine Learning Probabilities (MLP) First Guess Field, was subjectively evaluated each day during the experiment. The tool is a prediction system comprised of random forests which is trained with 11 years (January 2003 - August 2013) of Days 2 and 3 GEFS Reforecast (GEFS/R) data. More details on the system and the machine learning and decision tree aspects can be found in Appendix B. The output available during FFaIR displayed the Day 2 (60 hour) or Day 3 (84 hour) forecast of the probability of 24 hour precipitation exceeding the one or ten year 24 hour average recurrence interval (ARI) within 40 km of a point. For subjective evaluation, only the probability of exceeding the one year, 24 hour ARI was shown to participants and given a subjective score of 1 (very poor) to 10 (very good). The UFV system and 24 hour MRMS QPE were used as verification and participants also were able to see the Day 2 or Day 3 experimental ERO that was valid over the same time period. This method for verification was used rather than directly comparing to 24 hr QPE exceeding the 1 year 24 hr ARI in order to gather feedback on its utility as a first guess field for the ERO. Participants were asked to comment on the overall quality of the CSU-MLP First Guess Field. Specific areas of interest included the probabilistic representation as compared to both the verification and the experimental ERO forecasts as well as the overall spatial coverage of the probabilities. Participants were also encouraged to comment on whether this tool would be useful as a first guess field for a WPC forecaster that is tasked with creating the ERO product. Figure 41 shows an example of how the tool was presented to participants for subjective evaluation.

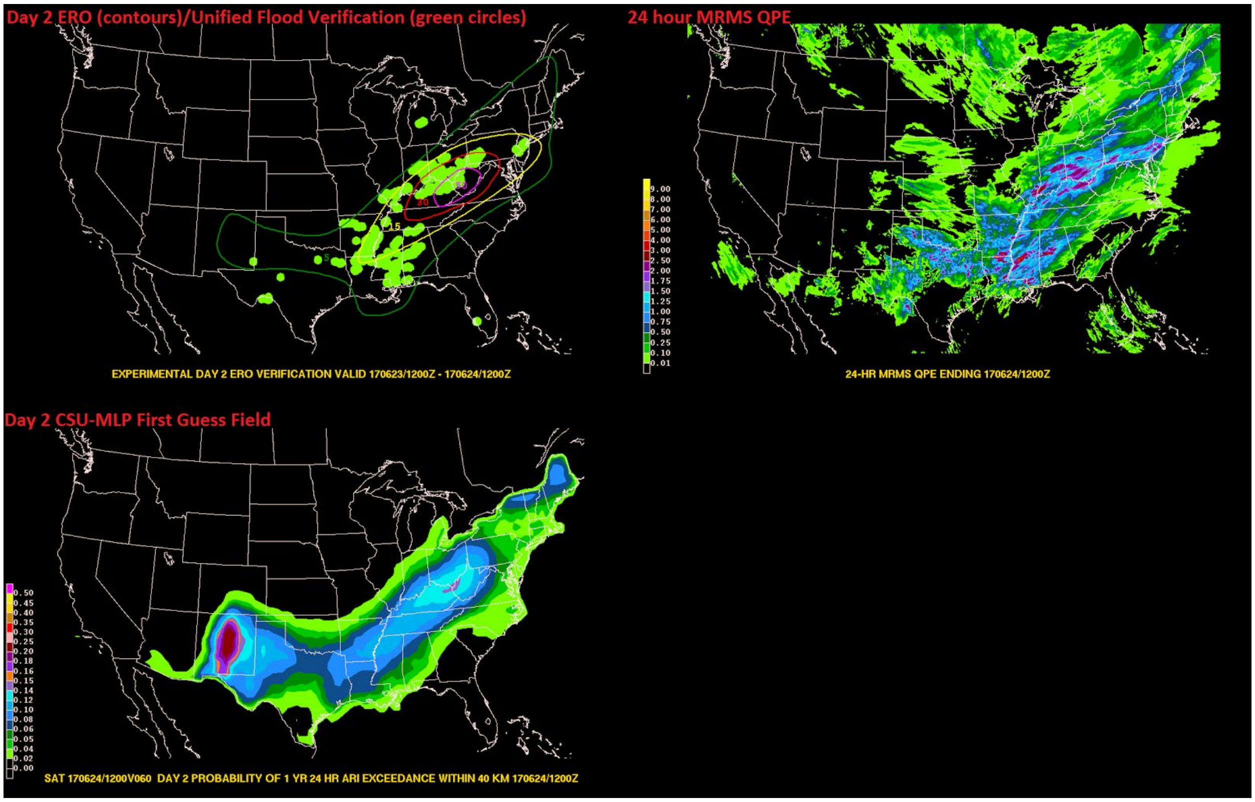

Day 2 ERO (contours)/Unified Flood Verification (green circles)
EXPERIMENTAL DAY 2 ERO VERIFICATION VALID 170623/1200Z - 170624/1200Z
24 hour MRMS QPE
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
24-HR MRMS QPE ENDING 170624/1200Z
Day 2 CSU-MLP First Guess Field
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.18
0.16
0.15
0.14
0.12
0.10
0.08
0.06
0.05
0.04
0.02
0.00
SAT 170624/1200V060 DAY 2 PROBABILITY OF 1 YR 24 HR ARI EXCEEDANCE WITHIN 40 KM 170624/1200Z

Figure 41. *Three panel image showing the Day 2 ERO (contoured) and UFV (green circles) (top left), 24 hour MRMS QPE (top right), and the Day 2 (60 hour) CSU-MLP First Guess Field forecast of the probability of 1 year 24 hour ARI exceedance within 40 km of a point (bottom left), all valid at 12 UTC June 24, 2017. For the ERO in this example a 50% probability contour was issued over Kentucky and West Virginia. Peak probabilities in the CSU-MLP CSU-MLP First Guess Field were between 20-25% in the New Mexico area.*

Figure 42 shows the average subjective score for both the Day 2 and Day 3 CSU-MLP First Guess Field. The Day 2 First Guess Field had an average subjective score of 5.82 and standard deviation of 1.35. The Day 3 First Guess Field scored higher with an average subjective score of 6.10 and a standard deviation of 1.25. Both these averages were below the experimental Day 2 and 3 ERO average subjective scores (more detailed results on the ERO can be found in the next section).

From the comments and feedback, participants generally felt the CSU-MLP tool was a viable first step in providing a first-guess-like field for the creation of the ERO product. A first guess field does not currently exist and has been highly requested by WPC forecasters. Participants found the overall spatial coverage of the 1 year 24 hour ARI exceedance probabilities to be quite good in most cases and a good indicator of areas of flood potential. There were areas, however, that had too much areal coverage on several days. These were mainly in the High Plains region and included areas of Montana, North Dakota, and South Dakota most often.

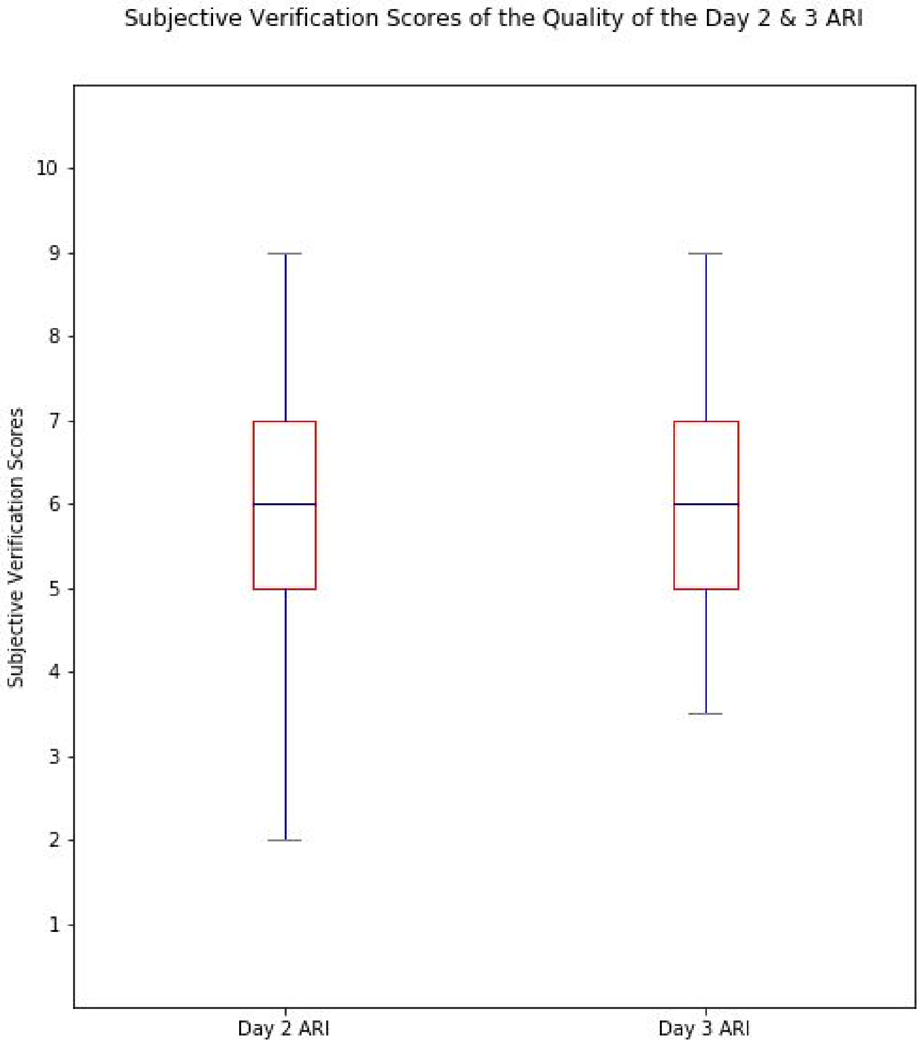


***Figure 42.** A box plot of all the subjective scores for the 24 hour Day 2 and Day 3 forecasts from the CSU-MLP First Guess Field over the course of the entire experiment.*

The probability scale for the product was matched to try and equal the probability scale of the ERO. When comparing to the experimental ERO issued by forecasters, the First Guess Field often had lower probabilities than the resulting experimental ERO created by the forecasters. One notable exception to this was in the Southwest region of the U.S., specifically over New Mexico. Considering both Day 2 and Day 3 forecast days, the First Guess Field produced the highest probabilities over a "bullseye" in New Mexico on two-thirds of the possible days. Figure 43 shows one such day where the probabilities were very high, in excess of 35%, in an area over New Mexico, but the forecasters went with lower probabilities and light QPE was observed. This figure also shows the positive aspects of the product previously mentioned. The overall spatial coverage is good and the higher probabilities in the Ohio River Valley and southeast Pennsylvania align well with reports. Overall, probabilities are lower than the ERO forecast, not exceeding 12% in those areas.

After using the product for a few days, the participants were able to adjust their thinking for some of the issues that appeared often and still found the guidance valuable as a starting point for the ERO forecast. Although not formally scored, participants found the 10 year 24 hour ARI exceedance probabilities to show no signal for marginal events and low signal for the stronger events. This made the 1 year 24 hour ARI product the more popular of the two to consult. Based on feedback from an end of the week survey question, participants overwhelmingly agreed that the CSU-MLP First Guess Field is an excellent first step in providing an initial starting point for WPC ERO forecasts, which has been a long requested tool. Many suggested further refinements and adjustments based on using this tool for the first time in a quasi-operational environment in order to further improve the tool.

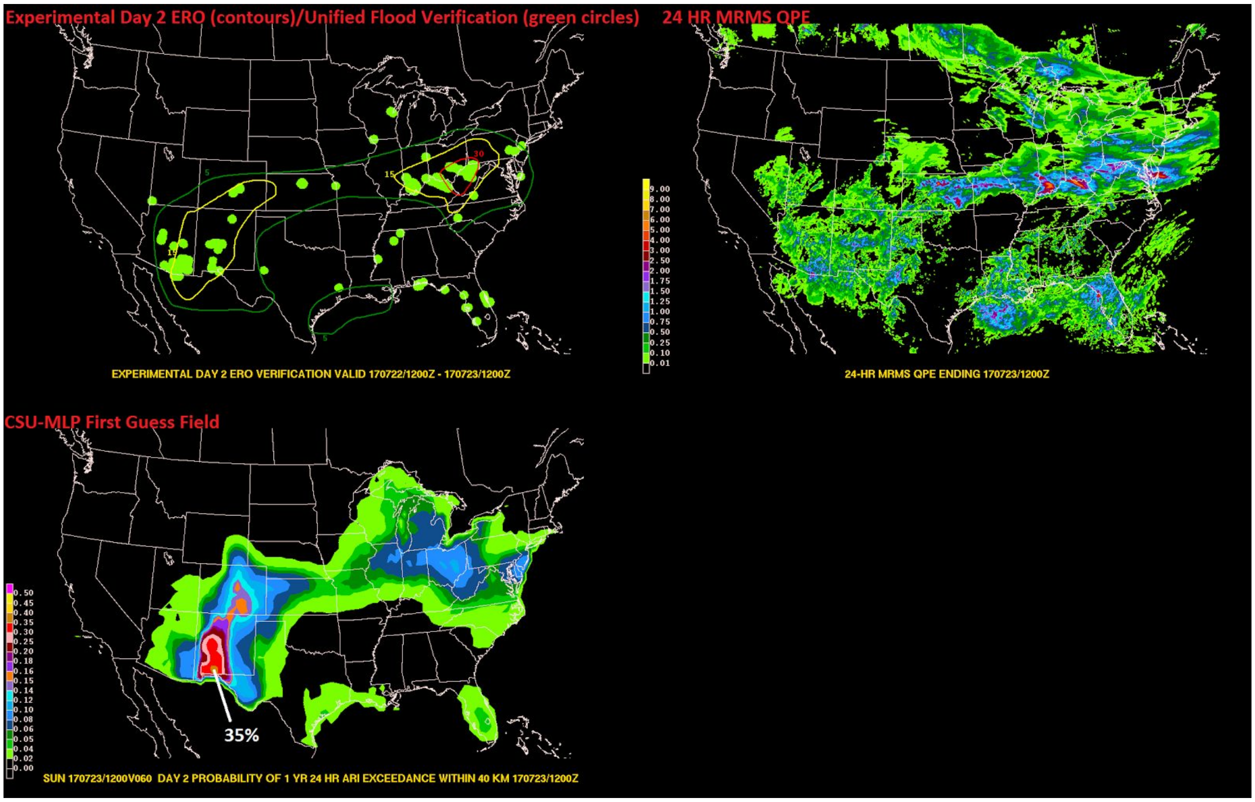

Experimental Day 2 ERO (contours)/Unified Flood Verification (green circles)
24 HR MRMS QPE
EXPERIMENTAL DAY 2 ERO VERIFICATION VALID 170722/1200Z - 170723/1200Z
24-HR MRMS QPE ENDING 170723/1200Z
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
CSU-MLP First Guess Field
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.18
0.16
0.15
0.14
0.12
0.10
0.08
0.06
0.05
0.04
0.02
0.00
35%
SUN 170723/1200V060 DAY 2 PROBABILITY OF 1 YR 24 HR ARI EXCEEDANCE WITHIN 40 KM 170723/1200Z

Figure 43. *Experimental Day 2 ERO (contours) overlaid on the UFV (green circles) (top left), 24 hour MRMS QPE (top right), the Day 2 (60 hour) CSU-MLP First Guess Field forecast of the probability of 1 year 24 hour ARI exceedance within 40 km of a point (bottom left) all valid from 12 UTC July 22 - 12 UTC July 23, 2017.*

6. Forecast Activities

<u>Experimental Probability of Flash Flooding Forecasts</u>

As part of the 2017 FFaIR Experiment, participants evaluated and subjectively scored the 6-hour experimental Probability of Flash Flooding (PFF) forecasts that were made each morning over a chosen limited domain of the CONUS, valid 18Z-00 UTC on Day 1. For verification, operational flash flood warnings, flood reports, and 6-hour MRMS QPE were used (please refer to the "Verification" section in section 3 for additional details). An example of a PFF forecast with verification is shown in Figure 44.

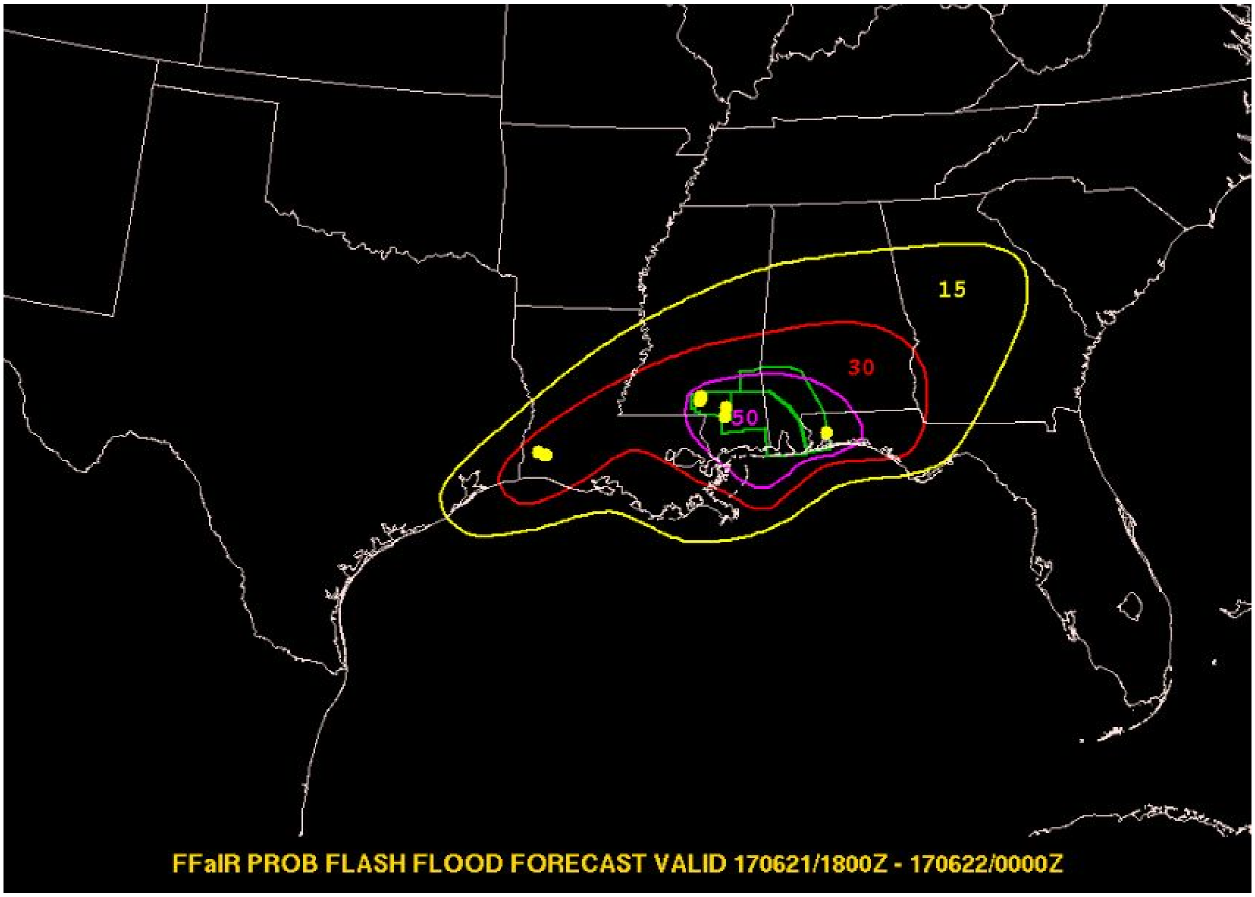

15
30
50
FFaIR PROB FLASH FLOOD FORECAST VALID 170621/1800Z - 170622/0000Z

Figure 44. *Experimental 6-hour PFF with probabilities of a 15% (slight/yellow), 30% (moderate/red), and 50% (high/magenta) probability of flash flooding from 18 UTC June 22 - 00 UTC June 23, 2017. The flash flood warnings are shown as green polygons and flash flood LSRs as yellow circles.*

Figure 45 shows a box plot of the all the subjective scores for the 6-hour experimental PFF which had an average subjective score of 6.84 out of 10 and a standard deviation of 1.90. The lowest score given by participants was a 2.5 (which was an outlier from the rest of the dataset as denoted by the red plus in the figure below) for the PFF, but it did receive an individual participant score of a perfect 10 on three separate occasions.

