2016 Flash Flood and Intense Rainfall Experiment

Final Report

REVISED November 8, 2016

TABLE OF CONTENTS

Introduction	2
Experiment Description	2
Experimental Forecast Activities	2
Data	10
Operational and Experimental Guidance	12
Experimental Ensemble Model Guidance	15
Other Experimental Tools	18
Synoptic Overview and Daily Impacts Throughout the Experiment	20
Atmospheric Guidance Results	25
Performance of the Deterministic Convection Allowing Models	25
Ensemble Performance During the 2016 FFaIR Experiment	30
Day 2 Model Guidance Performance	36
Hydrologic Guidance Results	43
Other Experimental Tools Results	46
Experimental Probabilistic Forecast Performance	51
Summary and Conclusions	57
References	59
Appendix A	60
Appendix B	61

1. INTRODUCTION

Over four weeks during a period from June 20 to July 22, 2015, the Hydrometeorology Testbed at the Weather Prediction Center (WPC-HMT) hosted the fourth annual Flash Flood and Intense Rainfall (FFaIR) Experiment. In an effort to support improvements to WPC's operational Excessive Rainfall Outlook (ERO) in both the Day 1 and Day 2 time periods and explore the utility and accuracy of shorter, 6-hourly probability of flash flood forecasts, the FFaIR Experiment brought together over 30 participants (Appendix A) from the forecast, research, and modeling communities to investigate methods for improving flash flood forecasting.

The focus of the 2016 experiment was to use and evaluate high resolution guidance to improve flash flood forecasts from a national perspective down to a more local, Weather Forecast Office (WFO) perspective. To simulate the flow of information that occurs from a national center (e.g. WPC) to the local forecast offices, this year's experiment was again held concurrently with the flash flood experiment hosted by the Hydrometeorology Testbed in Norman, OK (HMT-Hydro). While participants in FFaIR focused on issuing experimental flash flood forecasts from the next 6 hours out to 2 days, participants at the HMT-Hydro experiment issued experimental flash flood watches and warnings during the 0-6 hr period.

The goals of the 2016 FFaIR Experiment were to:

- Evaluate ways to maximize the utility of high resolution convection-allowing models and ensembles for short-term flash flood forecasts.
- Identify the most effective forms and proper usage of available hydrologic guidance for the assessment of flash flood risk.
- Explore proposed changes to WPC's operational Excessive Rainfall Outlook by evaluating the utility of probabilistic flash flood forecasts for Day 1 and Day 2 respectively.
- Enhance cross-testbed collaboration as well as collaboration between the operational forecasting, research, and academic communities on the forecast challenges associated with short-term flash flood forecasting.

This report will provide a summary of the activities, subjective evaluations, and potential enhancements to operations resulting from the experiment.

2. EXPERIMENT DESCRIPTION

Experimental Forecast Activities

This year for the first time in the FFaIR experiment we divided the participants into two teams for early morning assessment of the experimental guidance. "Team Hydro" examined the experimental hydrologic guidance such as soil saturation, runoff, probabilities of QPF exceeding recurrence intervals, and the National Water Model output for streamflow anomalies. “Team Atmo” examined the experimental guidance for the atmosphere including deterministic and ensemble QPF, probabilities of QPF exceeding various thresholds, model simulated radar reflectivity and standard fields such as integrated vapor transport and precipitable water. Figure 1 shows various images of the teams working together.

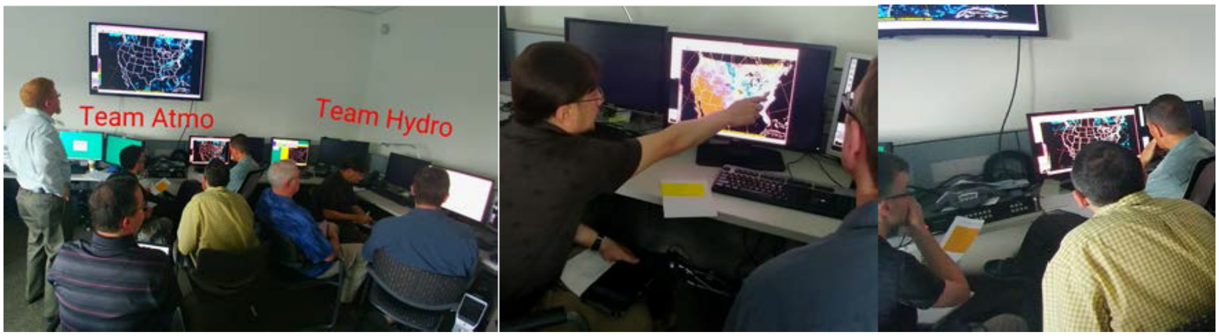


Figure 1. *Images of Team Atmo and Team Hydro working together during the experiment.*

The teams then chose a leader to brief the others on the fields-of-interest for that day, then the participants worked collaboratively using the experimental guidance to produce a 21-hour Day 1 Excessive Rainfall Outlook using contours of 2, 5, 10 and/or 30% probability of flash flooding occurring within 40 km of a point (Figure 2).

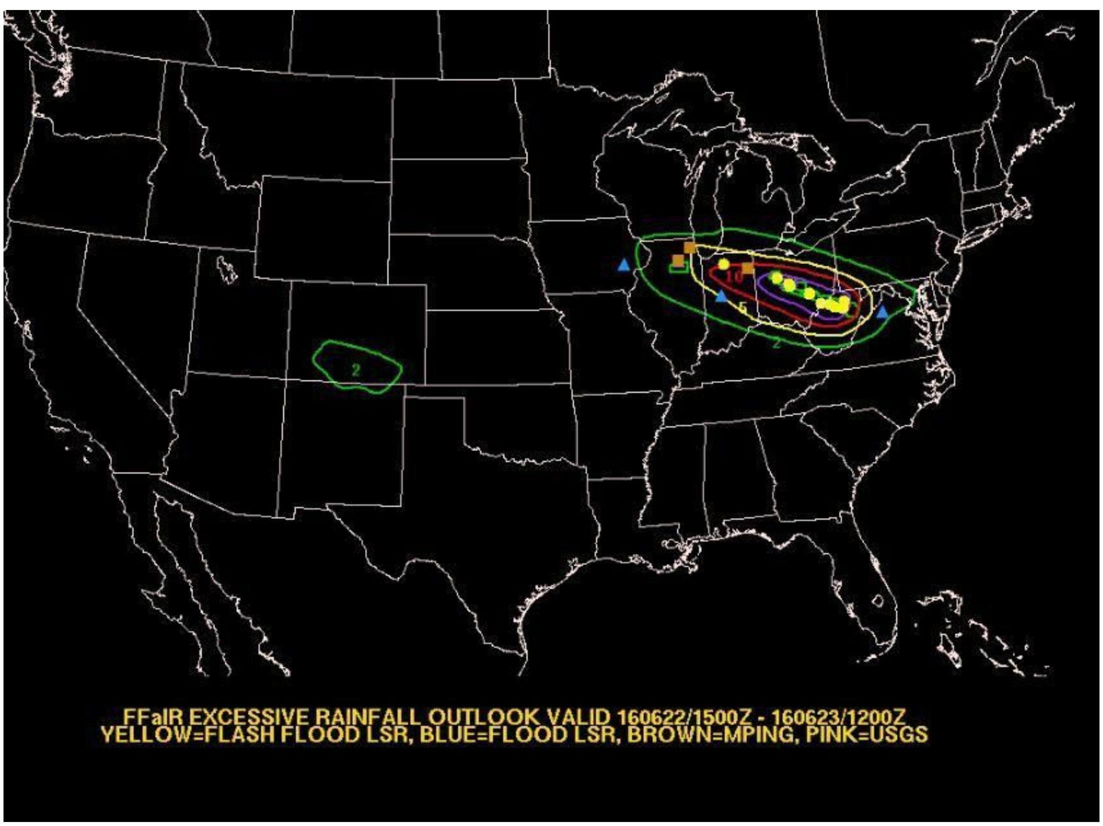


Figure 2. *Day 1 ERO issued during FFaIR overlaid with observations – Valid 15Z June 22 - 12Z June 23.*

Using updated guidance and current radar, the participants then issued a 6-hour forecast using contours of 10, 30 and/or 50% neighborhood probability of flash flooding (PFF) occurring within 40 km of a point between 18Z and 00Z. An example is shown in Figure 3. This forecast is most utilized by the HMT-Hydro team as it corresponds to their forecast window for issuing flood watches and warnings. It also serves to support improvements to the WPC MetWatch Desk operations by using experimental guidance in a shorter, 6 hour timeframe.