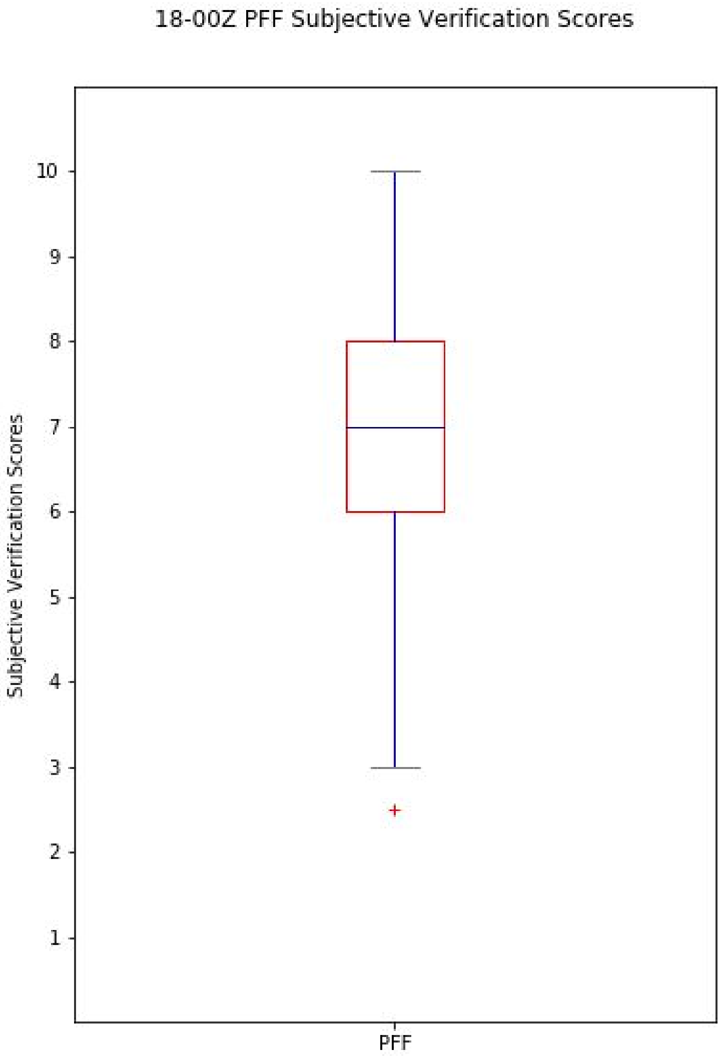


Figure 45. *A box plot of the overall subjective scores for the 6-hour PFF over the course of the 2017 FFaIR Experiment.*

The participants used experimental deterministic and ensemble CAM forecasts of QPF, reflectivity, and exceedance probabilities as well as hydrologic guidance from the National Water Model to produce the PFF. At times when experimental guidance was not plentiful, operational guidance was used. Of the 20 forecasts issued during the experiment, four of them were considered an over-forecast with contour values too high given the verification. Several forecasts did not have a large enough areal extent to capture all of the impacts. In two cases participants felt the coverage was too broad and should have been broken up into two separate contours. Otherwise the experimental PFF forecasts were generally regarded as well done.

Experimental Day 2 and Day 3 Excessive Rainfall Outlooks

Each day participants evaluated and subjectively scored the 24 hour experimental Day 2 and Day 3 EROs. For verification, the UFV system was used which takes into account FFG, ARI, and flood and flash flooding reports. Please refer to the "Verification" section in section 3 for additional details. Figure 46 shows an example of a Day 2 ERO overlaid with UFV reports. Twenty-four hour MRMS QPE was also available for reference in the evaluation. Participants could assign scores from 1 (very poor) to 10 (very good).

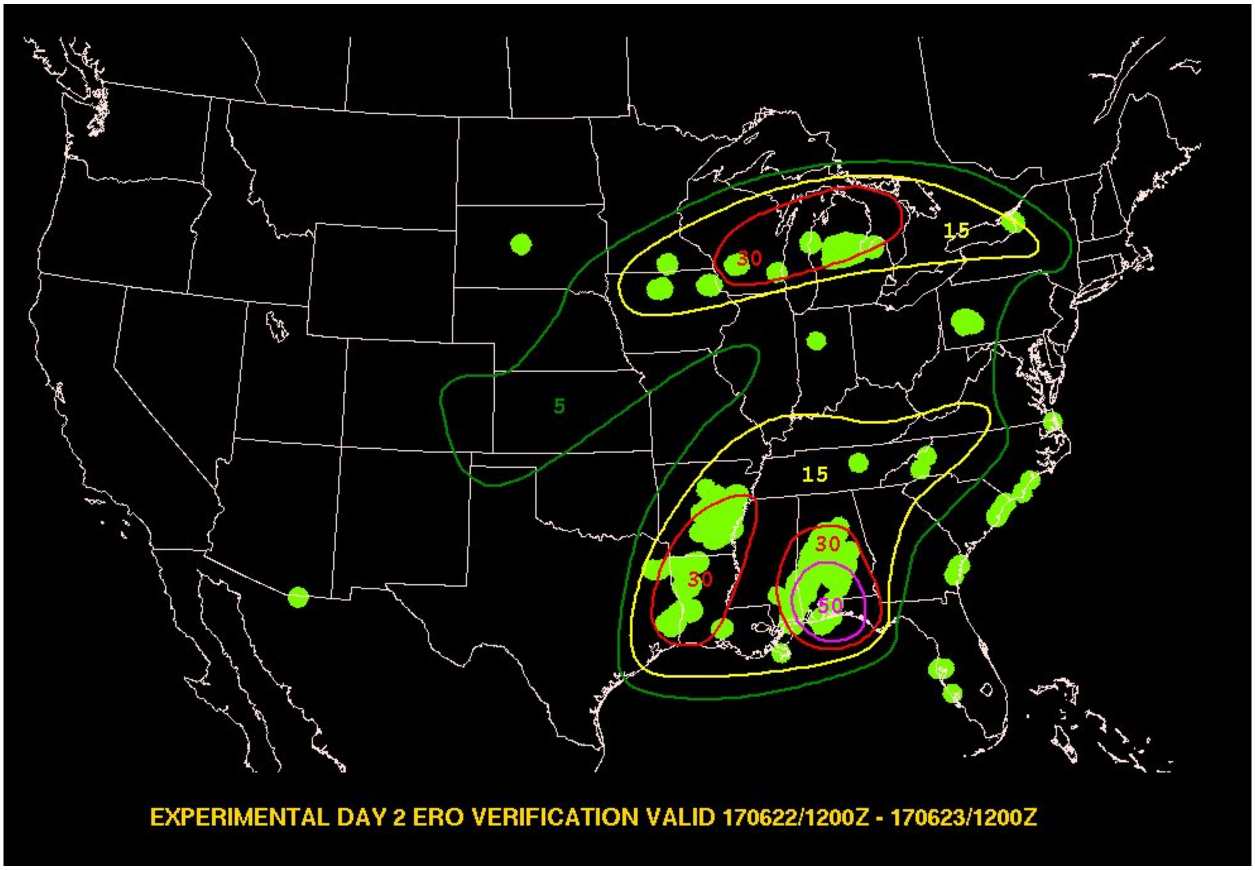

15
30
5
15
30
30
50
EXPERIMENTAL DAY 2 ERO VERIFICATION VALID 170622/1200Z - 170623/1200Z

Figure 46. *Experimental Day 2 ERO with probabilities of a 5% (marginal/green), 15% (yellow/slight), 30% (moderate/red), and 50% (high/magenta) chance of flooding rains occurring from 12 UTC June 22 - 12 UTC June 23, 2017. The UFV reports are shown by the green circles.*

Figure 47 shows a box plot of the all the subjective scores for the Day 2 and Day 3 experimental ERO. The Day 2/Day 3 EROs were subjectively scored the highest out of all forecasts and model guidance in the experiment. The Day 2 ERO had an average subjective score of 7.22 out of 10 and a standard deviation of 1.39. The Day 3 ERO had an average subjective score of 6.95 out of 10 and a standard deviation of 1.24.

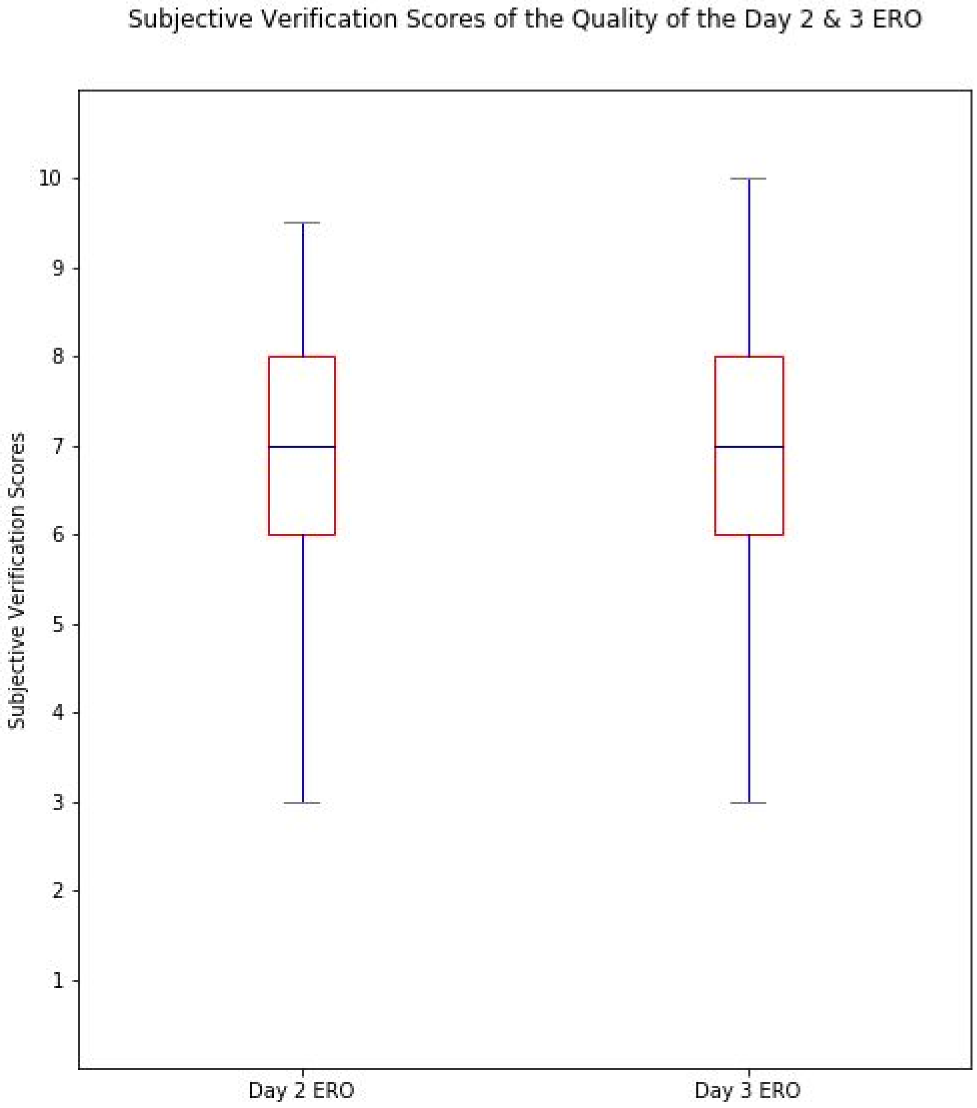
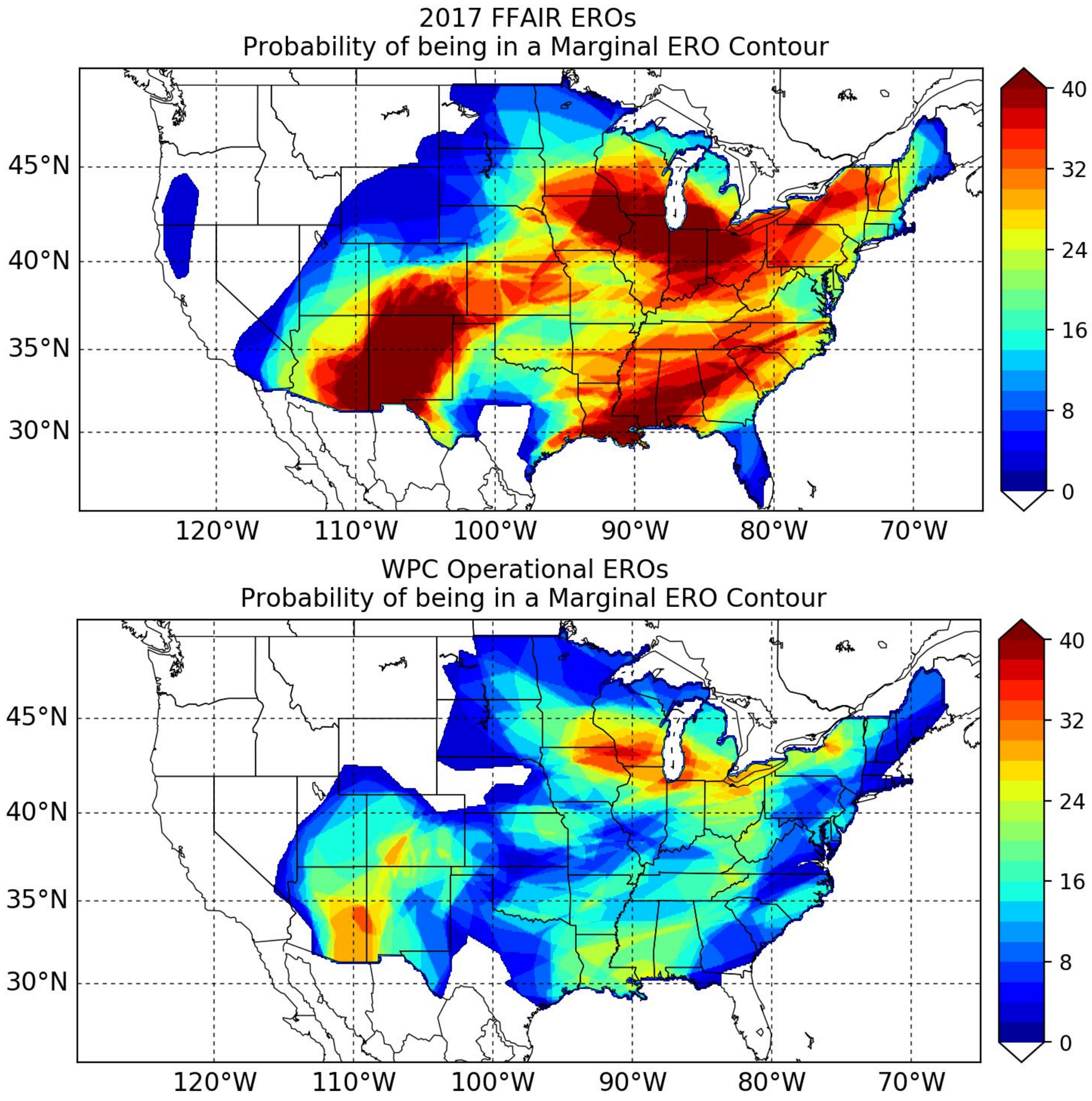


Figure 47. *A box plot of the overall subjective scores for the 24 hour experimental Day 2 and Day 3 ERO over the course of the entire experiment.*

As reflected by the average subjective scores, participants had overall positive feedback on the ERO forecasts. The most common feedback was that the shape and spatial coverage of the marginal (5%) probability contour usually captured well the areas where there were reports and rainfall. There were cases where smaller clusters of reports were missed and also times where, after seeing a large concentration of reports, the participants commented that the probability threshold should have been higher. On two days that had the lowest scores for Day 2, a four and a five out of 10, comments noted that in both cases the two separate marginal contours should have been connected due to a number of reports falling within the gap. For the day that scored a five out of 10, participants suggested a slight (15%) contour was needed in the Southwest whereas a slight contour in the Southeast was not necessary. The Day 3 ERO forecast that had the lowest average score, a four out of 10, failed to capture a large cluster of reports in Texas within any outlook, and the marginal areas that were drawn could have been expanded or connected. To summarize, participants felt that the experimental Day 2 and Day 3 ERO forecasts provided useful guidance by outlooking areas that had the potential for flooding rains. The most common issues with the forecasts were reports falling between separated marginal areas as well as probabilities that were too low based on the concentration of reports that occurred.