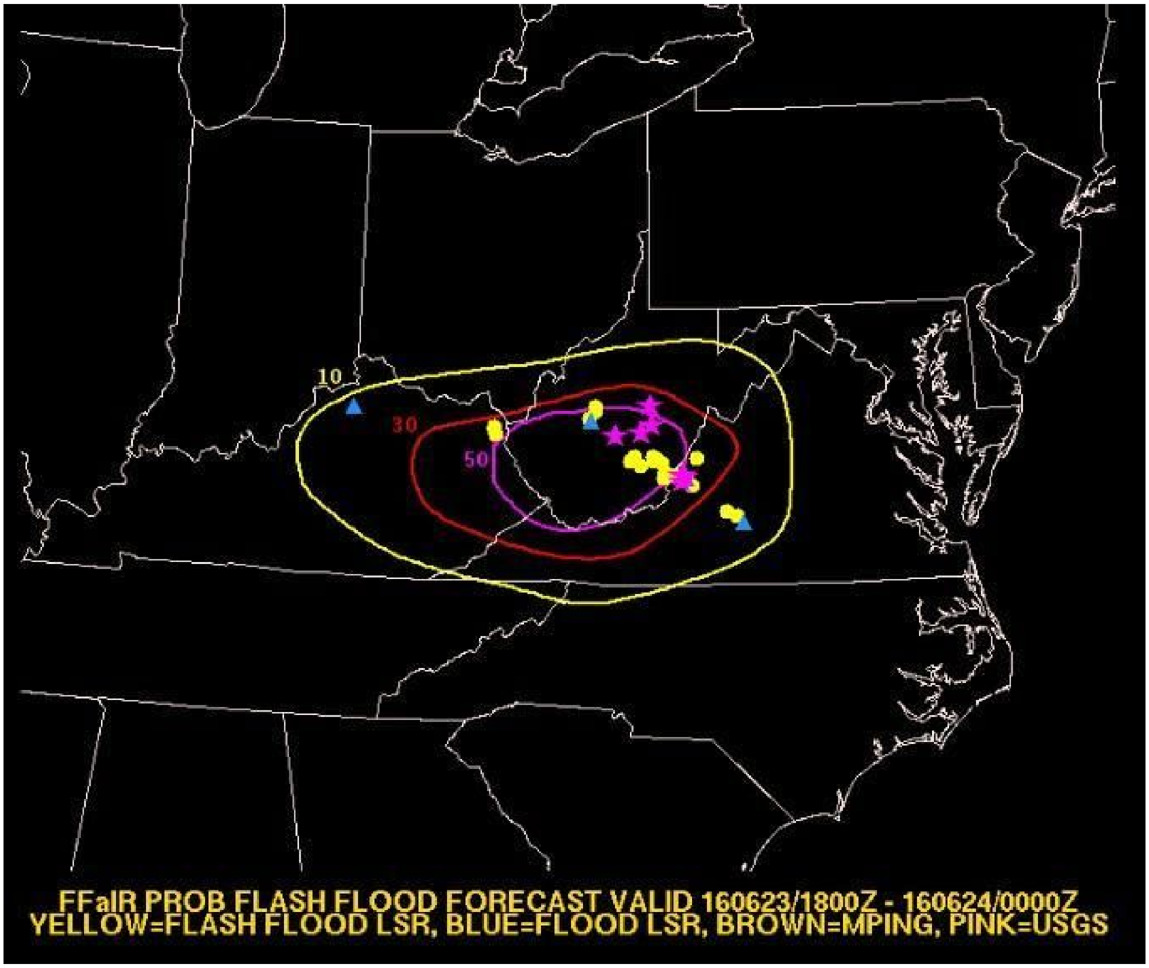


Figure 3. *PFF issued during FFaIR overlaid with observations – Valid 18Z June 23 - 00Z June 24.*

Lastly, the participants used the longer-range experimental guidance to issue an experimental Day 2 Excessive Rainfall Outlook, again using contours of 2, 5, 10 and/or 30% of neighborhood probability of flash flooding occurring within 40 km of a point, valid between 12Z and 12Z. (FIgure 4). Appendix B provides a breakdown of a typical day's schedule.

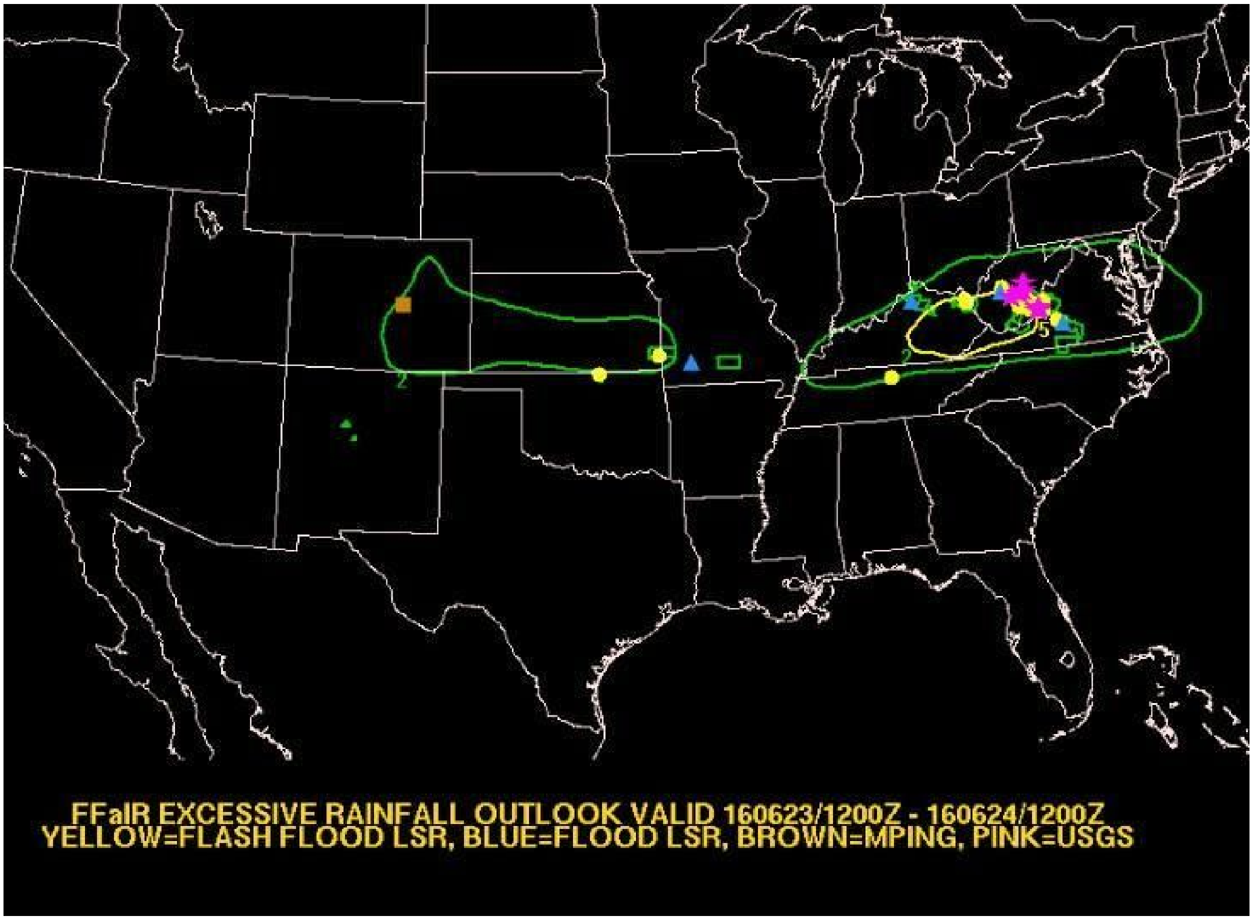


Figure 4. *Day 2 ERO issued during FFaIR overlaid with observations – Valid 12Z June 23 - 12Z June 24.*

Verification

Subjective evaluation consisted of 10 science questions presented to the participants by the testbed staff. These questions included evaluation of the forecasts participants created as well as other experimental models and tools used during the forecasting process. Participants used white boards to rank each experimental guidance or tool on a scale from 1 (very poor) to 10 (very good). Individual scores were then recorded and averaged to arrive at one score for the question. This was a change from previous years in which a 1 to 5 scale was used and participants were required to agree upon a common score. Figure 5 shows the participants ranking a question utilizing their white boards, which allowed each participant to have equal scoring representation.

Figure 5. *Participants evaluating experimental hydrologic and atmospheric model data and tools.*

A number of verification resources were used to score the experimental forecasts. Figure 6 displays an example of how a Day 1 ERO was evaluated. Panel (1) is the Day 1 ERO valid from 15Z June 22 -- 12Z June 23 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. 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 75% of the flash flood guidance. Participants were free to review all of these panels when coming up with a subjective score for each experimental forecast. For the other science evaluations, a combination of reports and MRMS QPE was typically used as verification.

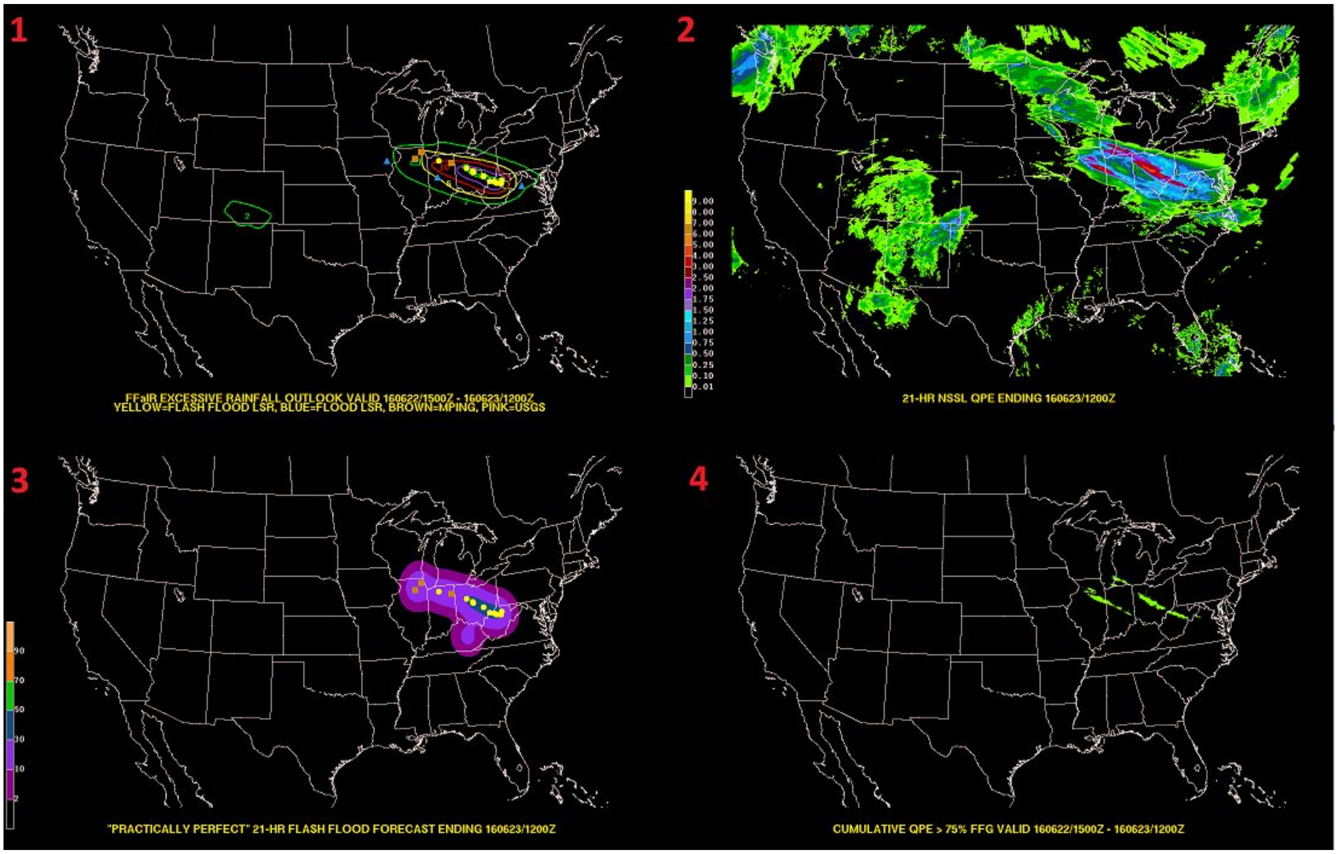


Figure 6. *Verification for Day 1 ERO valid 15Z June 22 -- 12Z June 23. (1) Display of the forecast contours with flash flood and flood reports and operational flash flood warnings overlaid. (2) MRMS QPE valid over the forecast period. (3) Practically perfect analysis. (4) Areas where MRMS QPE was greater than 75% of the flash flood guidance.*

Collaboration with NSSL's HMT-Hydro

Each day at 2 PM EDT we conducted a 2-way briefing with the HMT-Hydro group in Norman, OK. The FFaIR participants presented an atmospheric and hydrologic assessment of the current flood threats using the experimental guidance accompanying the experimental Day 1 ERO and the 18-00 UTC PFF, which serve as a starting point for the HMT-Hydro group's watch/warning issuances for the day. This collaboration simulates the flow of information from a National Center to a WFO for flood potential situational awareness and provides very valuable discussion opportunity. The exchange includes not only how we chose our forecast foci and the process that led decision making, but also about the origins of flash flood guidance, new tools, and different ideas on establishing verification (including media resources). The HMT-Hydro group also used this exchange to evaluate the previous days' forecasts offering feedback to the participants and commenting on the value of the forecasts to their experiment process (Figure 7).

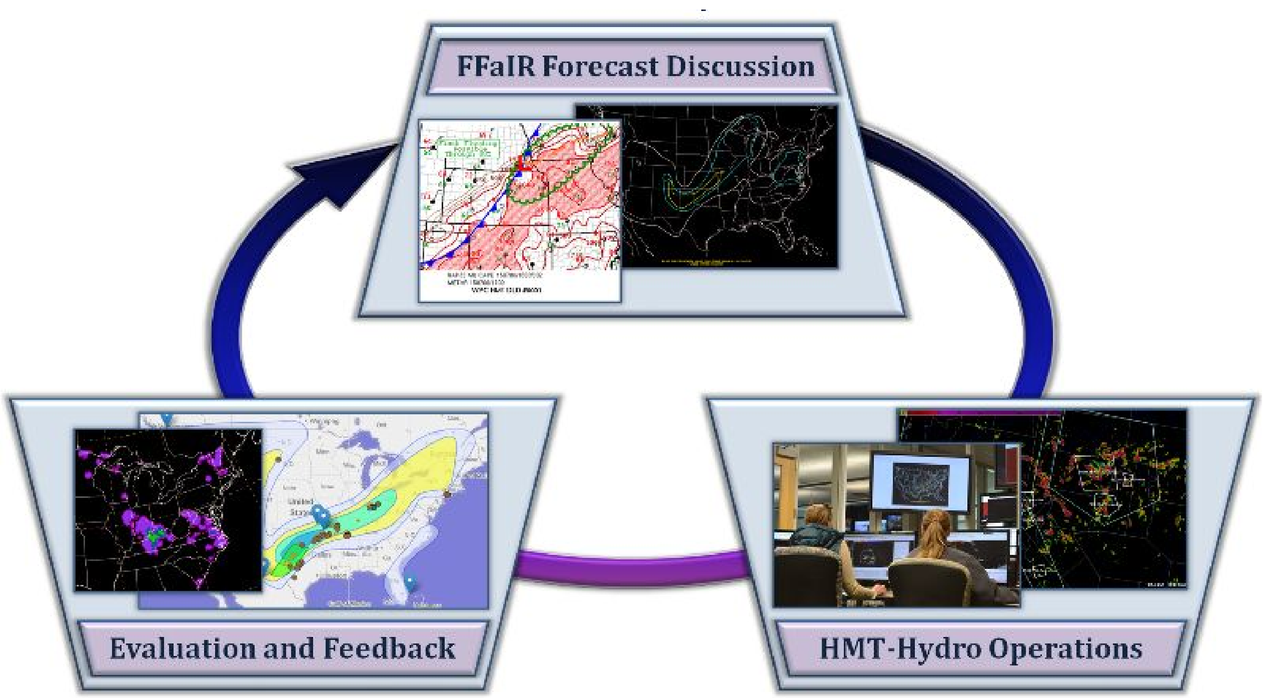


Figure 7. *The functional cycle of the collaboration with the HMT-Hydro group including forecast briefings and feedback (image courtesy HMT-Hydro)*.