Comparison of the Day 2 and Day 3 Experimental FFaIR EROs and WPC Operational EROs

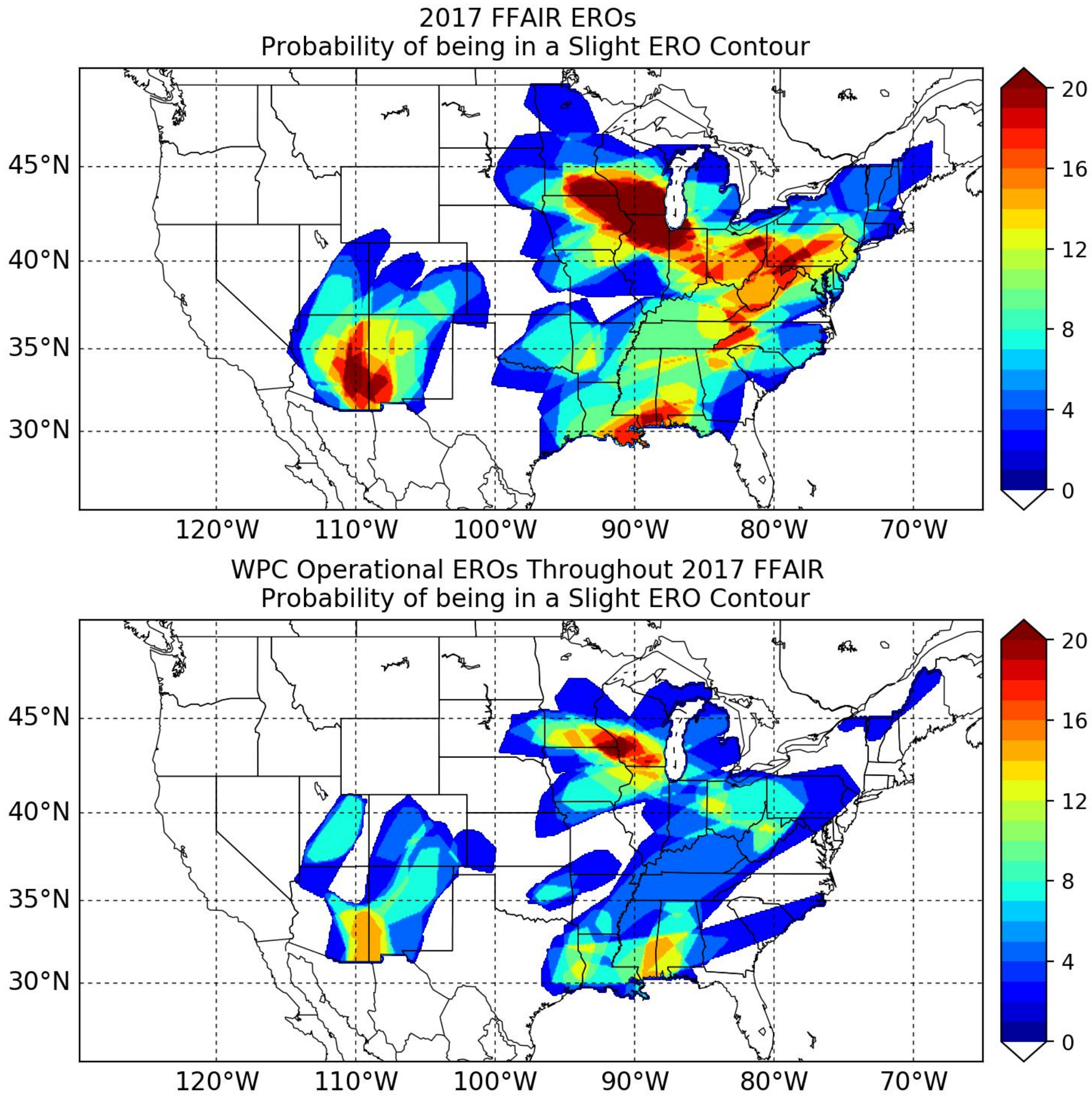
The Day 2 and Day 3 experimental FFaIR EROs issued over the four weeks of the experiment were compared to the operational WPC EROs issued over the same time period. The 09 UTC issuance of the Day 2 and Day 3 operational ERO was used for comparison. For verification, the UFV system was used with one important change: Stage IV QPE was used instead of radar-only-based MRMS QPE. The underlying probability contours for the FFaIR ERO differed from the operational ERO, so when assessing average fractional coverage, the FFaIR definition of marginal 5-10%, slight 15-30%, moderate 30-50%, and high 50%+ was applied to the operational ERO as well. Figures (48-51) show the probability of being in a “marginal risk” area, a “slight risk” area, a “moderate risk” area, and a “high risk” area from the operational/experimental EROs, respectively, over the four week experiment. From the figures it is immediately apparent that there are lower probabilities and lower overall coverage from the operational EROs at all probability thresholds. Zero Day 2 09 UTC high risk operational EROs were issued over the course of the experiment. Day 3 high risk EROs are currently not an option in the WPC operational product and thus, zero were issued as well. The area covered by the FFaIR ERO marginal, slight, and moderate contours was 207.7%, 206.9%, 394.9% greater than the operational ERO forecast, respectively. These maps also illustrate the major weather events during the experiment with higher probabilities in the Southeast from Tropical Storm Cindy, the Southwest from monsoonal moisture, and the upper Midwest/Great Lakes region from persistent MCS activity.



2017 FFAIR EROs
Probability of being in a Marginal ERO Contour

WPC Operational EROs
Probability of being in a Marginal ERO Contour

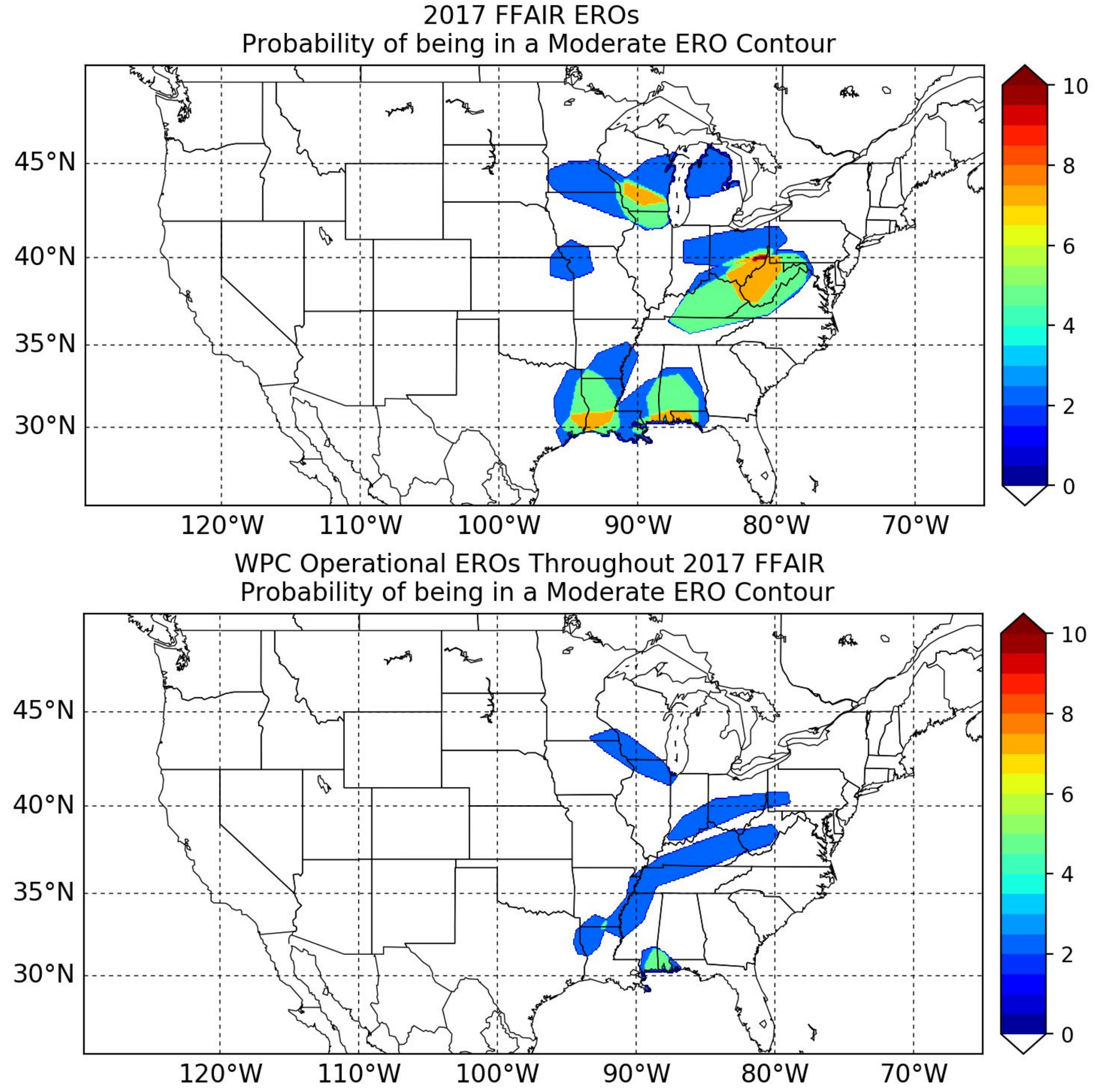
Figure 48. *Probability of being in a "marginal risk" experimental FFaIR ERO contour (top) and "marginal risk" operational ERO contour (bottom) over the four week experiment.*



2017 FFAIR EROs
Probability of being in a Slight ERO Contour

WPC Operational EROs Throughout 2017 FFAIR
Probability of being in a Slight ERO Contour

Figure 49. *Probability of being in a "slight risk" experimental FFaIR ERO contour (top) and "slight risk" operational ERO contour (bottom) over the four week experiment.*



2017 FFAIR EROs
Probability of being in a Moderate ERO Contour

WPC Operational EROs Throughout 2017 FFAIR
Probability of being in a Moderate ERO Contour

Figure 50. *Probability of being in a "moderate risk" experimental FFaIR ERO contour (top) and "moderate risk" operational ERO contour (bottom) over the four week experiment.*

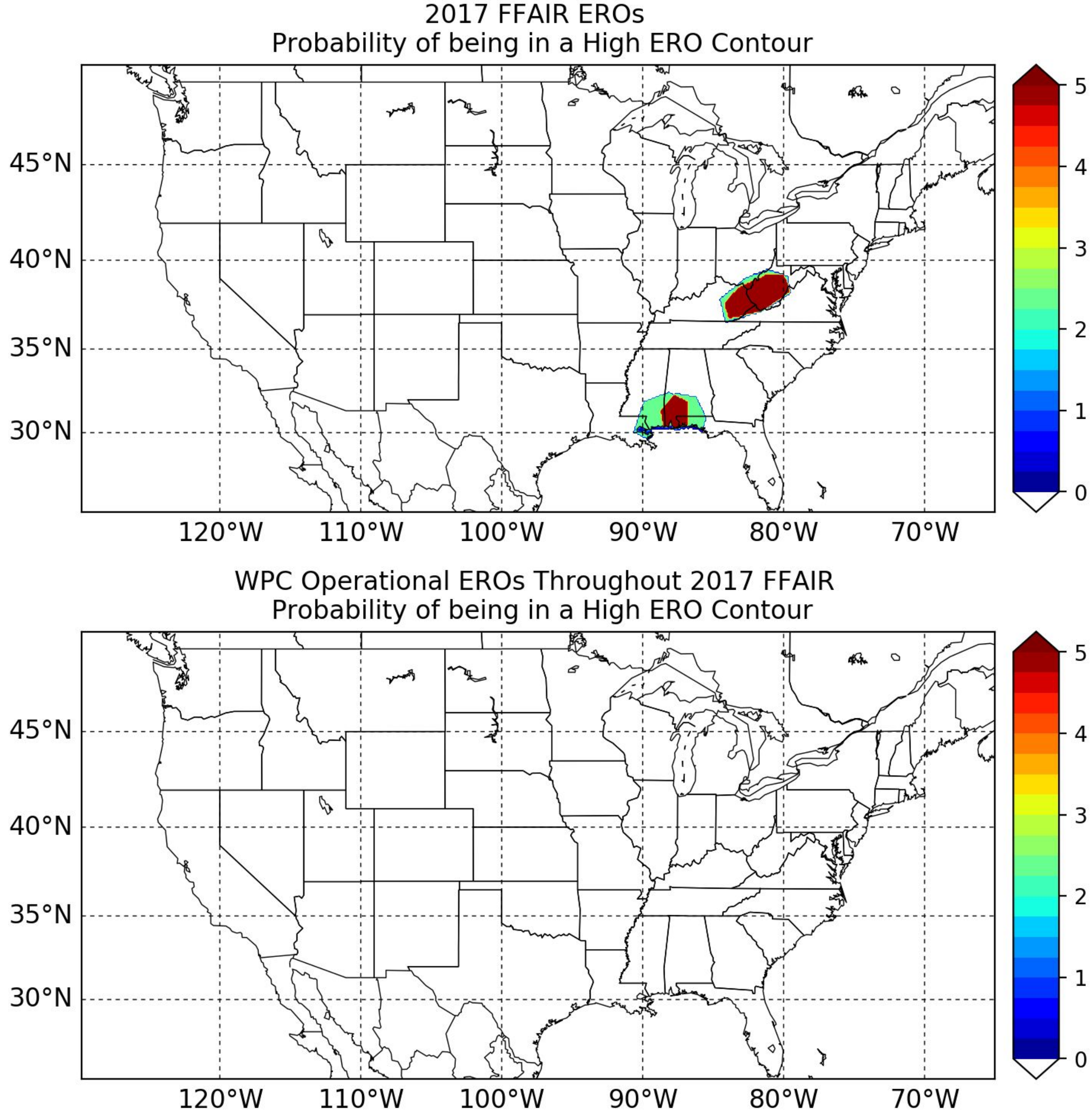

2017 FFAIR EROs
Probability of being in a High ERO Contour
45°N
40°N
35°N
30°N
120°W
110°W
100°W
90°W
80°W
70°W
5
4
3
2
1
0
WPC Operational EROs Throughout 2017 FFAIR
Probability of being in a High ERO Contour
45°N
40°N
35°N
30°N
120°W
110°W
100°W
90°W
80°W
70°W
5
4
3
2
1
0

Figure 51. *Probability of being in a "high risk" experimental FFaIR ERO contour (top) and "high risk" operational ERO contour (bottom) over the four week experiment.*

Figure 52 shows the bulk fractional coverage by threshold of the operational EROs and FFaIR EROs. The green and red horizontal lines represent the lower and higher bound of each probabilistic category. All thresholds, except for the operational moderate category, fell within the probabilistic definitions and can be considered calibrated. Due to the larger areal coverage of the FFaIR EROs, their average fractional coverage is lower when compared to the operational EROs. Figure 53 and Figure 54 show two examples of the WPC operational and FFaIR experimental Day 2 ERO valid 12Z June 22-23, 2017 and 12Z July 19-20, 2017, respectively. In both of these examples, the FFaIR ERO is more aggressive with the addition of moderate and high areas where the operational ERO had none or smaller moderate areas compared to the FFaIR. The verification reports are represented by the different symbols on each map.

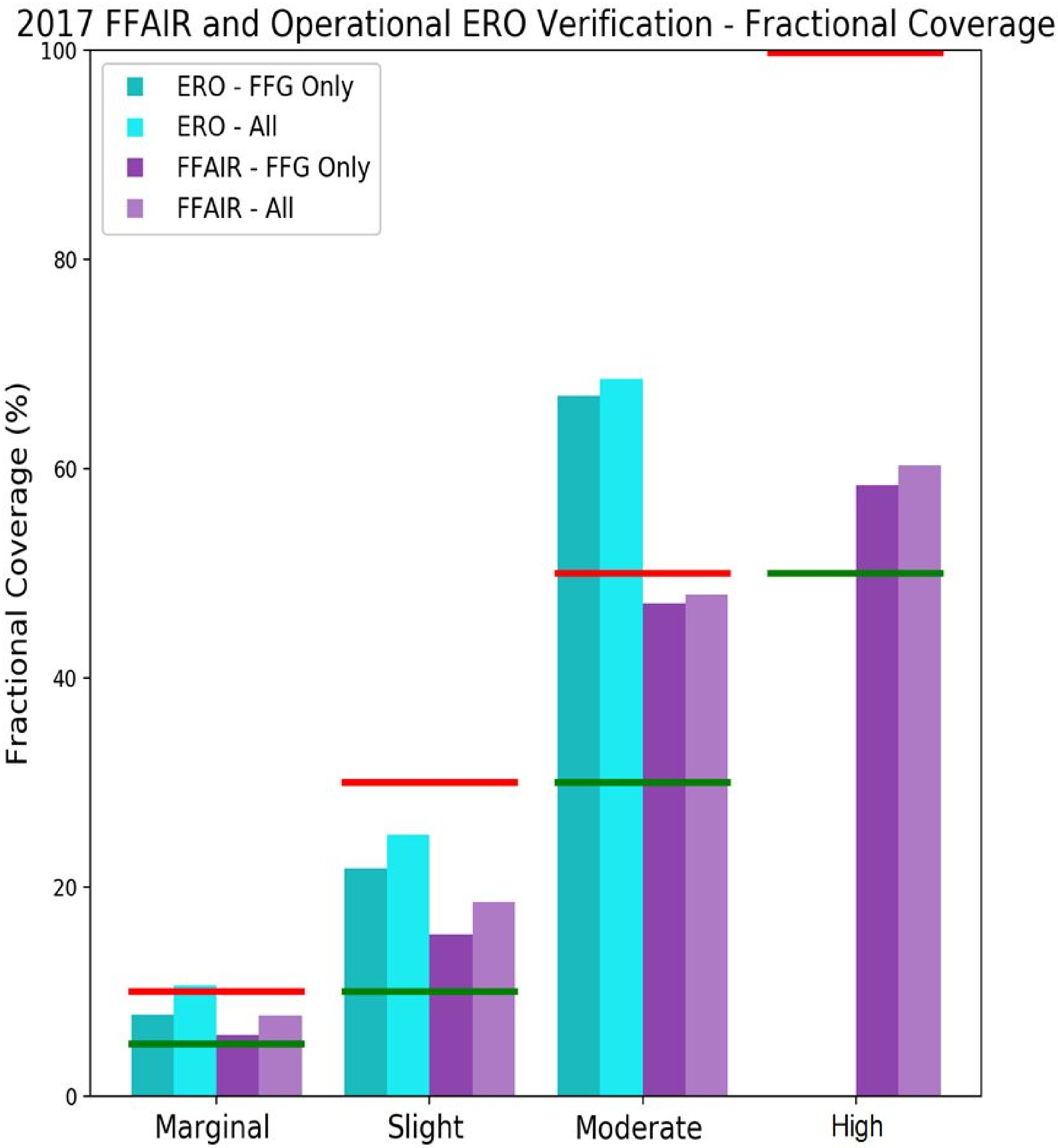


Figure 52. *Fractional coverage of the 2017 FFaIR EROs and operational EROs issued over the same time period for each probabilistic category. Green horizontal lines represent the lower defined bound for each threshold and red horizontal lines represent the highest defined bound.*