HMT-Hydro Group Evaluations of FFaIR Forecasts

The HMT-Hydro Team provided the FFaIR participants with scored and written evaluations of the experimental forecasts provided to them for feedback (Figures 8 and 9). The scores reflected skill in spatial accuracy, probabilistic contour values, and useful application to the watch/warning issuances. This feedback is used to fine-tune the guidance, forecast information and presentation of the briefing to the HMT-Hydro group.

#1 – Rate the following statement: The spatial accuracy of the Day 1 FFaIR Excessive Rainfall Outlook for the previous day was skillful.

Ranking	6/20	6/21	6/22	6/23	6/27	6/28	6/29	6/30	7/11	7/12	7/13	7/14	Total
1 - Strongly Disagree										1			1
2 - Disagree							4	4					8
3 - Neutral					1	2	1	1	4	2		3	14
4 - Agree	5			3	4	3			2	3	4	2	26
5 - Strongly Agree		5	5	2							2	1	15
Average	*4*	*5*	*5*	*4.4*	*3.8*	*3.6*	*2.2*	*2.2*	*3.333333*	*3.166666*	*4.333333*	*3.666666*	*3.71875*

#2 – Rate the following statement: The probability values of the Day 1 FFaIR Excessive Rainfall Outlook for the previous day were accurate.

Ranking	6/20	6/21	6/22	6/23	6/27	6/28	6/29	6/30	7/11	7/12	7/13	7/14	Total
1 - Strongly Disagree								1		1			2
2 - Disagree							1	3	3	3		3	13
3 - Neutral	2				4	2	3	1	2	2		1	17
4 - Agree	3	3		5	1	3	1		1		4	2	23
5 - Strongly Agree		2	5								2		9
Average	*3.6*	*4.4*	*5*	*4*	*3.2*	*3.6*	*3*	*2*	*2.666666*	*2.166666*	*4.333333*	*2.833333*	*3.375*

Figure 8*. An example of the scoring process at HMT-Hydro when evaluating FFaIR Forecasts.*

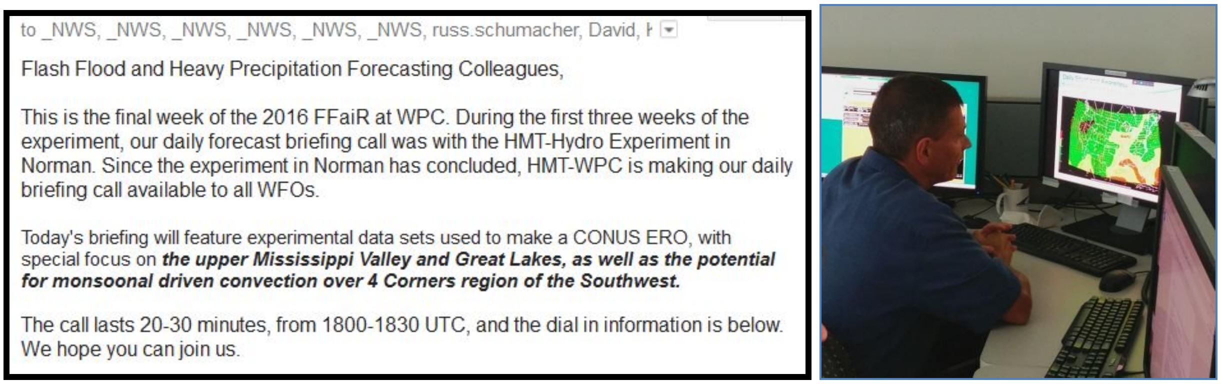
FFaIR Evaluation Comments **Forecast Date: 14 July 2016**

EVALUATION	COMMENTS
Day 1 Excessive Rainfall Outlook – Spatial Accuracy	They hit the areas, and talking about the main areas they exceeded flash flood guidance. The area of excessive rainfall was correctly covered.
Day 1 Excessive Rainfall Outlook – Probabilities	The area up in Maine was too high, the max in Ohio was incorrect. But looking at the QPE , they didn't have much.
Day 1 Probabilistic FF Forecast – Spatial Accuracy	They had the orientation incorrect where the heaviest rainfall occurred because in the briefing they were expecting more of a south to north track. Orientation was incorrect but in terms of focusing our attention, it did that.
Day 1 Probabilistic FF Forecast – Probabilities	Too high -- 30 too high, 10 was justified. Recognized that this is very subjective given the procedure you're supposed to follow in evaluating probabilities.
FFaIR Products/Briefings Benefitting SA	All the focus areas were where the concern areas were.
FFaIR Products/Briefings Benefitting Watch Decisions	
Performance of Day 1 ERO vs. Day 2 ERO	Spatially they did better but by magnitudes they did worse.

Figure 9. *An example of the written comments provided by HMT-Hydro regarding the utility of the FFaIR Experimental Forecasts.*

Collaboration with SOO-DOH Community

The HMT-Hydro ran for only three weeks. To supplement the briefing process during the fourth week, we instead featured emails to the SOO-DOH community stating the geographical flash flood problems of the day and offering GoToMeeting information. The FFaIR participants presented the SOO-DOH callers the same atmospheric and hydrologic assessment of the current flood threats using the experimental guidance as with the HMT-Hydro collaboration (Figure 10). Although the audience was typically small, collaborative discussions regarding the experimental data and forecast reasoning were useful.


to _NWS, _NWS, _NWS, _NWS, _NWS, _NWS, russ.schumacher, David, I

Flash Flood and Heavy Precipitation Forecasting Colleagues,

This is the final week of the 2016 FFaiR at WPC. During the first three weeks of the experiment, our daily forecast briefing call was with the HMT-Hydro Experiment in Norman. Since the experiment in Norman has concluded, HMT-WPC is making our daily briefing call available to all WFOs.

Today's briefing will feature experimental data sets used to make a CONUS ERO, with special focus on ***the upper Mississippi Valley and Great Lakes, as well as the potential for monsoonal driven convection over 4 Corners region of the Southwest.***

The call lasts 20-30 minutes, from 1800-1830 UTC, and the dial in information is below. We hope you can join us.

Figure 10. *An example of the email sent to the SOO-DOHs and WPC Forecaster, Bob Oravec, delivering the PPT briefing on the GoToMeeting Presentation.*

Data

In addition to the full multi-center suite of operational deterministic and ensemble guidance, the 2016 FFaIR Experiment featured four experimental ensemble systems: the Storm-Scale Ensemble of Opportunity (SSEO, Jirak et al., 2012) provided by the Storm Prediction Center (SPC), the experimental Storm-Scale Ensemble Forecast (SSEFX) from the University of Oklahoma (OU) and Center for Analysis and Prediction of Storms (CAPS), the Experimental High Resolution Ensemble Forecast (HREFX) provided by EMC, and a modified version of the SSEO provided by WPC. The experiment also featured several deterministic experimental high-resolution guidance systems: the experimental High Resolution Rapid Refresh (ESRL HRRRv2) provided by the Earth Systems Research Laboratory (ESRL), the experimental North American Mesoscale Model rapid refresh (NAMRR) provided by EMC, and the National Water Model provided by the Office of Water Prediction (OWP). Table 1 summarizes the model data that was the focus of the experiment. More information about each model is provided below.

Table 1. *Featured 2016 FFaIR deterministic and ensemble model guidance (Experimental guidance is in the darker shade)*

Provider	Model	Grid Spacing	Forecast Hours	Notes
NCEP	NAM	12 km (parent) 4 km (nest)	84 (parent) 60 (nest)	Operational NAM, includes 12 km parent model and 4 km CONUS nest
RFCs	Flash Flood Guidance	5 km	01, 03, 06, 12 and 24 hour values	CONUS mosaic grid created by compiling individual RFC-domain grids

NCEP	HRRR	3 km	15	High resolution, hourly updated, convection allowing nest of the Rapid Refresh (RAP) model
NCEP	HRRRv2	3 km	18	NCO parallel of HRRRv2, hourly updated, convection allowing
EMC/NSSL	NMMB ARW WRF-NSSL	4 km	48 36 (WRF-NSSL)	Hi 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	3 and 6 hr (2, 5, 10, 25 and 100 year intervals)	Precipitation frequency estimates based on historical observations.
EMC	NAM Rapid Refresh	3 km (nest) 12 km (parent)	18 hours (hourly) 60 hours (nest/00, 06, 12, 18 hours) 84 hours (parent/00 06, 12, 18 hours)	Features an hourly forecast and assimilation cycle for its 3 km CONUS nest and its North American 12 km domain. Uses hybrid 3DEnVar and incorporates radar reflectivity into its assimilation system via a complex cloud analysis approach.
ESRL/GSD	ESRL HRRRv2	3 km	Hourly out to 18 hours every hour Hourly out to 36 hours every 3 hours	Experimental version of the NCEP HRRRv2, hourly updated, convection allowing
ESRL/GSD	HRRR Time-lagged Ensemble (HRRR-TLE)	3 km	24	Neighborhood ensembling approach calculated over a 3km grid of time-lagged ESRL HRRRv2 deterministic members. Probabilities at a point refer to the chance of exceeding a given threshold somewhere with a 40-km radius around that point.
SPC/ESRL/WPC	WPC-SSEO (7 members)	4 km	24	Modification of the original SSEO provided by SPC; removes SPCWRF and adds HRRR
EMC	HREFX	5 km	36	Experimental version of HREF with 8 members (fewer NAM CONUS nest than operational) 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	15-member (14 ARW+1 NMMB) ensemble forecast (up to 60-h) starting at 0000 UTC
EMC	GEFSX	111 km	384	Beta version of the operational GEFS using a new Frequency Match calibration technique.

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. 11). 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.

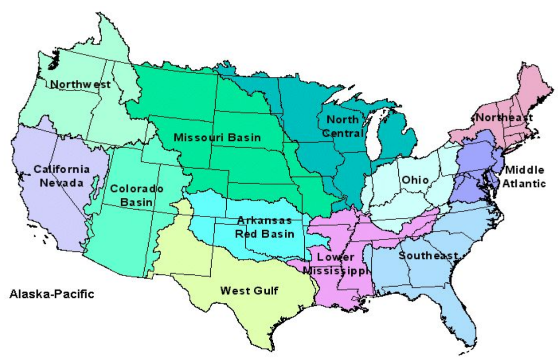


Figure 11. *The domains for each NWS River Forecast Center (NOAA/NWS (water.weather.gov).*

Precipitation Recurrence Intervals

Precipitation Recurrence Intervals (RIs) 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 the rarity of a rainfall event for a given area, alerting forecasters to abnormal or potentially extreme rainfall events. RIs are available for intervals of 2, 5, 10, 25, 100, 500 and 1000 years, and are measured in inches. Atlas-14 data over Texas and Pacific NW are still pending, and RIs do not account for antecedent conditions. An example of the 6 hour, 100 year recurrence interval for the CONUS is shown in Figure 12.

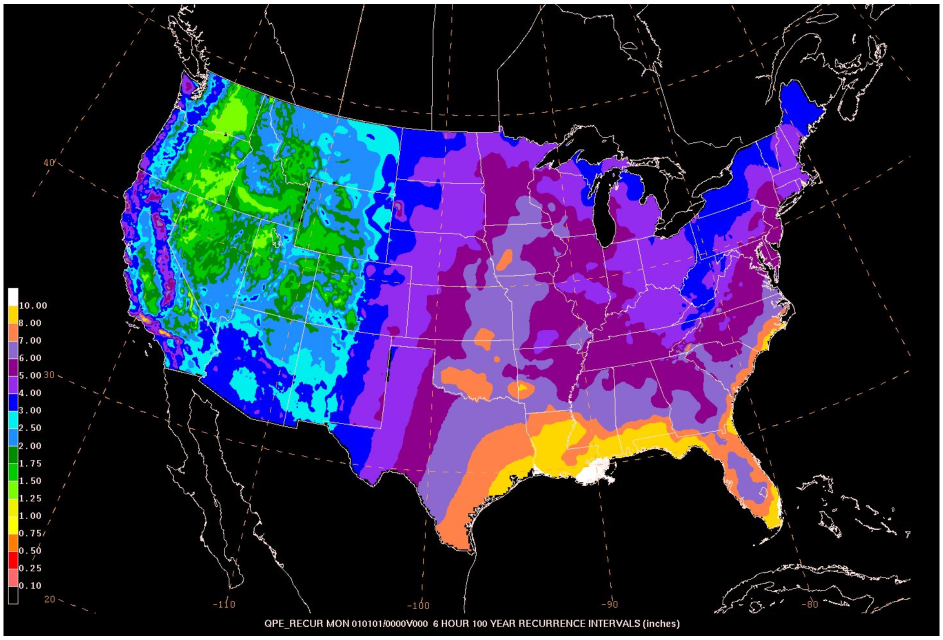


Figure 12. *An example of a full RI map (100 year RI over 6 hours) available to forecasters both operationally at WPC and in the FFaIR Experiment.*

<u>OWP's National Water Model (NWM)</u>

The NWM is an hourly uncoupled analysis and forecast system that will provide streamflow for 2.7 million river reaches and other hydrologic information on 1km and 250m grids.

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 NoahMP Land Surface Model (LSM) to simulate land surface processes. Separate water routing modules perform diffusive wave surface routing and saturated subsurface flow routing on a 250m grid, and muskingum-cunge channel routing down NHDPlusV2 stream reaches. 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). An example of the streamflow anomaly product is shown in Figure 13. 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 1,240 reservoirs. The NWM is run in four configurations:

1. Analysis and assimilation 1 hour snapshot
2. Short-Range 18 hour deterministic forecast
3. Medium-Range 10 day deterministic forecast
4. Long-Range 30 day ensemble forecast

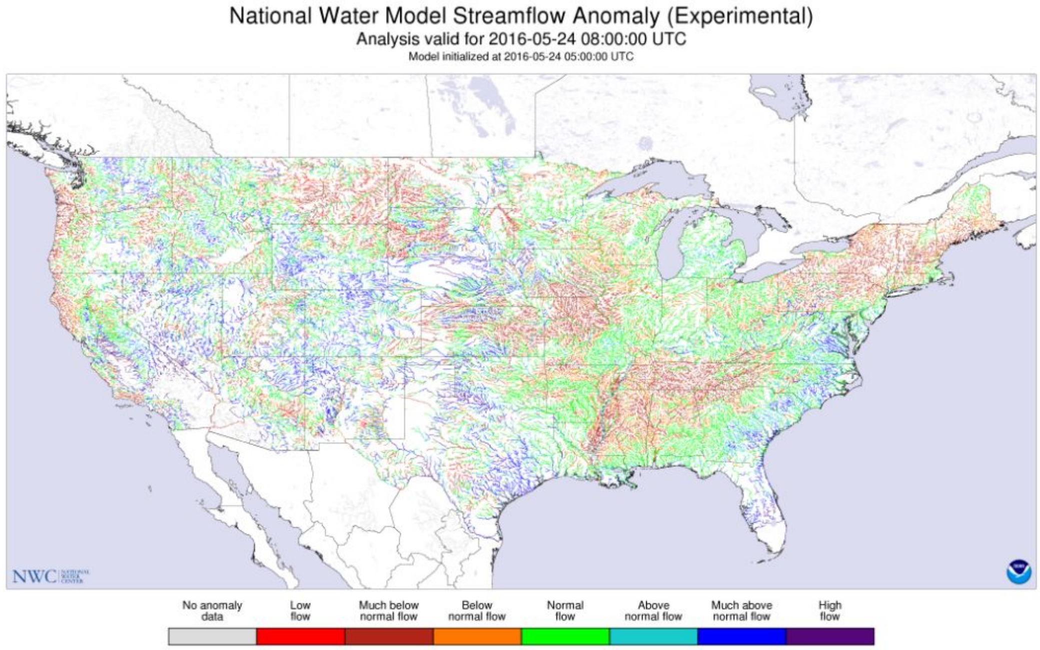


Figure 13. *An example of the National Water Model Streamflow Anomaly map. Note: For FFaIR 2016, NWM output will be provided as web graphics only.*