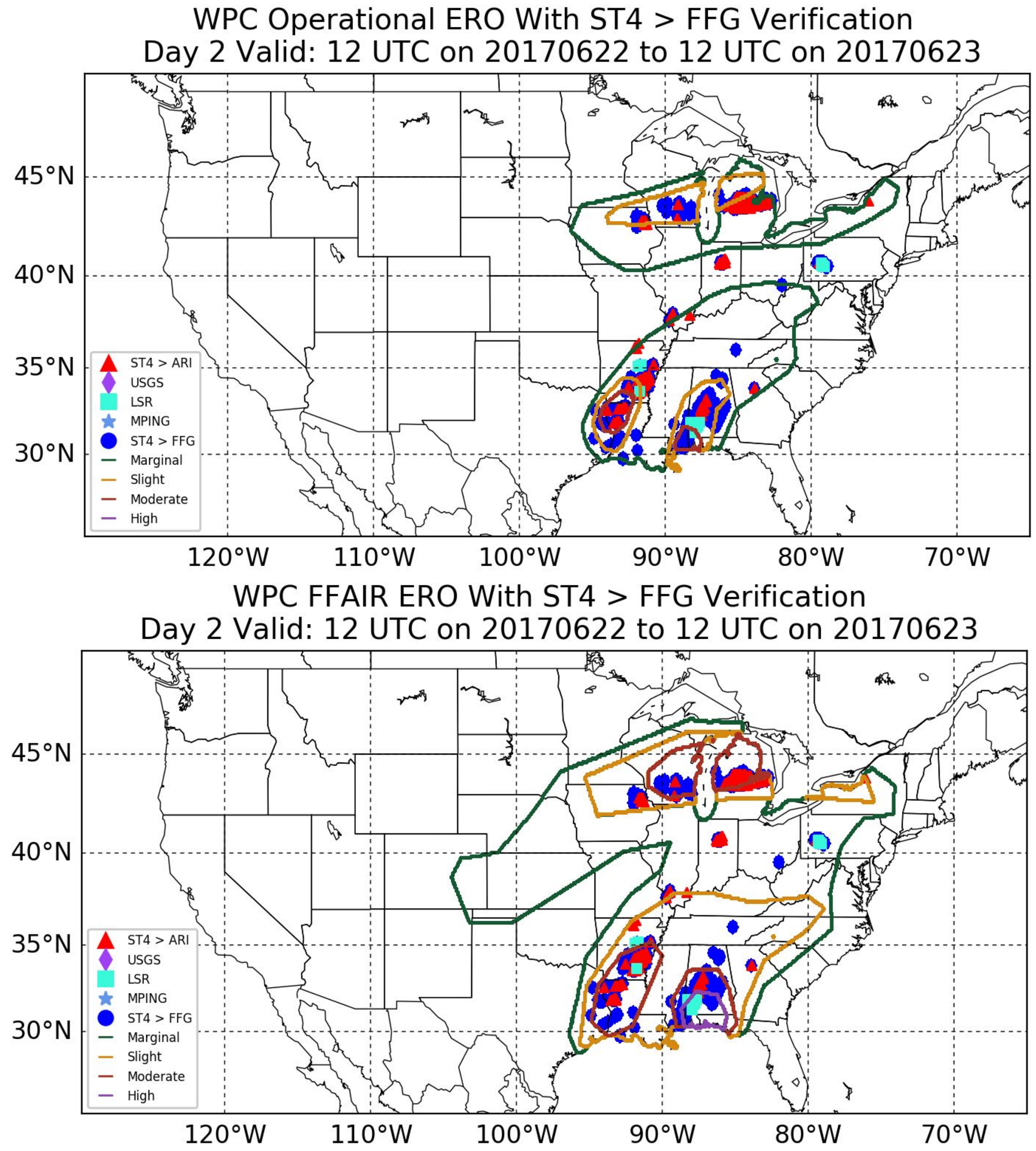


Figure 53. *Operational (top) and FFaIR experimental Day 2 ERO (bottom) valid 12 UTC June 22 to 12 UTC June 23, 2017 with verification.*

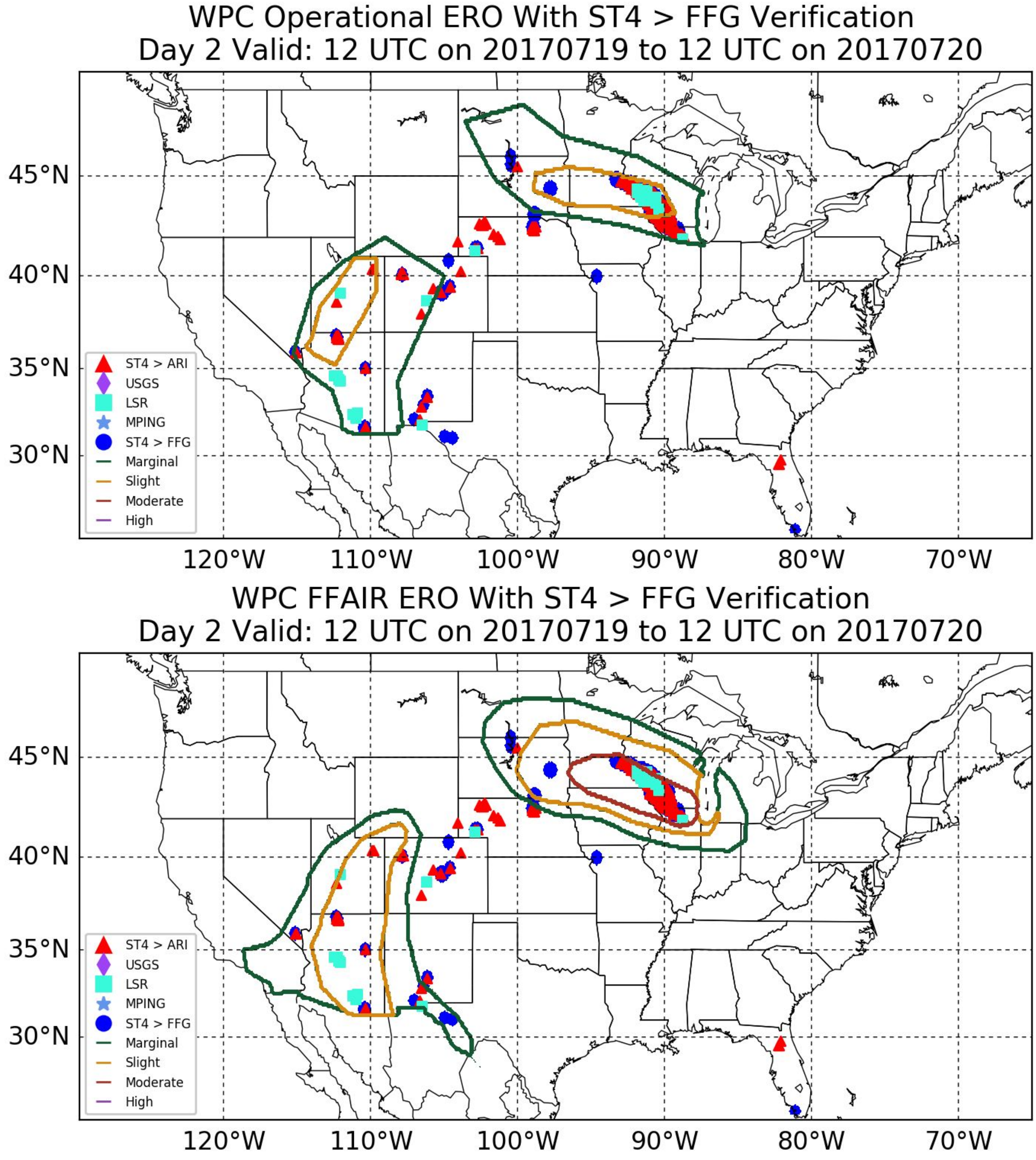


Figure 54. *Operational (top) and FFaIR experimental Day 2 ERO (bottom) valid 12 UTC July 19 to 12 UTC July 20, 2017 with verification.*

Figure 55A shows the daily Brier Skill Score (BSS) referenced against the operational EROs throughout the experiment and Figure 55B shows the bulk BSS referenced against the operational EROs for the Day 2 and Day 3 EROs. In both these figures, anywhere with positive values represents instances where the FFaIR ERO performed better than the operational ERO. There was considerable day to day variability, as is seen in Figure 55A, however in general, FFaIR EROs during portions of Week 1 and most of Week 2 performed better than the operational EROs. Weeks 3 and 4 the FFaIR EROs generally scored worse when compared to the operational EROs. In terms of the Day 2 and Day 3 EROs over the whole experiment (Figure 55B), the Day 3 ERO showed the most improvement in skill when compared to the operational ERO. Both FFaIR Day 2 and Day 3 EROs showed skill improvements over the operational EROs, except for the Day 2 ERO that was verified against FFG exceedance, which is the current operational definition.

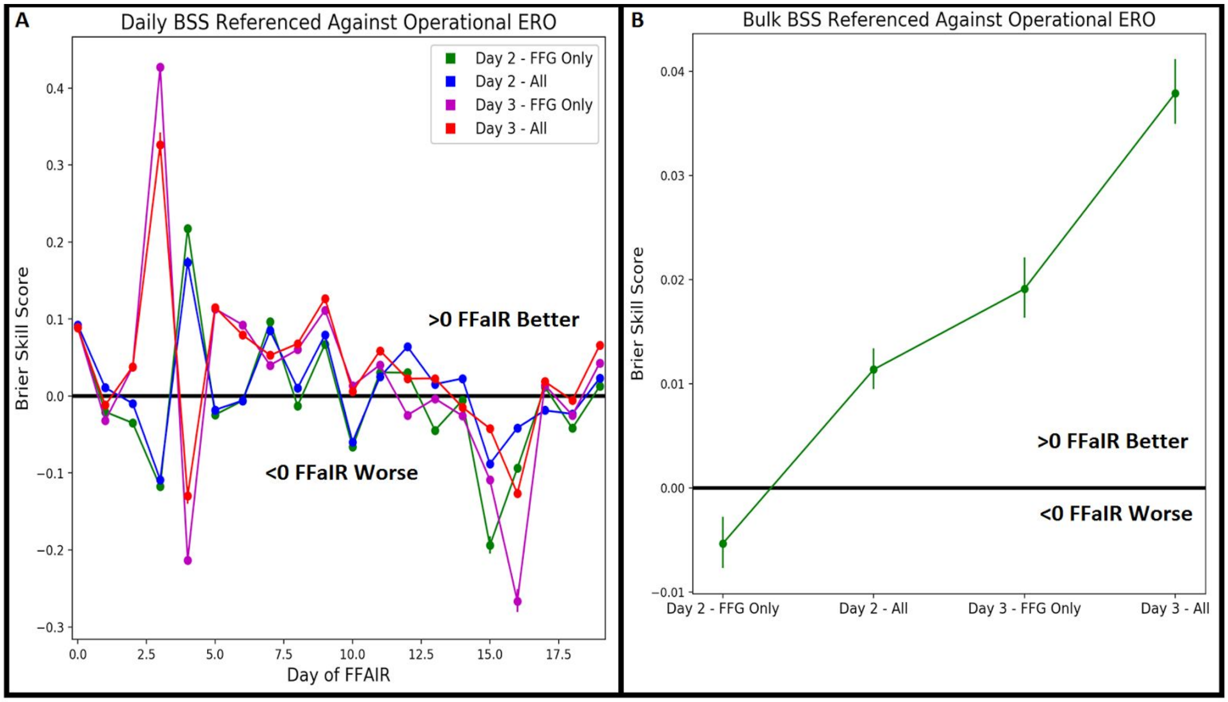


***Figure 55**. (A) Daily BSS referenced against operational EROs throughout the entire experiment. Positive values represent days the FFaIR ERO had better skill than the operational ERO, negative values represent worse skill. (B) Bulk BSS of the Day 2 and Day 3 FFaIR EROs referenced against the operational EROs over the whole experiment.*

Lastly, Figure 56 shows the area under the relative operating characteristic (AuROC) for both the FFaIR and operational Day 2 and Day 3 EROs. AuROC measures the hit rate against the false alarm rate, with higher values being better. According to this measure, the FFaIR EROs were consistently more skillful than the operational EROs during the experiment period. These results also show the Day 3 FFaIR ERO had the biggest skill improvement over the operational ERO. Despite the small sample size, all of these results combined show that despite the much larger areal coverage of the issued Day 2 and Day 3 FFaIR EROs, the objective statistics generally showed that the experimental EROs improved upon the skill of the operational EROs that were issued during the experiment.

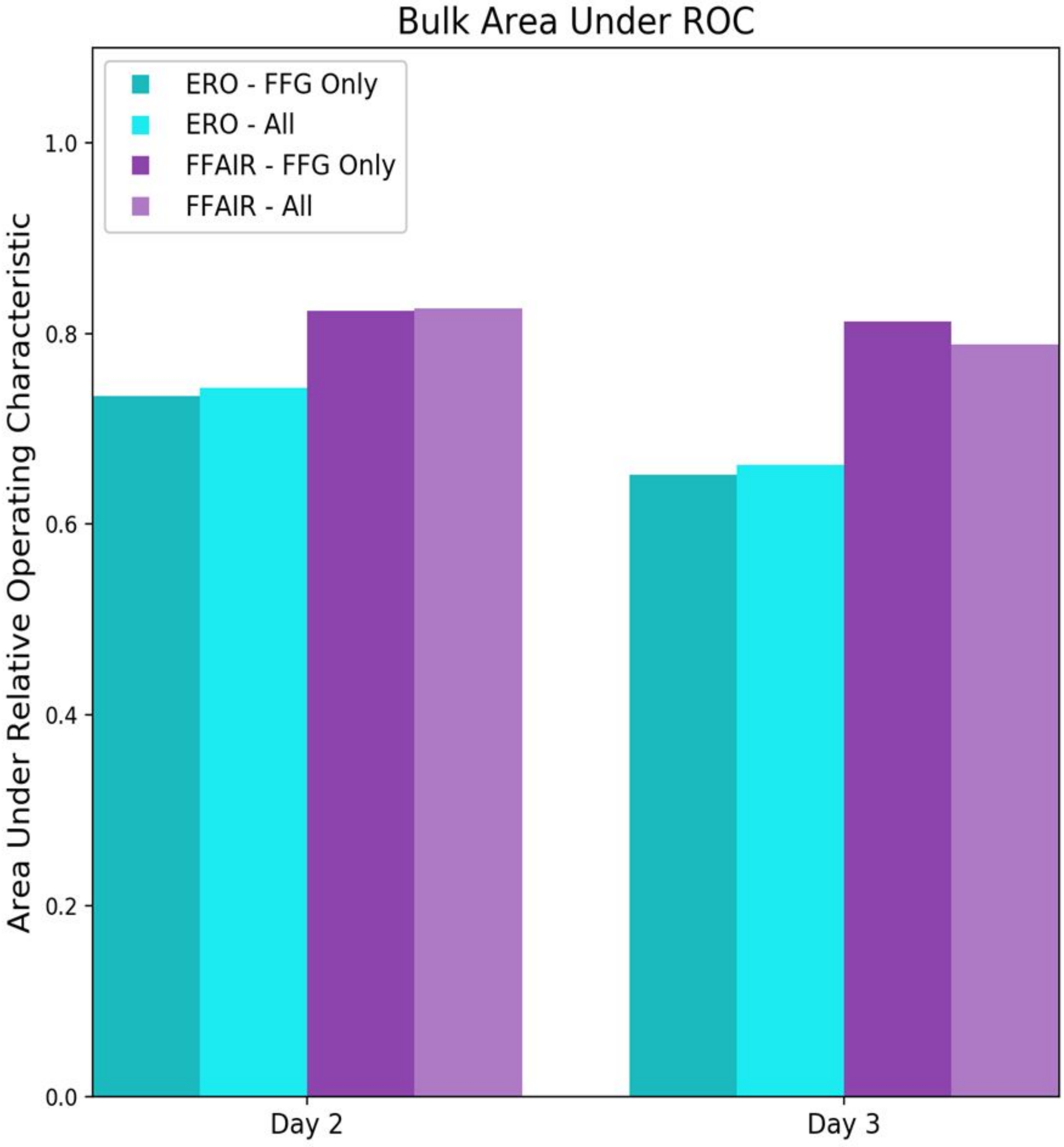


Figure 56. *Bulk area under the relative operating characteristic for both the Day 2 and Day 3 operational and FFaIR EROs.*

<u>Experimental Day 2 Flood Watches</u>

After the Day 2 ERO was completed by the participants, they were tasked with deliberating on Day 2 potential Flood Watches valid during 12Z-12 UTC, 12-, or 18-hour increments were an available option. The experimental watch region was defined by a contoured area in which operational flash flood warnings were expected to be issued during the specified valid time. On several of the days during the experiment, more than one experimental watch was issued on Day 2 with different valid times. All watches were subjectively evaluated and scored by the participants and again, the UFV system and MRMS QPE matching the valid time of the watch were used (please refer to the “Verification” section in section 3 for additional details).