EMC North American Model Rapid Refresh (NAMRR)

The North American Mesoscale Rapid Refresh system (NAMRR) is an hourly-updated version of the North American Mesoscale (NAM) forecast system and its data assimilation system. The current version of the NAM, upgraded in August 2014, currently performs an atmospheric analysis only every 3 hours on its 12 km North American domain - there is no cycling of its nested domains (forecasts are provided every 6 hours). The NAMRR, in contrast, features an hourly forecast and assimilation cycle for its 3 km CONUS and Alaska nest domains in addition to its North American 12 km domain. Similar to the NAM, the NAMRR's data assimilation system uses hybrid 3DEnVar, where the ensemble members come from the Global Data Assimilation System's Ensemble Kalman Filter. The NAMRR also incorporates radar reflectivity into its assimilation system via a complex cloud analysis approach. The use of radar reflectivity in the analysis and the institution of a distinct data assimilation cycle generally improve the utility of the short-term forecasts (0-12h) from 3 km CONUS and AK nests relative to their operational counterparts.

Like the NAM, the NAMRR is based upon the Non-hydrostatic Multiscale Model on the B-grid (NMMB) and runs the same suite of physics options. However the NAMRR has had several physics changes and some adjustments to its dynamics. The brief highlights include:

- Changes to the BMJ parameterized convection for the 12 km North American domain to improve QPF biases and threat scores in the cool season.
- Updates to the Noah LSM to reduce low 2m dewpoint biases in the cool season.
- Updates to the MYJ turbulence scheme to address maritime shallow cloudiness issues.
- Changes to RRTM radiation and Ferrier-Aligo microphysics to address 2m T warm bias during the warm season.
- Increase the frequency of calls to physics in the nests.

- Advect specific humidity at every model dynamics timestep (instead of every other).

Hourly cycles out to 18 hours were available for the Parent (12 km) and CONUS nest (3 km). At 00, 06, 12, and 18 UTC forecasts went out to 60 hours for the CONUS nest and 84 hours for the Parent 12 km domain.

High Resolution Rapid Refresh Operational NCEP (HRRRv1) and Experimental NCEP/ESRL (HRRRv2)

The operational HRRR version 1, run at NCEP, (http://rapidrefresh.noaa.gov/hrrr) is 3 km resolution which uses boundary conditions from the hourly updated, radar-DFI-assimilated Rapid Refresh (RAP) model. It features a WRF-ARW core, Thompson microphysics, and is fully convection allowing. The HRRR is initialized with latest 3D radar reflectivity and features a WRF-ARW core version 3.4.1, Thompson microphysics, and is fully convection allowing. The operational HRRRv1 is run every hour and produces hourly and sub-hourly forecasts out to 15 hrs.

During the FFaIR Experiment, the operational HRRR was in the process of being upgraded to version 2 (HRRRv2), which uses GSI hybrid data assimilation (instead of 3DVAR), WRF-ARW version 3.6.1 and remains on a 3 km grid. The RRTM/Goddard radiative schemes have been replaced with the RRTMG for both longwave and shortwave radiation. The HRRRv2 also features an enhanced boundary layer scheme that includes the radiative effects of subgrid scale (unresolved) clouds and a land surface scheme with seasonally-varying vegetation fraction and modified wilting points/increased roughness length values. The microphysics scheme is upgraded to an aerosol-aware Thompson microphysics for more explicit cloud formation. All of these changes help significantly reduce a low-level warm/dry bias during the daytime in the warm season along with a reduction in a high surface wind speed bias. A high bias in convective precipitation forecasts is also reduced.

HRRRv2 is run at NCEP (NCO parallel) every hour and produces hourly and sub-hourly forecasts out to 18 hrs. The HRRRv2 is also run experimentally at ESRL every hour with hourly and sub-hourly forecasts out to 18 hrs with the exception of every third hour where the hourly forecasts are extended to 36 hrs. The NCEP HRRRv2 and ESRL HRRRv2 are configured identically with the exception of the ESRL forecast length extension and minor differences in observation availability/use for data assimilation.

Experimental Ensemble Model Guidance

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

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 2016, uses 3 recent runs of the experimental ESRL HRRRv2 (see above). The HRRRv2 operates with a ~2h latency, and with hour zero set to the current time. For example, the 12z HRRR-TLE utilizes forecasts from HRRRv2 runs initialized at 8z, 9z and 10z. The precipitation and ARI exceedance probabilities available in FFaIR 2016 utilize bias corrected ensemble QPF exceedance within a 40 km neighborhood + 80 km spatial filter. The probabilistic runoff within a 40 km neighborhood + 80 km spatial filter is derived from ensemble QPF and the HRRR/RUC Land Surface Model soil state calculations.

<u>EMC FFaIR Experimental High Resolution Ensemble Forecast (HREFX)</u>

The HREFX is an ensemble product utilizing multiple cycles of operational convective allowing models of ~3-4 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 resolution of the HREFX is 5 km and the membership has been modified for the experiment, utilizing fewer NAM nest members than are used by the operational HREF. Probabilistic guidance has been enhanced with the addition of neighborhood probabilities (Harless et al. 2010) and Gaussian smoothing (Silverman, 1986) of probabilities. In addition to the conventional, probability-matched, and blended QPF means from the HREFX, probabilities of precipitation exceeding flash flood guidance (FFG) and return interval values, and probability matched (PM; Ebert 2001) mean fields also are generated by this version.

The HREFX ran for the 00Z, 06Z, 12Z, and 18Z cycles, generating output to 36 hours from the cycle time. Although the operational HREF is an 11-12-member ensemble (5-6 NAM nest members), the HREFX is an eight-member ensemble (two NAM nest, six HiresW from the three most recent cycles of the WRF-ARW and NMMB runs).

<u>SPC Storm-Scale Ensemble of Opportunity (SSEO)</u>

The SSEO is a high-resolution, multi-model, multi-physics, convection-allowing ensemble system produced by the Storm Prediction Center. Issued at 00 and 12 UTC, it is composed of seven deterministic high-resolution members (Table 2). At WPC, the ensemble mean is displayed at 4 km, although each member can be viewed independently at its native resolution (Table 2). Two of the members (the operational ARW and NMM hi-res windows) are time-lagged by 12 hours to provide additional initial condition diversity (Jirak et al, 2012). The NSSL WRF-ARW and EMC WRF-NMM are non-operational and can be subject to outages; the four hi-res window members (HRW-ARW and HRW-NMM) are operational, but can be supplanted with other hi-res runs (e.g. hurricane models) if the need arises (Jirak et al, 2012).

At WPC, a modified version of the SSEO is also employed, which replaces the EMC WRF-NMM member (member 6) with the latest cycle of the HRRR. This was done to mitigate the high QPF bias that has been observed with the EMC WRF-NMM member. Additionally, the WPC-SSEO is run at 06 and 18 UTC; the 06 and 18 UTC cycles feature five-time lagged members (members 1, 2, and 4 are time-lagged 6 hours, members 3 and 5 time-lagged 18 hours) along with the 06 UTC cycles of the HRRR and NAM nest.