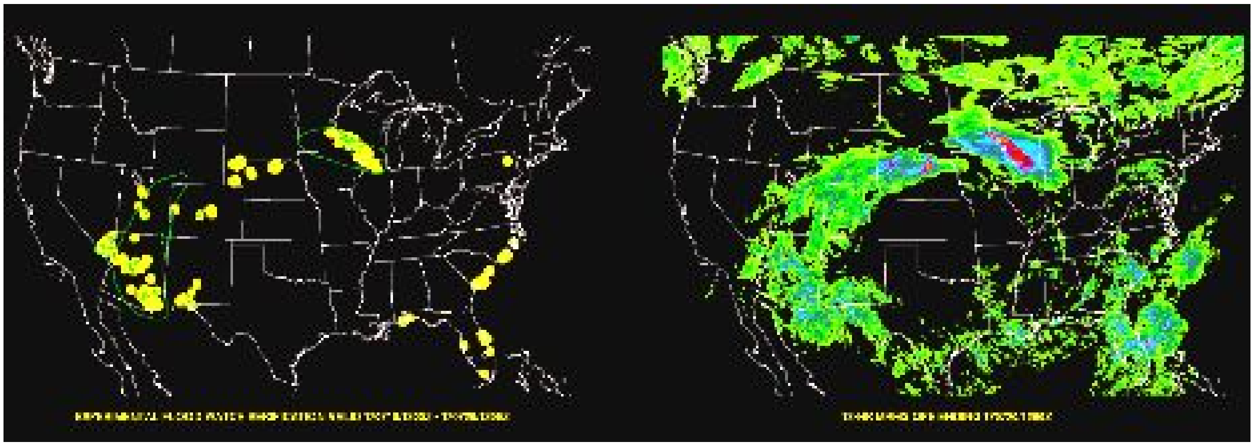


Figure 57. *An example of two experimental flash flood watch areas drawn in the FFaIR Experiment with flood reports (yellow circles) and operationally-issued flash flood warnings (green polygons) valid 18Z July 19 - 12Z July 20, 2017 and its corresponding NSSL QPE valid for that same 18-hour time period (right).*

The experimental flood watches received an average subjective score over the experiment of 6.2 out of 10 by participants with a standard deviation of 2.02. Figure 58 shows a box plot of the all the subjective scores for the watches. The lowest average score given during the experiment was a one-time 3 out of 10. The highest average score was a 9, achieved for three separate watch evaluations. Much of the subjective feedback from participants about issued experimental watches referenced “under-watching”, or contours being too limited in areal extent.

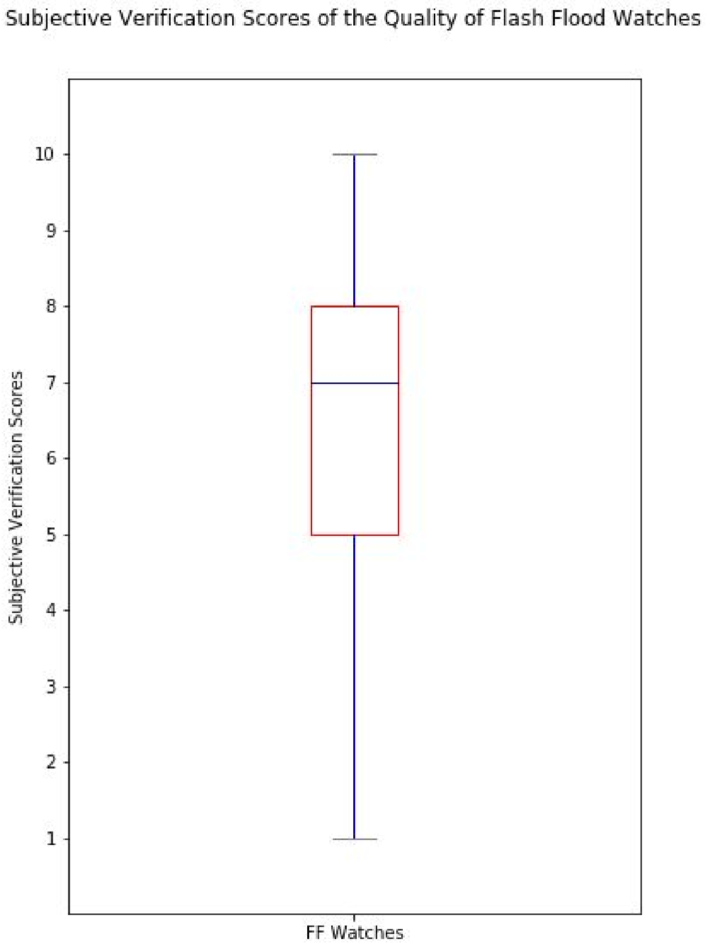


Figure 58. *A box plot of the overall subjective scores for the experimental flood watches over the course of the 2017 FFaIR Experiment.*

The various groups approached the creation of the CONUS-wide experimental watches differently. It was expected that participants would start their approach with the NWS field office threshold of at least 50% forecaster confidence that flash flooding will occur within 48 hours as justification for a watch. This standard should then have aligned with the 50% probability of flash flooding contours drawn in the corresponding experimental Day 2 ERO. However, we found that this was rarely the case when additionally factoring potential impacts into the decision making process (Figure 59). It was discovered that watch decisions are very complex, and tend to be more subjective than objective, driven by impacts and messaging. Conversely, the ERO tends to rely on more objective accuracy thus seemingly cannot directly inform a flash flood watch within its current paradigm.

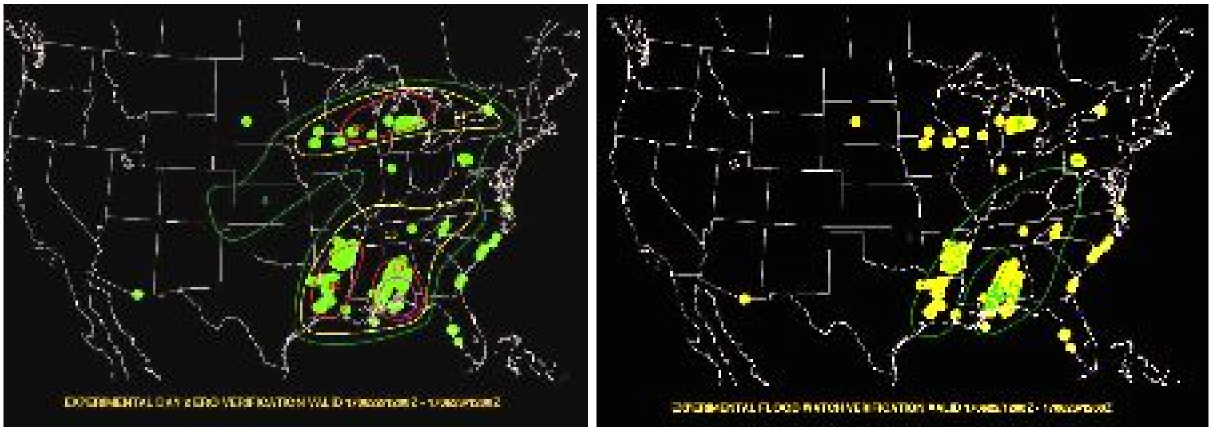

EXPERIMENTAL DAY 2 ERO VERIFICATION VALID
EXPERIMENTAL FLOOD WATCH VERIFICATION VALID

Figure 59*. An example of an ERO with 30% and 50% contours drawn along the Gulf Coast (left) with a corresponding large flash flood watch (right) that extended northward beyond those contours to include Kentucky and West Virginia based on impacts information such as the 1-year anniversary of the deadly 2016 West Virginia floods and population density.*

Creating flood watches on a CONUS scale sparked interesting discussions throughout the experiment. A common thread was the need for local knowledge of basins, infrastructure, terrain, soil type, and other flood-related impacts which could be achieved through deep collaboration between a national center and the field offices and the incorporation of a collaborative tool such as the National Water Model. The process of issuing a flash flood watch valid during the Day 2 period (as opposed to operational watches which are more often issued for Day 1) was fairly unconventional and posed new challenges which the experiment attempted to meet by offering longer-range high-resolution QPF and hydrologic guidance to increase confidence. Overall the participants resonated that issuing flood watches over the CONUS was certainly feasible with tighter definitions, local collaboration and proper staffing.

7. Other Forecasting Tools and Evaluations

CIRA Layered Precipitable Water

During the 2017 FFaIR Experiment, we asked participants to use and subjectively summarize the layered precipitable water (LPW) product. The LPW developed at the Cooperative Institute for Research in the Atmosphere (CIRA) uses microwave sensors on polar orbiting satellites and is mapped to a 16km grid to visualize the vertical distribution of water vapor in the atmosphere which may lead to heavy rainfall and flash flooding.

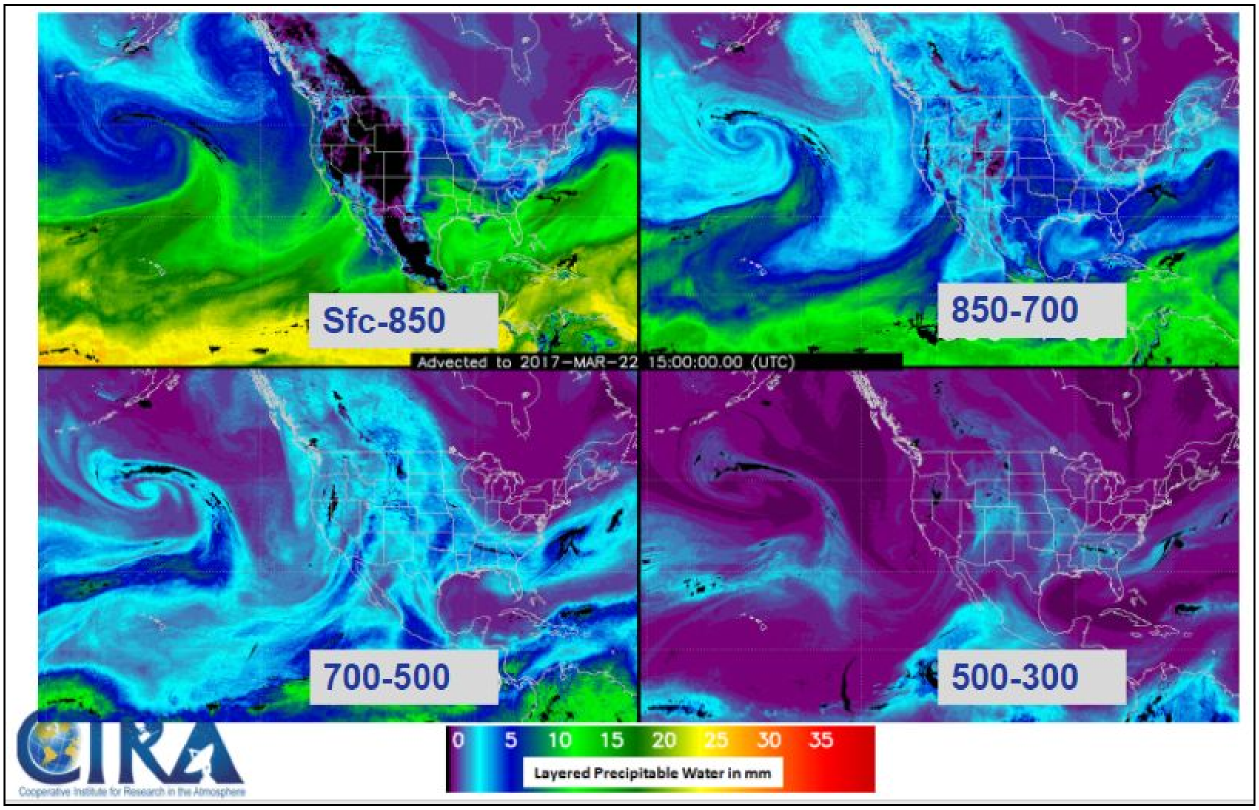

Sfc-850
850-700
700-500
500-300
Advected to 2017-MAR-22 15:00:00.00 (UTC)
CIRA
Cooperative Institute for Research in the Atmosphere
0 5 10 15 20 25 30 35
Layered Precipitable Water in mm

Figure 60. *An example of the CIRA LPW available on the web and to the NWS National Centers NAWIPS systems every 3 hours.*

The participants generally found the CIRA LPW to be very useful for situational awareness and very short-range flood forecasting as it is not a prognostic field. The availability and real-time nature of the product made it more difficult to apply to longer-range or flood outlook products. Analyzing the moisture content and transport at level intervals was particularly useful for areas of complex terrain, tropical systems, atmospheric rivers, and monsoonal fetches. For example, during the experiment, participants were able to identify moisture in the mid and upper levels from Tropical Storm Eugene in the Pacific interacting with the mountainous terrain in the southwest US. The correlation of moisture plumes aided the development of flood-producing monsoonal convection. Additionally, the CIRA LPW showed dry air being pulled into Tropical Storm Cindy’s circulation and led forecasters to trim flooding potential along the Gulf Coast.

Some participants felt the CIRA LPW did not offer any new information over soundings, model data or other precipitable water tools. As with other satellite-derived products, there was frustration with the areas of dense cloud cover shrouding the heavy rain and flood areas of concern. Although the visualization and data post-processing of the CIRA LPW was highly praised, the time constraints of longer-range flood forecasting often deemed this product not particularly vital to the forecast process.

8. Summary and Research-to-Operations Recommendation

The 5th annual Flash Flood and Intense Rainfall Experiment was conducted within the Hydrometeorology Testbed at WPC (HMT-WPC) from June 19-July 21, 2017 bringing together NWS meteorologists, hydrologists, and the development and research communities for the advancement of research into WPC and NWS field operations. The FFaIR Experiment focused on the use of high resolution guidance to improve flash flood prediction in both the short range (6-12 hours) and longer range (48-72 hours). The experiment also focused on engagement with the NWS Science and Operations Officer (SOO) community to enhance flash flood situational awareness and collaborate on potential experimental flood watches. Subjective and objective data was successfully collected and analyzed. The conclusions drawn are as follows:

- The **CSU Machine-Learning First Guess field** for the ERO showed great potential and was scored well by participants. It is recommended that the CSU developers work to reduce recurring biases and continue to refine the tool and reintroduce it into the testbed next year for further evaluation.
- The **ensemble QPF blended mean** was favorably scored by participants and was generally successful in increasing QPF magnitude of the composite mean and decreasing the magnitude of the probability matched mean. Although deeper objective evaluations of the components (composite and probability matched means) are recommended in future experiments to determine the best performance, there is justification for transitioning the QPF blended means for use in operations.
- The Time-Lagged HRRR **(HRRR-TLE) QPF on saturated soil** product performed as expected and was a useful tool for visualizing high probabilities of precipitation falling on soil that is already near or at saturation. Participants would like more opportunities to use this product in forecast operations to evaluate its utility in the flood forecast process. Due to its ease of transition into WPC operations and its availability on the ESRL/GSD (https://rapidrefresh.noaa.gov/hrrr/hrrrtle/), it is recommended that the QPF on saturated soil be available operationally to forecasters for further evaluation.
- The Experimental **Time to High Flow** product from the National Water Model (NWM) has high potential for use by forecasters in assessing flood risk. The information provided in this product is adequately detailed and can be used not just for prediction but in decision support messaging. It is recommended that this product be developed further to improve representation of time (for example, animation) and coupled with the capability to overlay the QPF forecast. The Time to High Flow should then be reintroduced to the testbed next year for further evaluation.
- The Experimental **Ponded Water** from the NWM is not recommended for operations at this time. Although the product does depict where the model QPF is accumulating water in grid cells that is not running off or absorbing into the soil, it does not provide information regarding flood risk. It has been suggested that the Ponded Water product be a step toward a more useful dataset such as inundation or a representation of accumulated precipitation that is reaching and/or exceeding bankfull.
- The Experimental **High Flow Probability** product from the NWM is a welcome first step into probabilistic hydrologic guidance. Though this product has utility for the forecasters, it is recommended for further development to ensure consistent messaging

among the NWM product suite. It is also recommended that greater ensemble membership be introduced to increase model spread and refine the probabilistic output. Probabilities from the NWM should continue to be tested in future experiments.

- Testing **Day 2 and Day 3 high resolution deterministic QPF** was a primary goal of the 2017 FFaIR Experiment, and results were mixed. Common weaknesses in convective allowing models, including difficulty capturing events in more weakly-forced synoptic regimes and a low bias in the QPF magnitude and spatial extent, propagate from Day 1 into Days 2 and 3. There is slight degradation in skill from Day 2 to Day 3. However, participants found value in the higher resolution models, especially reflectivity fields, and can apply known biases to the forecast process. It is recommended that developers continue work on the schemes that improve QPF and reflectivity fields in high-resolution models so they may further benefit flood forecasting at longer time ranges.
- It continues to be strongly recommended that **water models and precipitation models continue to experiment with fusion** (not only in model forcings, but also in visualization) to create products best suited for flood forecasters with issuance deadlines. It continues to be the desire of forecasters to be able to overlay the model QPF onto the hydrologic response data, both deterministically and probabilistically, with animating time steps for full situational awareness and improved prediction and messaging.
- The area covered by the FFaIR ERO marginal, slight, and moderate **contours was significantly larger than the Operational ERO** for every probabilistic threshold at Days 2 and 3, thus the FFaIR EROs had lower average fractional coverage of flooding proxies than the Operational EROs. FFaIR forecasters were **more aggressive issuing higher risk** probabilistic contours, indicating increased confidence with higher resolution guidance.
- Although considerable day to day variability is noted between FFaIR and Operational ERO performance, The **FFaIR ERO exhibits lower probabilistic error** when the Brier Skill Score is greater than zero. FFaIR ERO performed better than the Operational ERO first two weeks (strong synoptic events such as T.S. Cindy, frontal systems), but less so over second two weeks (marginal events, scattered convection, monsoons). **The FFaIR ERO consistently achieved higher Area Under the ROC values than operational EROs, with the greatest improvement on Day 3.**

Acknowledgements

The 2017 FFaIR Experiment was only possible through the hard work and dedication of the HMT and WPC staff. Box plots, performance diagrams and objective MODE verification were generated by **Joshua Kastman** (WPC/CIRES). Comparisons of experimental ERO and operational ERO provided by **Michael Erickson** (WPC/CIRES). Special thanks to WPC Science Operations Officer, **Mark Klein**, for technical and management support. Special thanks to Branch Chief, **James Nelson**. And many thanks to the operational WPC forecasters that lent their time and expertise to the experiment.

References

Herman, G., R. Schumacher (2016), Extreme Precipitation in Models: An Evaluation. *Wea. Forecasting*, **31**: 1853-1879.

Hershfield, D. M., 1961: Rainfall frequency atlas of the United States: For durations from 30 minutes to 24 hours and return periods from 1 to 100 years. U.S. *Weather Bureau Tech. Paper* **40**, 61 pp.

Miller, J., R. Frederick, and R. Tracey, 1973: *Precipitation-Frequency Atlas of the Western United States*. NOAA Atlas 2, Vol. 3, 43 pp.

Dey, S. R. A., Plant, R. S., Roberts, N. M. and Migliorini, S. (2016), Assessing spatial precipitation uncertainties in a convective-scale ensemble. Q.J.R. Meteorol. Soc., 142: 2935–2948. doi:10.1002/qj.2893

Ebert, E. E., 2001: Analysis of a Poor Man's Ensemble to Predict the Probability and Distribution of Precipitation. *Mon. Wea. Rev*., **129**, 2461-2480.