Table 2*. Membership characteristics of the SSEO and WPC-SSEO. Members denoted by the asterisk (*) are time lagged by 12 hours. For the WPC-SSEO, member six is changed from the EMC WRF-NMM to the HRRR. Adapted from Jirak et al (2012).*

SSEO Member	Model	Provider	Resolution	PBL	Microphysics
01	WRF-ARW	NSSL	4 km	MYJ	WSM6
02	HRW-ARW	EMC	4.2 km	YSU	WSM6
03	HRW-ARW*	EMC	4.2 km	YSU	WSM6
04	HRW-NMMB	EMC	3.6 km	MYJ	Ferrier
05	HRW-NMMB*	EMC	3.6 km	MYJ	Ferrier
06 *06*	EMC WRF-NMM *HRRR*	EMC *ESRL*	4 km *3 km*	MYJ *MYNN*	Ferrier *Thompson*
07	NAM-NMMB Nest	EMC	4 km	MYJ	Ferrier

Probabilities within a 40-km neighborhood were available for the WPC-SSEO in hourly intervals out to 24 hours. The probability of the 3 hour and 6 hour QPF exceeding certain notable thresholds are derived by determining how many members predict precipitation to exceed the relevant threshold using the neighborhood maximum QPF value at each individual grid point. The probability of QPF exceeding various precipitation recurrence interval values are derived by subtracting the recurrence interval value from the neighborhood maximum QPF of each member at each individual grid point, then determining how many members predict precipitation to exceed the recurrence interval. The neighborhood probability of the 3 hour and 6 hour QPF exceeding FFG values are derived by subtracting the FFG value from the neighborhood maximum QPF of each member at each individual grid point, then determining how many members predict precipitation to exceed FFG.

OU/CAPS WRF-ARW SSEFX

The experimental Storm-Scale Ensemble Forecast (SSEFX) was generated with the Weather Research and Forecast (WRF) modeling system (Version 3.7.1), with the Advanced Research WRF (ARW) core, and the NCEP operationally used NMMB modeling system. CAPS produced 15 ARW (or ARW + NMMB) members to support the FFaIR Experiment. Major features for 2016 include:

A. **3-km** horizontal grid spacing over the CONUS domain (1680×1152, Figure 1), same as in 2015
B. **WRF version 3.7.1** (coupled with ARPS v5.4)
C. **15 members** (14 ARW + 1 NMMB)

EMC One-Degree, Calibrated, Experimental Global Ensemble Forecasting System (GEFSX)

The one-degree calibrated GEFSX is a beta version of the current operational GEFS provided by EMC specifically for the purposes of the experiment. It consists of the same twenty members plus one control as the operational GEFS and it forecasts out to sixteen days at a one-degree grid spacing provided once a day at 00Z. New in this beta version is a new calibration technique applied to QPF. The calibration technique employed in the GEFSX is a frequency matched method. The GEFSX provides both calibrated six and twenty-four hour precipitation as well as probabilities of exceeding various QPF thresholds.

Other Experimental Tools

Excessive Precipitation with Elevated Convection (EPEC) Index

The EPEC tool is a diagnostic tool and was designed to help aide in identifying where heavy precipitation associated with elevated convection may occur. EPEC is calculated as follows:

EPEC = KINX + PWAT + (DIV_{250} x 100,000)

KINX = K-index
PWAT = Precipitable water (mm)
Div = Divergence at 250 hPa (s-1)

The index can be applied to any numerical model from which the above variables are able to be derived or outputted. The variables that make up the equation demonstrated a strong signal and low variability. The divergence term is scaled because it is many orders of magnitude smaller than the KINX or PWAT. 50^{th} percentile minimum values for the variables are:

- K-index → 35
- PWATs → 37 mm
- Divergence → 5 x 100,000 s^{-1}

Percentile values for the EPEC Index are as follows:

- 74 → 25th percentile
- 86 → 50th percentile
- 98 → 75th percentile

Using the above percentiles as a basis, the EPEC Index will be defined in the experiment by the raw index values that translate to the chance of elevated convection occurring. The index values are as follows:

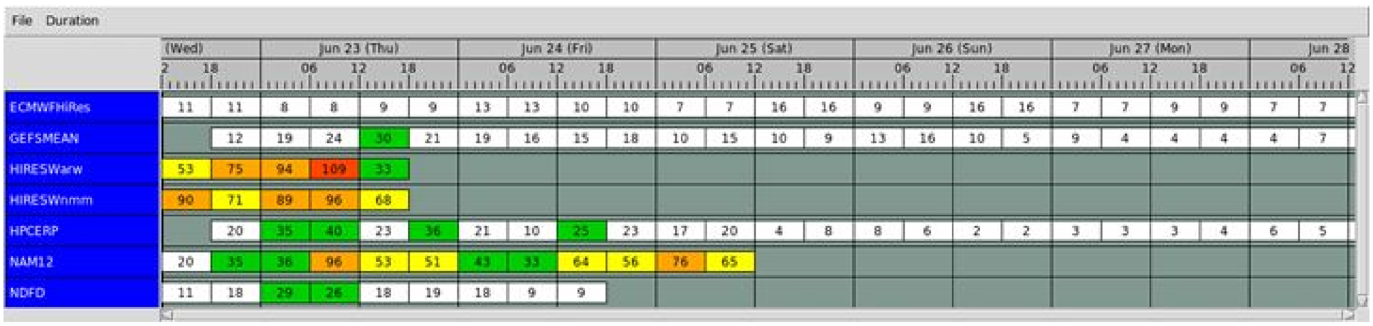
- 70-80 -- Low chance
- 80-90 -- Medium chance
- 90+ -- High chance

For more information, please see the National Weather Association presentation by Foscato and Market (2015) available at: http://www.nwas.org/meetings/nwas15/presents/2571.pptx

Extreme Precipitation Forecasting Table (EPFT)

The EPFT is an AWIPS II Graphical Forecast Editor (GFE) procedure designed as a contribution to situational awareness of extreme precipitation events. Events classified as “extreme” have a 100-year Annual Recurrence Interval (ARI), meaning that they have a 1% probability of occurring in any given year for any location. The procedure compares QPF to precipitation estimates from the NOAA Atlas 14 to highlight when and where the models are predicting rainfall nearing or exceeding an extreme event.

Running the procedure produces the EPFT, which is populated with the maximum ratio in the grid of the model QPF to the 100-year ARI precipitation estimate, expressed as a percentage, for a user-specified area and duration. The version of the EPFT running at WPC for the 2016 FFaIR experiment was for the six and twenty-four hour 100-year ARIs over the entire CONUS. The EPFT was designed to allow the forecaster to quickly identify and compare models that indicate potential for heavy or extreme rainfall according to an extreme precipitation climatology. Figure 14 shows an example of the EPFT that was used during the experiment along with the approximate interpretations of the ratios in Table 3.


File Duration

	(Wed)		Jun 23 (Thu)				Jun 24 (Fri)				Jun 25 (Sat)				Jun 26 (Sun)				Jun 27 (Mon)				Jun 28	
ECMWFHiRes	11	11	8	8	9	9	13	13	10	10	7	7	16	16	9	9	16	16	7	7	9	9	7	7
GEFSMEAN		12	19	24	30	21	19	16	15	18	10	15	10	9	13	16	10	5	9	4	4	4	4	7
HIRESWarw	53	75	94	109	33																			
HIRESWnmm	90	71	89	96	68																			
HPCERP		20	35	40	23	36	21	10	25	23	17	20	4	8	8	6	2	2	3	3	3	4	6	5
NAM12	20	35	36	96	53	51	43	33	64	56	76	65												
NDFD	11	18	29	26	18	19	18	9	9															

Figure 14. An example of the EPFT table from Wednesday, June 22 with the various model guidance.

Table 3. Approximate interpretation for the various QPF/ARI ratios.