Harless, A. R., S. J. Weiss, R. S. Schneider, M. Xue, and F Kong, 2010: A Report and Feature-based Verification Study of the CAPS 2008 Storm-Scale Ensemble Forecasts for Severe Convective Weather, Preprints, 25th Conference on Severe Local Storms, Denver, CO, Amer. Meteor. Soc., 13B.2

Roebber PJ. 2009. Visualizing multiple measures of forecast quality. *Weather Forecasting* **24**: 601–608.

Silverman, B.W., 1986: *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 175 pp.

APPENDIX A

Participants

* Denotes participant was an observer
EMC-M indicates Environmental Modeling Center - Mesoscale Branch
EMC-G indicates Environmental Modeling Center - Global Branch
Note: Some weeks participants from EMC and WPC forecasters shared/split weeks

Week	WPC Forecaster	WFO/RFC/Other	Research/Academia	EMC
June 19 – 23	Patrick Burke	Leonard Vaughn (WFO CAE) Jeremy Michael (WFO ILK)	Diana Stovern* (WPC) Nate Snook (OU/CAPS) Trevor Alcott (ESRL-GSD) Pam Heinselman* (NSSL)	Ben Blake (EMC-M) Corey Guastini (EMC-M)
June 26 – 30	Marc Chenard	Jonathon Thornburg (NCRFC) Keith Fenwick (Met Office, UK)	Chandra Kondragunta* (OWAQ) Ryan Sobash* (NCAR) Katie McGee (MDL) Jeff Craven* (MDL) John Forsythe (CIRA) Kelly Mahoney (ESRL-PSD) Russ Schumacher (CSU)	Jacob Carley (EMC-M) Matthew Pyle (EMC-M)
July 10 – 14	Rich Bann	Neal Strauss (NERFC)	Phil Schafer* (MDL) Kate Abshire (OWP) Jill Hardy (WDTD) Terra Ladwig (ESRL-GSD) Greg Herman (CSU) Brian Tang (SUNY-ALB)	Tracey Dorian (EMC-G)
July 18 – 22	Robert Oravec	Melissa Huffman (WFO HGX) Tim Axford (WFO PBZ)	Peter Stone* Cammey Simms (MDL) Curtis Alexander (ESRL-GSD) Steve Martinaitis (NSSL) Fernando Salas* (OWP/NWC) Keith Brewster (OU/CAPS)	Mallory Row (EMC-G)

APPENDIX B

Operational and Experimental Deterministic Guidance

RFC Flash Flood Guidance

Flash Flood Guidance (FFG) is produced by each individual NWS River Forecast Center (RFC) in accordance with each RFC domain (Fig. 61). There are four methods currently employed to create FFG: Lumped Flash Flood Guidance (LFFG), Gridded Flash Flood Guidance (GFFG), Distributed Flash Flood Guidance (DFFG), and the Flash Flood Potential Index (FFPI). Therefore, the method of producing FFG is inconsistent across RFCs. WPC compiles the guidance from each RFC to create a CONUS 5-km resolution mosaic FFG grid. The CONUS mosaics are time-stamped every 6 hours (00, 06, 12, 18 UTC), but are updated hourly to account for the latest guidance issued by RFCs.

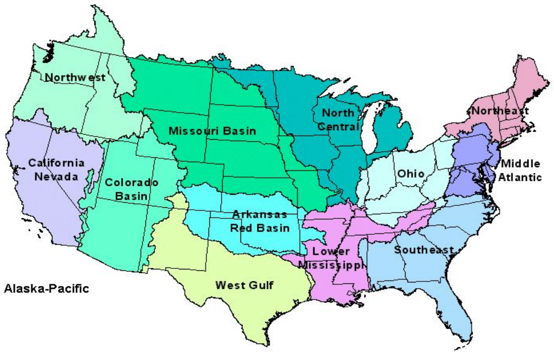


Figure 61 *Showing domain for each NWS River Forecast Center (NOAA/NWS (water.weather.gov)*

Precipitation Average Recurrence Intervals

Precipitation Average Recurrence Intervals (ARIs) are frequency estimates generated mainly from NOAA Atlas-14 Climatology of USGS rain gages. Statistical analyses are applied to the precipitation climatology to generate precipitation amounts representing the approximate frequency of occurrence (e.g. 1 year, 5 years, 100 years, etc.) for various accumulation periods (e.g. 5 minutes, 30 minutes, 3 hours, 24 hours, etc.). RIs can help to identify how rare a rainfall event is for a given area, alerting forecasters to abnormal or potentially extreme rainfall events. Standard ARIs are available for intervals of 2, 5, 10, 25, 100, 500 and 1000 years, and are measured in inches, and do not account for antecedent conditions. An example of the 6 hour, 100 year recurrence interval is shown in Figure 62.

For 2017 FFaIR ensemble model developers and participants, we will be providing fully-stitched grids (Herman and Schumacher, 2016). Thresholds come from NOAA Atlas 14 for most of CONUS. This includes the New England area which received an Atlas 14 update in autumn 2015, the most recent update to Atlas 14. Two regions of CONUS have not yet received NOAA Atlas 14 updates: Texas and the northwest, which is comprised of Washington, Oregon, Idaho, Montana, and Wyoming. For Texas, thresholds from Technical Paper 40 (TP-40; Hershfield 1961) were used; digital grids of the selected ARIs between 1 and 100 years were included for 6- and 24-hour precipitation accumulations. and so no additional processing was required. For the northwest, TP-40 did not provide coverage; instead, NOAA Atlas 2 (Miller et al. 1973) was used for these thresholds. The only grids that had been digitized were 2- and 100-year ARIs for 6- and 24-hour accumulations. However, the two frequency thresholds at each point for these five states combined with the knowledge that these threshold estimates were originally derived from a (two-parameter) Gumbel distribution, a Gumbel distribution could then be fitted to each point (two equations, two unknowns), and estimates for the 1-, 5-, 10-, 25-, and 50-year ARIs derived from those Gumbel fits. These different threshold estimate sources were then stitched together to form CONUS-wide grids.

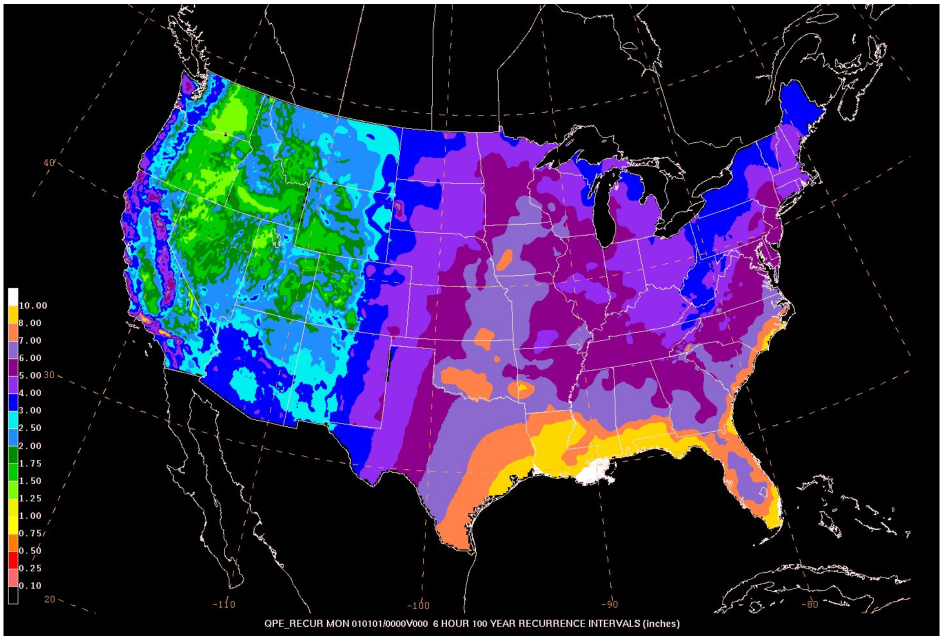


Figure 62. *An example of a full Average Recurrence Interval map (100 year ARI over 6 hours) available to forecasters both operationally at WPC and in the FFaIR Experiment.*

<u>Flooded Locations and Simulated Hydrographs (FLASH)</u>

The *Flooded Locations And Simulated Hydrographs Project* (FLASH) was launched in early 2012 largely in response to the demonstration and real-time availability of high-resolution, accurate rainfall observations from the MRMS/Q3 project. FLASH introduces a new paradigm in flash flood prediction that uses the MRMS forcing and produces flash flood forecasts at 1-km/5-min resolution through direct, forward simulation. The FLASH team is comprised of researchers and students who use an interdisciplinary and collaborative approach to achieve the goal.

In the first implementation, FLASH relies on rainfall forcing from MRMS/Q3 to the CREST model run with *a priori*, physically based parameters. The CREST hydrologic model is used to produce surface water fluxes at 1 km/5 min resolution. These discharges are converted to return periods using a long-term hindcast simulation with forcing from the gridded NEXRAD rainfall archive. All hydrologic forecast products are produced at the same 1-km/5-min resolution as the rainfall forcing. The availability of long-term, gridded rainfall archives provides for retrospective hydrologic simulation. At each grid point, the time series of historic simulations are used to evaluate the rarity or severity of model forecasts.

National Water Model (NWM) Experimental Post-Processed Products

The National Water Model (NWM) is a hydrologic model that simulates real-time and forecasted streamflow, as well as other key water budget variables, over the entire continental United States (CONUS). Operating on the NHDPlus river network, the NWM simulates streamflow for ~2.7 million river segments by coupling land-surface processes, defined on a 1 km grid, with terrain routing processes defined on a 250 m grid. An example of the streamflow analysis is shown in Figure 63.

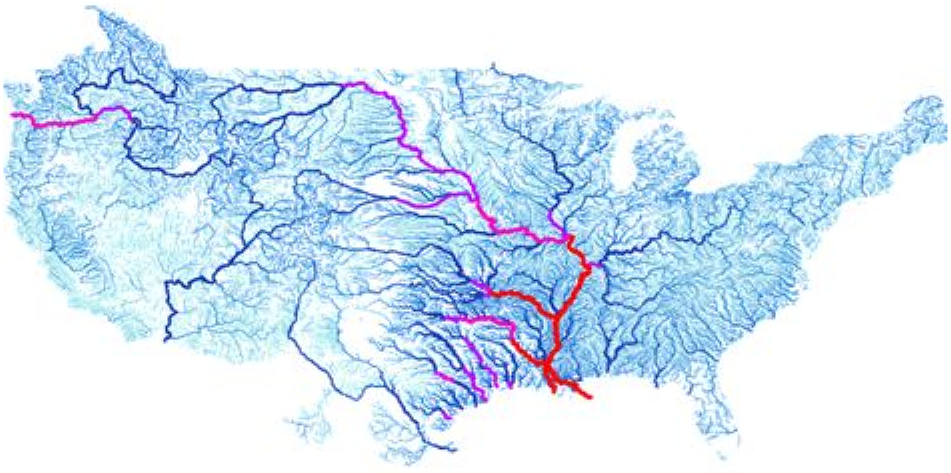


Figure 63. *National Water Model Streamflow Analysis.*

The core of the system is the NCAR-supported community WRF-Hydro hydrologic model. It ingests forcing from a variety of sources including MRMS radar-gauge observed precipitation data and HRRR, RAP, GFS and CFS NWP forecast data. WRF-Hydro is configured to use the Noah-MP Land Surface Model (LSM) to simulate land surface processes, separate terrain routing modules to perform diffusive wave surface routing and saturated subsurface flow routing on a 250m grid, and muskingum-cunge channel routing down NHDPlus stream segments. River analyses and forecasts are provided across a domain encompassing the CONUS and hydrologically contributing areas, while land surface output is available on a larger domain that extends beyond the CONUS into Canada and Mexico (roughly from latitude 19N to 58N). The system includes an analysis and assimilation configuration along with three forecast configurations. USGS streamflow observations are assimilated into the analysis and assimilation configuration, and all four configurations benefit from the inclusion of 1,260 reservoirs.

The NWM is run in four configurations:

1. Analysis and assimilation (updated every hour)
2. Short-Range (18 hour deterministic forecast updated every hour)
3. Medium-Range (10 day deterministic forecast updated every six hours)
4. Long-Range (30 day multi-member ensemble forecast update every six hours)

In order to quickly synthesize the spatial and temporal patterns within the NWM output, the NWC plans to post-process and publish a suite of dynamic map services that allow users to zoom in and and pan around areas of interest. To test experimentally during FFaIR, these NWM products may include:

- **High flow potential** (updated each hour): *This map will identify reaches that are at or above a defined high flow threshold** using the analysis/assimilation configuration of the NWM. This map will communicate near real-time conditions of high flow.*
- **Time to high flow/return to normal** (updated each hour): *This map will use the 18 hour short range forecast from the NWM to communicate the time (or time until) each river segment exceeds a high flow threshold** (e.g. 1 hour ahead, 2 hours ahead, 3 hours ahead, 4-6 hours ahead, 7-9 hours ahead etc.) In addition, this map will communicate the time at which the high flow is predicted to return to normal.*
- **Time to high flow** (updated every 6 hours): *This map will use the 10 day medium range forecast from the NWM to communicate the time (or time until) each river segment exceeds a high flow threshold**. This map will be similar to the above but would be symbolized using a coarser time scale (i.e. 18-24 hours ahead, 24-30 hours ahead, 30-36 hours ahead etc.). This map will only extend out 3 days.*
- **Time-lagged probabilistic forecast** (updated each hour): *This map will communicate the probability of exceeding a high flow threshold** 6 hours ahead, for each river segment. Probability will be based on time-lagged ensembles using the 18 hour short range forecast from the NWM. Depending on the zoom level, probabilities will be aggregated in space to improve visualization.*
- **Ponded water forecast** (updated each hour): *This map will depict areas where water is potentially ponding within the 250 m NWM grid. This map is to be used as guidance and not interpreted as actual depth of ponded water.*
- **Time enabled NWM QPF** (update each hour): *This map will display depth of precipitation as a function of lead using the 18 hour short range forecast. The precipitation grid presented here represents the forcing used to drive the NWM.*
- **Accumulated QPF** (updated each hour) *This map will display the cumulative precipitation depth as a function of lead time (as described above) over the short range forecast time horizon (0-18 hours).*

EMC 3km North American Model CONUS Nest (NAM)

Version four of the North American Mesoscale Forecast System (NAM) features many changes to its model and data assimilation components that serve to improve forecast quality from its convection-allowing, nested domains. A subset of these changes include: (a) updates to the Ferrier-Aligo microphysics scheme to reduce noted high-precipitation biases and improve stratiform precipitation, (b) improved consistency between model dynamics and physics through calling the physics routines more frequently, and (c) advecting humidity every dynamics timestep. In addition, the CONUS, Alaska, Hawaii, and Puerto Rico nest grid-spacings have all been unified at 3 km while the requestable, on-demand Fire Weather domain will use a grid-spacing of 1.5 km.

Prior versions of the NAM forecast system have featured a three hourly forecast-analysis assimilation cycle done just on the 12 km parent domain. The NAM package has been updated with an hourly forecast-analysis assimilation cycling period, which includes radar and lightning observations, prior to initializing the free forecasts at the traditional times of 00, 06, 12, and 18Z. Furthermore, the NAM features a distinct data assimilation cycle for its 3 km CONUS and Alaska domains in addition to the traditional 12 km North American domain.

The NAM system changes targeted the greatest deficiencies in the previous NAM system, such as excessive precipitation from the high-resolution CONUS nest and risks of failure exposed by Hurricane Joaquin. The development and testing of NAM targeted these deficiencies through 1) upgrading the forecast model, 2) improvements in the data assimilation techniques applied for efficient use of all observations (especially NEXRAD data) and 3) assimilation of new observation types. Specifically, the changes include increasing the horizontal resolution of the convective-allowing CONUS nest from 4 to 3 km, steps to ensure unstable thermodynamic structures do not develop, updated microphysics to improve storm structure and radar reflectivity signatures, land-surface model changes to improve near-surface fields and visibility, and building a new NAM data assimilation system for the CONUS and Alaska nests which allowed for effective use of NEXRAD radar and lightning data.

The NAM changes led to greatly improved warm-season precipitation forecasts from the 3 km model simulations, improved predictions of low visibility conditions in coastal regions, improved cool-season precipitation forecasts from the 12 km simulation, improved short-term (0-12 h) simulation of convective storms in the 3 km simulation, and a reduction of the moist bias for CONUS cool-season surface dew point temperatures and a reduction of the warm bias for CONUS summer surface temperatures.

The NAM CONUS Nest provides runs hourly out to 60 hours at 00Z, 06Z, 12Z, and 18Z cycles.

United Kingdom Met Office Experimental Unified Model (UM)

The Met Office Unified Model (UM) is the name given to the suite of numerical modelling software used by the Met Office. A fully operational, nested limited-area high resolution version of the UM at (2.2km horizontal resolution) running twice per day will be supplied to the 2017 FFaIR Experiment. These operational, nested, hi-res versions will incorporate the latest UM settings that are used over the UK. The 2.2-km model has 70 vertical levels (spaced between 5m and 40 km) across a slightly sub-CONUS domain (Figure 64).

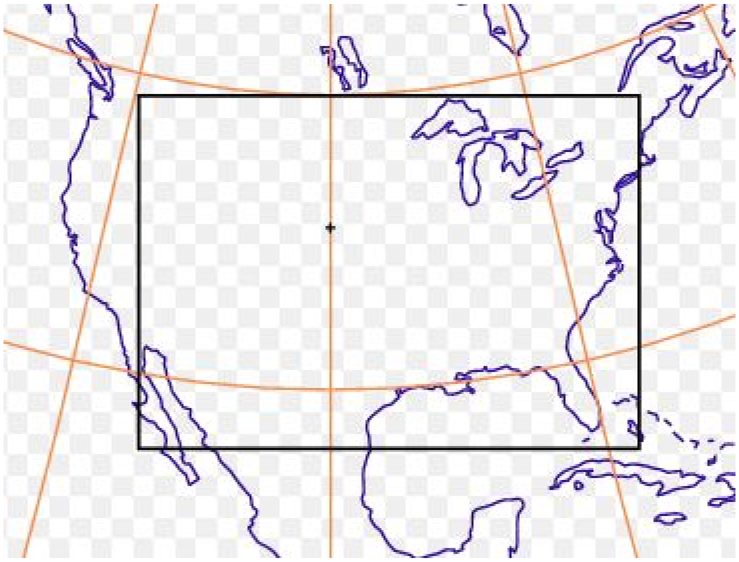


Figure 64. *The Unified Model sub-CONUS domain.*

Taking its initial and lateral boundary conditions from the 00z/12z 17-km horizontal grid spacing global configuration of the UM, the 2.2-km model initializes without data assimilation and runs out to T+120. This model configuration uses a 3D turbulent mixing scheme using a locally scale-dependent blending of Smagorinsky and boundary layer mixing schemes, stochastic perturbations are made to the low-level resolved-scale temperature field in conditionally unstable regimes (to
encourage the transition from subgrid to resolved scale flows) and the microphysics is single moment. Partial cloudiness is diagnosed assuming a triangular moisture distribution with a width that
is a universally specified function of height only. There is no convection parameterization in this or any
of the high-resolution UM configurations.

Details at a glance:

- Horizontal Resolution: 0.02 degrees (approx 2.2km at equator)
- Vertical levels: 70 (up to 40km)
- Time-Step: 100s
- Initialisation: From UK Met Global analysis at 0z and 12
- Boundary Conditions: Hourly Global UK Met
- Run Length: 5 days

Finite Volume Cubed-Sphere Dynamical Core (FV3) Versions

The GFDL Finite Volume Cubed-Sphere Dynamical Core (FV3) is a scalable and flexible dynamical core capable of both hydrostatic and non-hydrostatic atmospheric simulations. The full 3D hydrostatic dynamical core, the FV core, was constructed based on the Lin-Rood (1996) transport algorithm and the Lin-Rood shallow-water algorithm (1997). The pressure gradient force is evaluated by the Lin (1997) finite-volume integration method, derived from Green's integral theorem based directly on first principles, and demonstrated errors an order of magnitude smaller than other well-known pressure-gradient schemes. Finally, the vertical discretization is the "vertically Lagrangian" scheme described by Lin (2004).

The most unique aspect of the FV3 is its Lagrangian vertical coordinate, which is computationally efficient as well as more accurate given the same vertical resolution. Recently, a more computationally efficient non-hydrostatic solver is implemented using a traditional semi-implicit approach for treating the vertically propagating sound waves. This faster solver is the default. The Riemann solver option is more efficient for resolution finer than 1km, and also more accurate, because sound waves are treated nearly exactly. FV3 has been chosen as one of the candidates for the Next Generation Global Prediction System project (NGGPS), designed to upgrade the current operational Global Forecast System to run as a unified, fully-coupled system in NOAA's Environmental Modeling System infrastructure. For more information, please see https://www.gfdl.noaa.gov/fv3/.

For the 2017 FFaIR Experiment, we will be evaluating three different models that all utilize the FV3 dynamical core. The first will be referred to as the **FV3-GFS** and is a 13 km quasi real-time model in which 4 cycles per day are available and initialized with the operational GFS analyses and provided by EMC. The **FV3-CAPS** was provided by the CAPS/OU team. This version is a 3 km CAM that forecasts out to 84 hours and uses the Thompson microphysics scheme, GFDL planetary boundary layer (PBL) physics, and is initialized by the GFS. Finally, the **FV3-GFDL** was provided by GFDL and is also a 3 km CAM that forecasts out to 84 hours and is also initalized by the GFS and then nested over the CONUS. Some important differences are that the FV3-GFDL uses GFDL microphysics and an adjusted GFLD PBL scheme to balance some GFS issues.

ESRL High Resolution Rapid Refresh - Experimental (HRRRv3)

The operational HRRR model (version 2) (https://rapidrefresh.noaa.gov/hrrr/HRRR/Welcome.cgi?dsKey=hrrr_ncep_jet) is on a 3 km grid and uses boundary conditions from the hourly updated, radar-DFI-assimilated Rapid Refresh (RAP) model. The HRRR uses GSI hybrid data assimilation (instead of 3D-VAR), is initialized with latest 3-D radar reflectivity and features a WRF-ARW core version 3.6.1, Thompson microphysics, and is fully convection allowing. The operational HRRR is run every hour and produces hourly and sub-hourly forecasts out to 18 hrs.

We will be featuring the experimental version, HRRR version 3 (HRRRv3; https://rapidrefresh.noaa.gov/hrrr/HRRR), in the 2017 FFaIR Experiment. This version runs every hour with output to 18 hrs (01z, 02z, 04z, 05z, ...), 36 hours (00z, 03z, 06z…) or 48 hrs (12z). The experimental HRRRv3 remains on a 3-km grid with hourly runs that are changed to the forecast lengths listed above. The HRRRv3 is initialized with an hour of 3-D radar reflectivity using a latent-heating specification technique including some refinements in this latent-heating from the parent RAPv4 model. The HRRRv3 uses grid-point statistical interpolation (GSI) hybrid GFS ensemble-variational data assimilation of conventional observations. Building upon the advancements in the operational HRRRv2 at NCEP, HRRRv3 includes assimilation of TAMDAR aircraft observations, refines assimilation of surface observations for improved lower-tropospheric temperature, dewpoint (humidity) winds and cloud base heights and places more weight on the ensemble contribution to the data assimilation. HRRRv3 adds assimilation of lightning flash rates as a complement to radar reflectivity observations through a similar conversion to specified latent heating rates during a one-hour spin-up period in the model. HRRRv3 also contains numerous model changes including an update to WRF-ARW version 3.9 including the Thompson microphysics, transition to a hybrid sigma-pressure vertical coordinate for improved tropospheric temperature, dewpoint and wind forecasts along with a higher resolution (15 second) land use dataset. Physics enhancements have also been made to the MYNN planetary boundary layer (PBL) scheme and RUC land surface model along with additional refinements to shallow cumulus/sub-grid-scale cloud parameterizations including enhanced interactions with the radiation and microphysics schemes for greater retention of cloud features.

HRRRv3 is run hourly and provides forecasts as follows:

- Hourly output out to 36 hrs from runs at 00z, 03z, 06z, etc…
- Hourly output out to 18 hrs from runs at 01z, 02z, 04z, 05z, etc…
- Sub-hourly output to 15 hrs from all runs

Experimental Ensemble Guidance

ESRL/GSD HRRR Time-lagged Ensemble (HRRR-TLE)

Time-lagged ensembles are a computationally inexpensive substitute for full ensembles, using the "free" uncertainty information provided by a single, rapidly-cycled deterministic model (or a small set of deterministic models). Rather than running a large number of simultaneous simulations where initial conditions are perturbed based on uncertainty information from a data assimilation routine, time-lagging simply combines forecasts from deterministic model runs initialized at different times. Differences in the initial conditions from one run to the next are used in lieu of uncertainty estimates of the initial atmospheric state.

The HRRR-TLE combines forecasts from multiple deterministic HRRR runs, initialized at different times but valid at the same time. The current version, frozen for the duration of FFaIR 2017, uses the 3 most recent runs of the experimental ESRL HRRR, Version 3. The HRRRv3 operates with a ~2-h latency, and we chose to set hour zero of the HRRR-TLE forecast to the current time. For example, the 12z HRRR-TLE utilizes forecasts from HRRRv3 runs initialized at 8z, 9z and 10z.

Beyond HRRR-TLE forecast hour 12, we rely on a once-every-3-hours extension of the HRRRv3 to 36 hours. These HRRRv3 runs have a longer latency than the hourly 18-h HRRRv3. This means the longer-range HRRR-TLE forecasts have to "reach back" to much older HRRRv3 runs to construct probabilities from 3 members. For example, the 12z HRRR-TLE probabilistic QPF for hours 18-24 uses the 09z HRRRv3 (FHR 21-27), 06z HRRRv3 (FHR 24-30) and 03z HRRRv3 (FHR 27-33). This method could potentially result in a significant shift between the 12 and 13-h HRRR-TLE forecasts -- a topic worthy of discussion/evaluation during FFaIR.

The probabilistic QPF exceeding 1, 2, 5, 10, 25, 50, and 100-year Average Return Interval (ARI) are valid only for 6-hour intervals, are produced from bias corrected 6-h QPF (see previous section), are 40-km neighborhood probabilities, and also use an 100-km spatial filter to increase spread. So for example, a 50% value can be interpreted as, "a 50% chance of 6-h precipitation exceeding the 100-year ARI threshold, somewhere within 40 km."

The probabilistic QPF on saturated soil values represent the probability of 1-h calibrated rainfall exceeding various hourly thresholds at locations where the soil is saturated. For this application, "saturated soil" is defined where the soil moisture availability (ratio of top-soil-layer [0-1 cm] volumetric soil moisture to the volumetric soil capacity) is at least 95 percent. High hourly precipitation rates on saturated soil indicate high potential for initiation or worsening of localized urban or stream flooding (and potentially later impacts to mainstem rivers). These products use a 40-km neighborhood (e.g., 50% chance of 1" of QPF over saturated soil somewhere within 40-km), and benefit from increased spread through 60-km spatial and 3x1-h temporal filters, as employed for 1-h PQPF products. A summary of these products is in (Table 5). Web site: http://rapidrefresh.noaa.gov/hrrr/hrrrtle/

Table 5. *Probabilistic fields available from the HRRR-TLE for FFaIR 2017.*

HRRR-TLE Probabilistic QPF	Probability of 6-h precip > 0.5, 1.0, 2.0, 3.0, 6.0 in Probability of 3-h precip > 0.5, 1.0, 2.0, 3.0 in Probability of 1-h precip > 0.5, 1.0, 2.0 in
HRRR-TLE Probabilistic ARI Exceedance	Probability of 6-h precip exceeding 1, 2, 5, 10, 25, 50 and 100-yr ARI
HRRR-TLE Probabilistic QPF on Saturated Soil	Probability of 1-h rainfall on saturated soil > 0.5, 1.0, 2.0 in

ESRL/GSD Experimental HRRR Ensemble (HRRR-E)

The HRRR ensemble (HRRR-E) is initialized at 09 UTC each day from a combination of atmospheric RAP mean and GFS data assimilation ensemble (GDAS) perturbations along with HRRR land surface data. A total of 36 3-km HRRR members are initialized and then cycled hourly through 00 UTC using an Ensemble Kalman filter to assimilate conventional and radar observations each hour followed by the application of the HRRR cloud analysis and soil adjustment to each member. For the experiment, the ensemble will be run once a day at 00 UTC when the nine members produce 36 hr forecasts over a full CONUS domain. Stochastic soil moisture perturbations are introduced across all members at 09 UTC and boundary layer parameter perturbations are applied at 00 UTC along with lateral boundary perturbations and inflation during the cycled data assimilation to promote spread and represent both initial condition and model forecast uncertainties. An ensemble post-processing system is applied to the nine HRRRE forecast members to produce all-season weather hazard probabilities including heavy rainfall as is done with the time-lagged HRRR. HRRR-E uses WRF-ARW version 3.9 with the same physics configuration as the HRRRv3.

EMC Experimental High Resolution Ensemble Forecast Version 2 (HREFv2)

The HREF is an ensemble product generator utilizing multiple cycles of operational convective allowing models of ~3 km horizontal scale: namely the High-Resolution Window (HiresW; both the Weather Research and Forecast (WRF) Advanced Research WRF (ARW) and Non-hydrostatic Multiscale Model on the B-grid (NMMB) members) and the NAM CONUS nest. The membership of the HREF for the 2017 FFaIR experiment will match the parallel HREFv2 system with eight members: the two most recent runs of the NAM CONUS nest, the HiresW NMMB, and of two different HiresW ARW members (Figure 65). Probabilistic guidance has been enhanced with the addition of neighborhood probabilities (Harless et al. 2010) and Gaussian smoothing (Silverman 1986) of probabilities. Probabilities of precipitation exceeding and return interval values, and probability-matched (PM; Ebert 2001) mean fields also are generated by this version.

This experimental HREF will be run for the 00Z, 06Z, 12Z, and 18Z cycles, generating output to 36 hours from the cycle time. A description of the products most likely to be useful for the FFaIR experiment:

- Ensemble mean precipitation in three forms (as a conventional mean, as a PM mean, and as a “blend” of the conventional and PM means).
- Precipitation probability of exceedance (POE) at various fixed thresholds over several duration periods (e.g., percentage of the ensemble exceeding 3” over a 6 h period). This product and the other POE fields are expressed as neighborhood probabilities computed over a ~40 km radius neighborhood, and also are Gaussian smoothed.
- Precipitation POE of return interval values (e.g., percentage of ensemble exceeding the 50 year return interval value for a 6 h accumulation period).

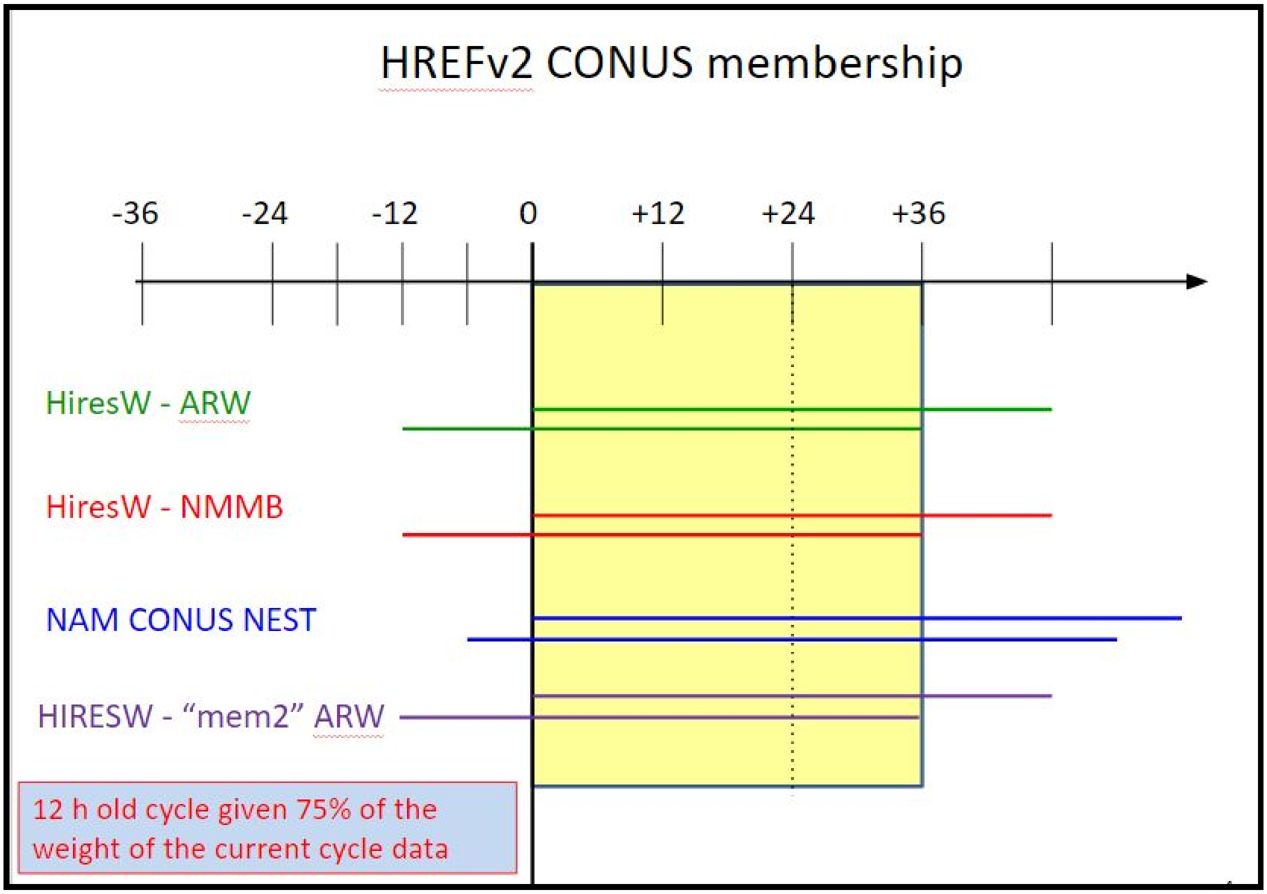


Figure 65. *The HREFv2 ensemble membership including one real-time and one time-lagged ARW, one real-time and one time-lagged NMMB, one real time and one time-lagged NAM CONUS Nest, one real-time and one time-lagged "mem2" ARW.*

In addition to a suite of fields relative to precipitation forecasting, derived probabilities will be available from a special HREFv2 run for the 2017 FFaIR Experiment (Table 6).

Table 6. *Probabilities available from the special HREFv2 for FFaIR 2017*

HREFv2 Probabilistic QPF	1-h, 3-h QPF > 0.5, 1, 2 inches 6-h, 12-h, 24-h QPF > 1, 2, 3 inches
HREFv2 Probabilistic FFG Exceedance	3-h QPF > 3-h, 6-h, 12-h, 24-h FFG
HREFv2 Probabilistic ARI Exceedance	QPF > 2, 5, 10, 25, 100-year ARIs
HREFv2 Fixed 100-km Filter Prob QPF*	3-h, 6-h QPF > 0.5, 1 inches
HREFv2 Variable (EAS) Filter Prob QPF*	3-h, 6-h QPF > 0.5, 1 inches

*Traditional ensemble probabilities are computed as the number of members that exceed a threshold divided by the total number of members at a given point. Such methods have been employed for many years in comparatively coarse resolution models. However, with high-resolution ensembles, i.e. those where convection is not parameterized, spatial displacement at the grid point level is large. This issue has motivated the development of methods which account for this spatial uncertainty. As part of the USWRP project, 2017 FFaIR will feature two different types of fractional coverage approaches for the generation of point probabilities will be evaluated (e.g. "neighborhood methods"). The first method expands the traditional point probability calculation to cover a 100-km radius around a given point. In other words, the probability represents the total number of points from each member in a 100-km radius around each grid point that exceed a threshold divided by the total number of points. The second method, which is considered experimental, attempts to account for the fact that a uniform radius is not always appropriate, i.e. orographically forced precipitation. In such cases, the traditional fractional coverage approach can reduce the probabilities of these often well handled events. Therefore, a variable radius approach has been developed based upon ensemble agreement scale (EAS) similarity criteria outlined in Dey et al. (2016). This approach varies the neighborhood radius size according to member-member similarity criteria. In this method, the radius sizes range from 10-km, for member forecasts that are in good agreement (e.g. lake effect, complex terrain, very short forecasts, etc.), to 100-km.

OU/CAPS WRF-ARW+FV3 SSEFX

The experimental Storm-Scale Ensemble Forecast (SSEFX) is generated with the Weather Research and Forecast (WRF) modeling system (Version 3.8.1), with the Advanced Research WRF (ARW) core, and the experimental GFDL FV3. CAPS will produce 10 (9 ARW, 1 FV3) members and one control member to support the FFaIR Experiment. The 3-km FV3 will feature Thompson microphysics. Membership details can be found in Table 7. Major features for 2017 include:

- 3-km horizontal grid spacing over the CONUS domain (1620×1120)
- WRF version 3.8.1 is used for 2017 season. (coupled with ARPS v5.4)
- 15 members, including 14 ARW and 1 FV3 members
- 00 UTC 60-hour forecast (FV3 deterministic member out to 84 h)
- ARPS 3DVAR analysis of radar data

Table 7. *Membership characteristics of the SSEFX for the 2017 FFaIR Experiment. For all ARW members: ra_lw_physics= RRTMG; ra_sw_physics=RRTMG; cu_physics=none*

Member	IC	BC	Radar	Microphy	LSM	PBL
arw_cn	00Z ARPSa	00Z NAMf	yes	Thompson	Noah	MYJ
arw_m2	arw_cn + arw-p1_pert	21Z SREF arw-p1	yes	P3	Noah	YSU
arw_m3	arw_cn + arw-n1_pert	21Z SREF arw-n1	yes	MY	Noah	MYNN

arw_m4	arw_cn + arw-p2_pert	21Z SREF arw-p2	yes	Morrison	Noah	MYJ
arw_m5	arw_cn + arw-n2_pert	21Z SREF arw-n2	yes	P3	Noah	MYNN
arw_m6	arw_cn + nmmb-p1_pert	21Z SREF nmmb-p1	yes	MY	Noah	MYJ
arw_m7	arw_cn + nmmb-n1_pert	21Z SREF nmmb-n1	yes	Morrison	Noah	YSU
arw_m8	arw_cn + nmmb-p2_pert	21Z SREF nmmb-p2	yes	P3	Noah	MYJ
arw_m9	arw_cn + nmmb-n2_pert	21Z SREF nmmb-n2	yes	Thompson	Noah	MYNN
arw_m10	arw_cn + arw-n3_pert	21Z SREF arw-n3	yes	Thompson	Noah	MYJ
fv3	GFS	-	no	Thompson	GFDL	GFDL

New for 2017, the SSEFX will be producing an experimental Localized Probability Matched Mean to be tested in FFaIR. The localized probability-matched mean (LPM) calculates the probability-matched mean over small patches (typically 6×6 or smaller) of the domain, using calculation regions with substantial overlap (typically around 60×60 gridpoints for each 6×6 patch), and then smooths the resulting field with a Gaussian smoother. The result is a forecast field that provides many of the advantages of the probability-matched mean (PM) while retaining small-scale structures in the resulting LPM field that may be informative or of meteorological interest. The LPM also does not suffer from potential errors resulting from considering all precipitation from a full CONUS domain for each smaller patch, limiting the influence to the nearest 100 km or so, ensuring that values used are from local storms and the local near-storm environment.

An example is shown in Figure 66, for a 3-hour rainfall forecast where rain was present over much of the southern Great Plains. The PM field (Fig. 66a) exhibits a typical highly-smoothed distribution of rainfall amounts; this is typical of PM forecasts. In contrast, the LPM field (Fig. 66b) retains much more small-scale structure, particularly for storms in Kansas, and the predicted rainfall in Wisconsin.

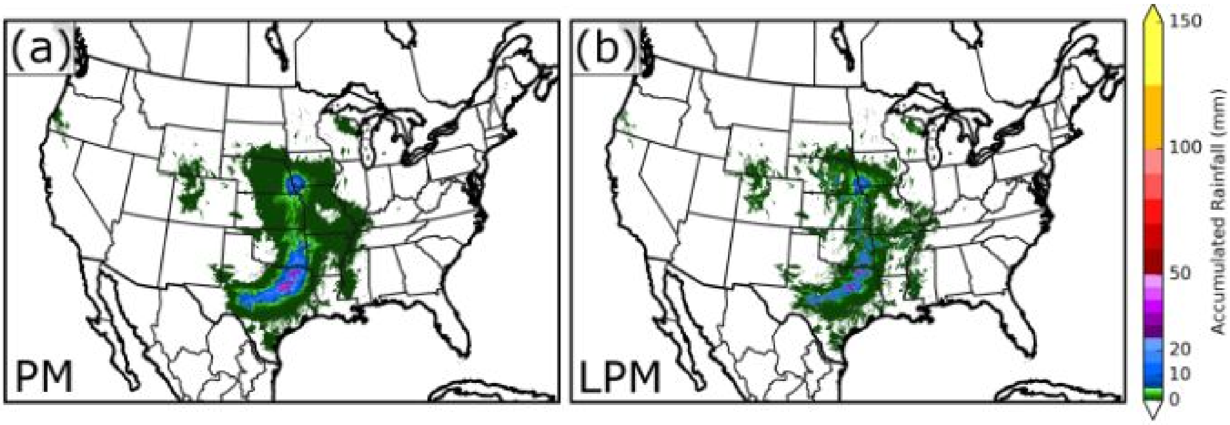


Figure 66. *Sample PM (a) and LPM (b) means for 3-hour accumulated rainfall forecasts.*

In addition to a suite of fields relative to precipitation forecasting (including conventional, probability-matched, and maximum QPF), derived probabilities will be available (Table 8).

Table 8. *Probabilities available from the SSEFX during FFaIR 2017*

SSEFX Probabilistic QPF	1-h, 3-h QPF > 0.5, 1, 2 inches 6-h, 12-h, 24-h QPF > 1, 2, 3 inches
SSEFX Probabilistic FFG Exceedance	3-h QPF > 3-h, 6-h, 12-h, 24-h FFG
SSEFX Probabilistic ARI Exceedance	QPF > 2, 5, 10, 25, 100-year ARIs

National Blend of Modes Version 3 (NBMv3):
The NBMv3 runs every hour with 15 different deterministic and ensemble systems. For the CONUS, typically 4 to 6 new model runs each hour, with up to 7 or 8 (~50% new) on four cycles. The NBMv3 uses TOD (Time of Day) concept, rather than the "model" cycle. Therefore, a 12z run of NBM V3.0 does not contain a single 12z model run. It is a data cutoff time (newest models are from 10-11z in this example; several 00z to 06z models included). NBM V3.0 Will be run each hour on top of hour, and available 50-60 minutes later. Models included are found in Table 9.

Table 9. *Data dependencies for the NBMv3.*

Global Models	Mesoscale Models
GFS - 0.25 degree	HRRR - 3 km
GEFS mean + members - 0.5 degree	NAM Low-Res - 12 km
CMC deterministic (PoP12/QPF only) - 1 degree	NAM High-Res - 3 km (pending)

CMC ensemble mean + members - 1 degree	HIRESW (NMMB and ARW cores) - 3 km
FNMOC - 1 degree	RAP - 12 km
GMOS - 2.5 km	SREF - 40 km
EKDMOS - 2.5 km	GLMP - 2.5 km
CCPA (used in NBM Precip SQM) - ~ 13 km	URMA (CONUS, AK, HI, PR) - 2.5 km
NBM Precip Stochastic Quantile Mapping - 2.5 km	

Other Experimental Tools

Experimental ERO "First-Guess" Field Using Reforecast Data, ARIs, Machine Learning

Developed by Greg Herman and Russ Schumacher of Colorado State University, this first-guess field is a prediction system comprised of random forests which is trained with 11 years (January 2003 - August 2013) of Days 2 and 3 GEFS Reforecast (GEFS/R) data. While forecasts are made on the GEFS/R grid, verification is first done on the Stage IV grid (in tabulating where Stage IV 24-hr QPE exceeds NOAA Atlas-14 ARI precipitation thresholds) and then projected onto the nearest GEFS/R grid point. The system then applies the latest 00Z GEFS/R run using select fields such as precipitation, convective available potential energy, and precipitable water, as examples. The product is issued daily for two 24-hr (12-12 UTC) forecast periods (Day 2 is hours 36-60, Day 3 is hours 60-84).

Through machine learning and decision trees, the system can produce the probability that an Average Recurrence Interval (ARI) exceedance within 40 km will occur. The 1000 trees, each created with a different predictor at the root, make up the "Random Forest." When a new set GEFS data is put into the system, the data traverses down a tree until a terminal node is reached, giving a deterministic prediction as the outcome. In the model, this value is either a "2" for a 10+ year ARI exceedance, a "1" for a 1-year ARI exceedance without a 10-year ARI exceedance, or a "0" for a non-exceedance. All of trees indicating 2's and 1's are tallied up and divided by the total numbers of trees (B=1000), which is essentially the probability of occurrence at that point for the two ARI levels.

The product issued is a CONUS grid indicating the following (Figure 67):

- Probability of at least one 1-yr ARI exceedance
- Probability of at least one 10-yr ARI exceedance

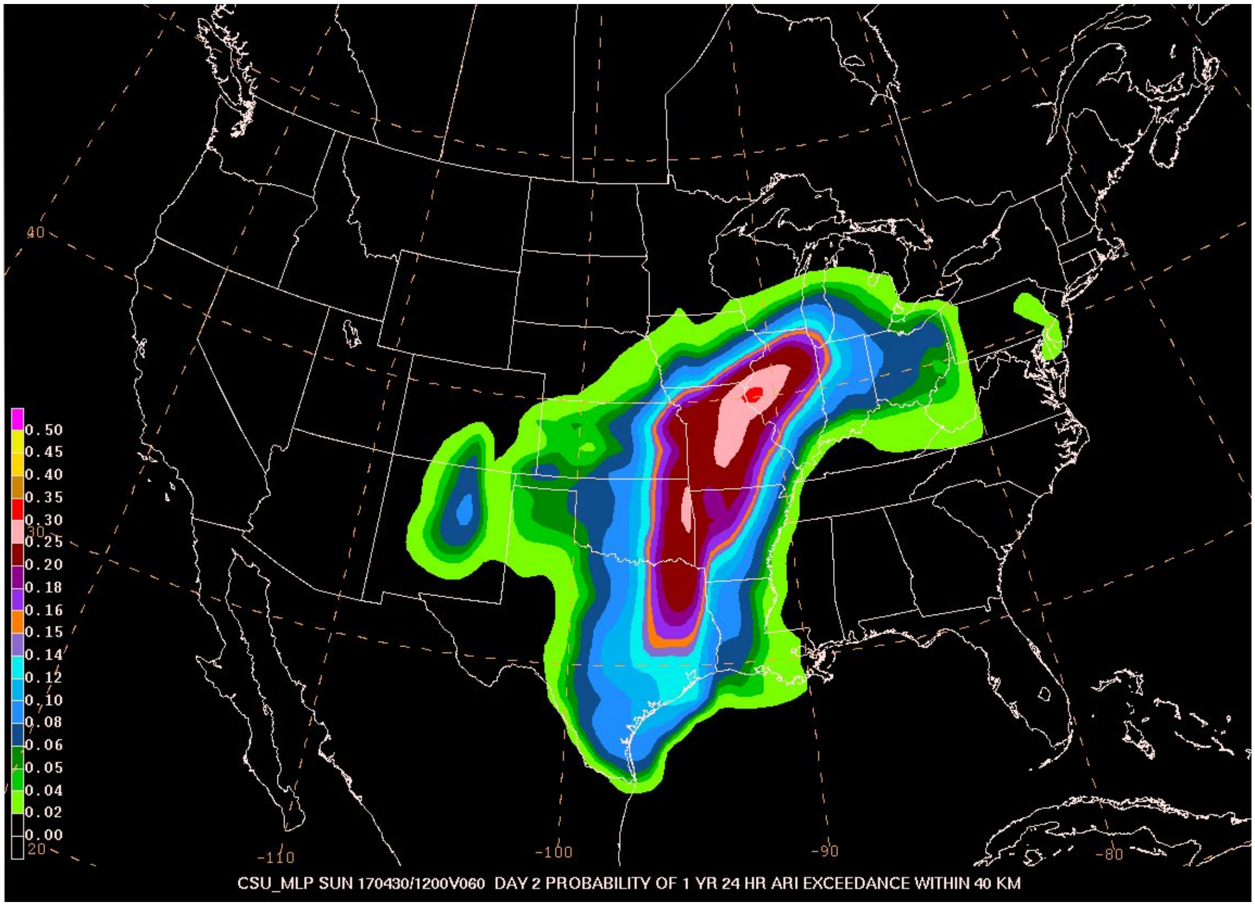


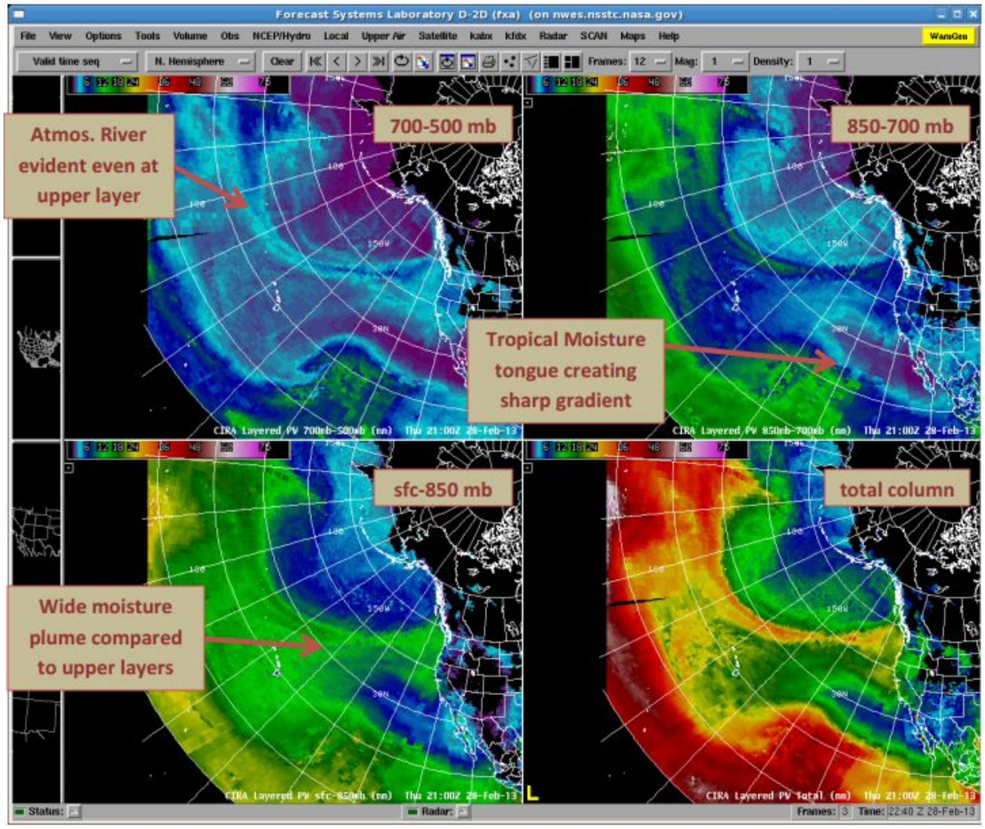
Figure 67. *An example of the Herman-Schumacher ARI exceedance product to be tested as a first-guess field for the experimental Days 2 and 3 Excessive Rainfall Outlooks. Here shown is the Day 2 probability of 1 year 24 hour ARI exceedance within 40 km of a point.*

Experimental Remotely-Sensed Products

CIRA LPW:

The total precipitable water products are developed by John Forsythe, Stanley Kidder, and Andrew Jones at the Cooperative Institute for Research in the Atmosphere (CIRA) as well as Sheldon Kusselson at NOAA NESDIS / Satellite Analysis Branch in Washington D.C. CIRA has partnered with the SPoRT Center to use SPoRT's expertise in transitioning products to the National Weather Service. This partnership has allowed CIRA to transition two total precipitable water products and the suite of layered precipitable water products to SPoRT's partner National Weather Service offices for evaluation and comment. A key component of this partnership is SPoRT's knowledge of transitioning products to the forecaster's native display environment, AWIPS/AWIPS-II (Figure 68). A major emphasis of CIRA is to create observation based products that can provide information in traditionally data poor regions, such as the oceans. This feature has made the precipitable water products widely popular with SPoRT's coastal National Weather Service partners. Each of the total PW products are available every six hours and has a

10 km resolution, while the layered PW product suite updates every three hours and is mapped to a 16 km Mercator grid. https://nasasporttraining.files.wordpress.com/2017/02/layered-precipitable-water-quick-guide-by-nasa-sport.pdf


Forecast Systems Laboratory D-2D (fxa) (on nwes.nsstc.nasa.gov)
File View Options Tools Volume Obs NCEP/Hydro Local Upper Air Satellite kabx kfdx Radar SCAN Maps Help
WarnGen
Clear
700-500 mb
850-700 mb
Atmos. River evident even at upper layer
Tropical Moisture tongue creating sharp gradient
sfc-850 mb
total column
Wide moisture plume compared to upper layers
6 12 18 24 36 48 58 75
30N

Valid time seq · N. Hemisphere · Frames: 12 · Mag: 1 · Density: 1

CIRA Layered PV 700mb-500mb (mm) Thu 21:00Z 28-Feb-13

CIRA Layered PV 850mb-700mb (mm) Thu 21:00Z 28-Feb-13

CIRA Layered PV sfc-850mb (mm) Thu 21:00Z 28-Feb-13

CIRA Layered PV Total (mm) Thu 21:00Z 28-Feb-13

Status: · Radar: · Frames: 3 · Time: 22:40 Z 28-Feb-13

Figure 68. *Example of CIRA Layer Precipitable Water (LPW) in AWIPS*

GOES-16 TPW/LPW:

The Total Precipitable Water (TPW) product is computed from the retrieved atmospheric moisture profiles and represents the total integrated moisture in the atmospheric column from the surface to the top of the atmosphere. This product will provide useful information to weather forecasters and hydrologists to improve their situational awareness for a number of situations that require forecasting of events, such as heavy rain, flash flooding, onset of Gulf of Mexico return flow, and the onset of the Southwest United States monsoon. The TPW product also serves to initialize the moisture field used in numerical weather prediction models.

APPENDIX C

WPC MODE Settings for Objective Verification

- 60 HR & 84 HR QPF verified against Stage IV QPE
    - 00Z forecast cycle used
    - Both QPF and QPE re-gridded to a common 5km lat/lon grid
    - CONUS mask applied to common grid
    - Thresholds of 0.5”, 1.0”, 2.0”, 4.0” and 6.0” investigated
- MODE
    - Grid stats harvested from MODE CTS
    - Circular convolution radius of 3 grid squares used
    - Double thresholding technique applied
- MODE Analysis
    - Summary of all forecasted vs. observed shapes throughout experiment
    - Describes centroid distance, angle, and interest