QPF/ARI*	Potential for Damaging Impacts to Life and Property**
< 25%	Minimal
25-50%	Low
50-75%	Moderate
75-100%	High
100-125%	Very High
125-150%	Extreme
150-175%	Very Extreme

*Compared to the 100-year ARI rainfall estimates for durations of 6- and 24-hours
**Impacts from flooding also depend on soil moisture and type, terrain, areal coverage of rainfall, and geography (urban vs. rural).

GOES-R Advanced Baseline Imager (ABI) Algorithm for Rainfall Rate/QPE

During the 2016 FFaIR Experiment, the GOES-R ABI Algorithm for Rainfall Rate/QPE was evaluated. Currently, a proxy from GOES-13 and GOES-15 for the official GOES-R QPE algorithm is being used ahead of the official launch of GOES-R. The product tested during FFaIR had a resolution of 4 km and used a combination of infrared (IR) and microwave (MW) satellite data to compute estimated QPE. It refreshes every 15 minutes and recalibrates hourly using the MW satellite data allowing for QPE amounts in time-scales from 15 minutes to 24 hours. A potential use for this tool is to fill in radar coverage gaps, such as in the mountains, with satellite QPE estimates to get a more accurate representation of the rainfall. Figure 15 shows an example of the 24-hour GOES-R QPE product.

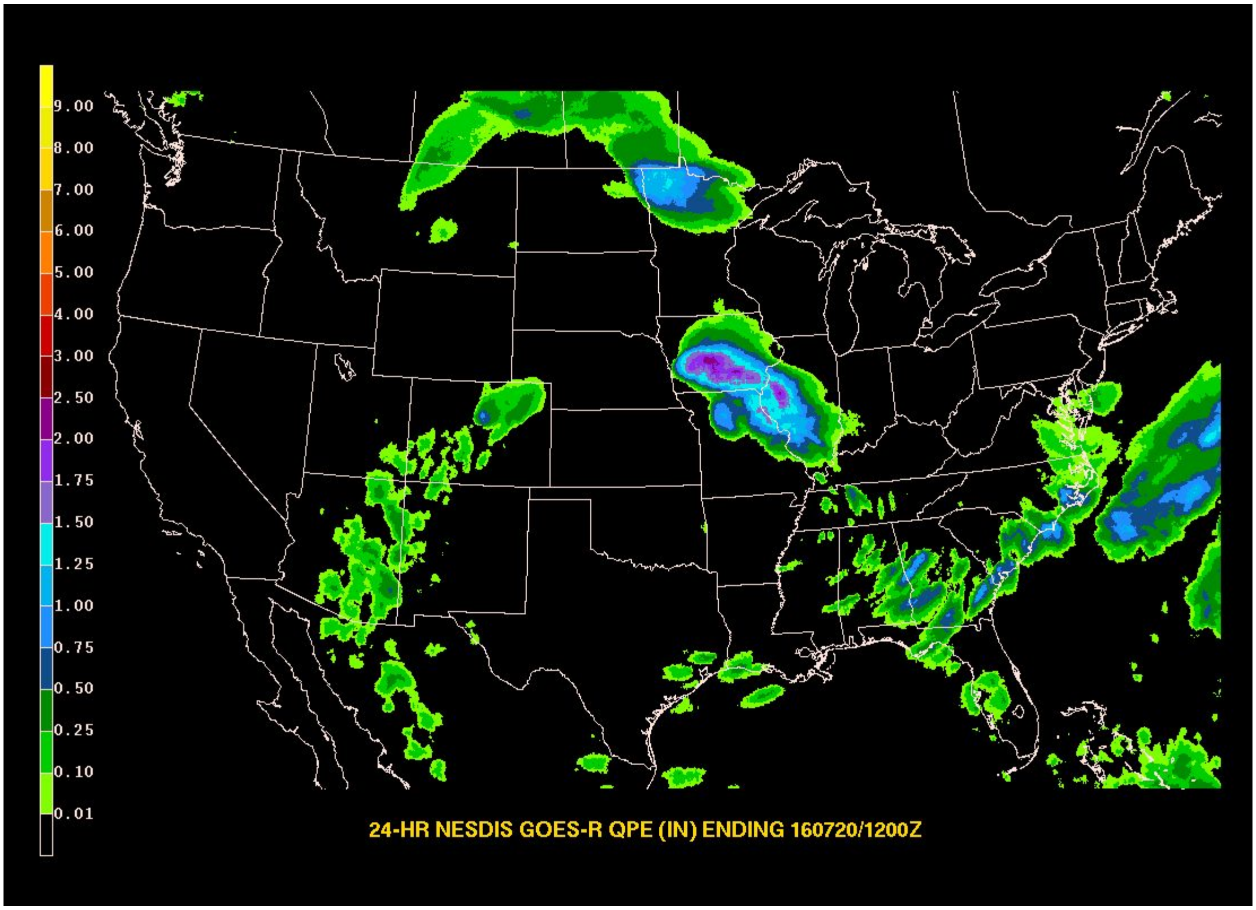


Figure 15. 24-HR GOES-R QPE estimation valid from 12Z July 19 -- 12Z July 20.

3. SYNOPTIC OVERVIEW AND DAILY IMPACTS THROUGHOUT THE EXPERIMENT

Over the four weeks of the 2016 FFaIR experiment, a wide variety of weather and flooding challenges affected the United States. There was a mix of very active weather at times with a few large scale flood events and quieter weather and only isolated flood risks. Figure 16 shows the 500 hPa mean geopotential height over the United States during the first half of FFaIR (June 20 - July 2 in panel 1) and second half (July 11 - 23 in panel 3). In both, a large ridge dominated much of the country centered over the Four Corners Region and western Texas. The ridge was more expansive during the second half of FFaIR than the first half, during which a weaker trough was positioned off the East coast and higher heights across the Southeast. Panels 2 and 4 in Figure 16 display the 500 hPa geopotential height composite anomalies for the first half of FFaIR and second half of FFaIR, respectively. During the first half of the experiment, 500 hPa heights were anomalously high over much of the country except over the Great Lakes region and the Northeast where they were near normal. In the last two weeks of FFaIR, 500 hPa heights were just slightly above normal over much of the country except for below-normal heights in the Northwest.

Figure 17 displays the column (1000- 500 hPa) precipitable water anomalies over both halves of the experiment. Over the first half of the experiment, precipitable water anomalies were highest in the Southwest United States and spreading into Kansas and Missouri in the central Plains. Much of the Eastern U.S. along with the Great Lakes and northern Plains had below normal 1000-500 hPa precipitable water anomalies. During the second half of the experiment, precipitable water anomalies were above normal along the Mississippi River from Minnesota south to Louisiana. Large areas of below normal were found over California and the Southwest as well as in the Mid-Atlantic and along the eastern Southeast coast.

The highest impact event to occur during the 2016 FFaIR Experiment was the devastating flash flooding event that took place from June 23-24 and affected areas of West Virginia the most. Between 8-10 inches of rain fell due to several rounds of heavy convection in the hardest hit Greenbrier County, WV. In total for the event, over 1,200 homes were either damaged or completely destroyed in the event and 23 people lost their lives making it the deadliest flash flood in the United States since May 2010.

Other notable events also occurred throughout the four weeks including two strong monsoonal events that took place on July 1st and 2nd. On July 1, widespread flash flooding hit in and around the Las Vegas, NV metro area. Numerous swift water rescues were performed and there was one reported fatality. The next day on July 2, Tucson, AZ broke their daily rainfall record which stood for 118 years with a total of 0.82 inches of rain. Several roads were flooded, motorists stranded, and the city's streetcar system had to be shut down due to high water. The first day of the second half of the FFaIR Experiment saw widespread flash flooding in Minnesota and Wisconsin on July 12-13. Up to 10 inches of rain fell and there were two reported fatalities. Finally, the western part of West Virginia was hit with flash flooding on July 14-15 when numerous roads were overrun with water in Huntington, WV and there was one death when a vehicle was swept away. Table 4 lists these events and others over the course of the entire experiment. Overall, the 2016 FFaIR participants experienced a wide variety of events over the four weeks from relatively quiet days with zero or very few isolated events to the worst flash flooding event in six years in terms of fatalities in the West Virginia floods on June 23-24.

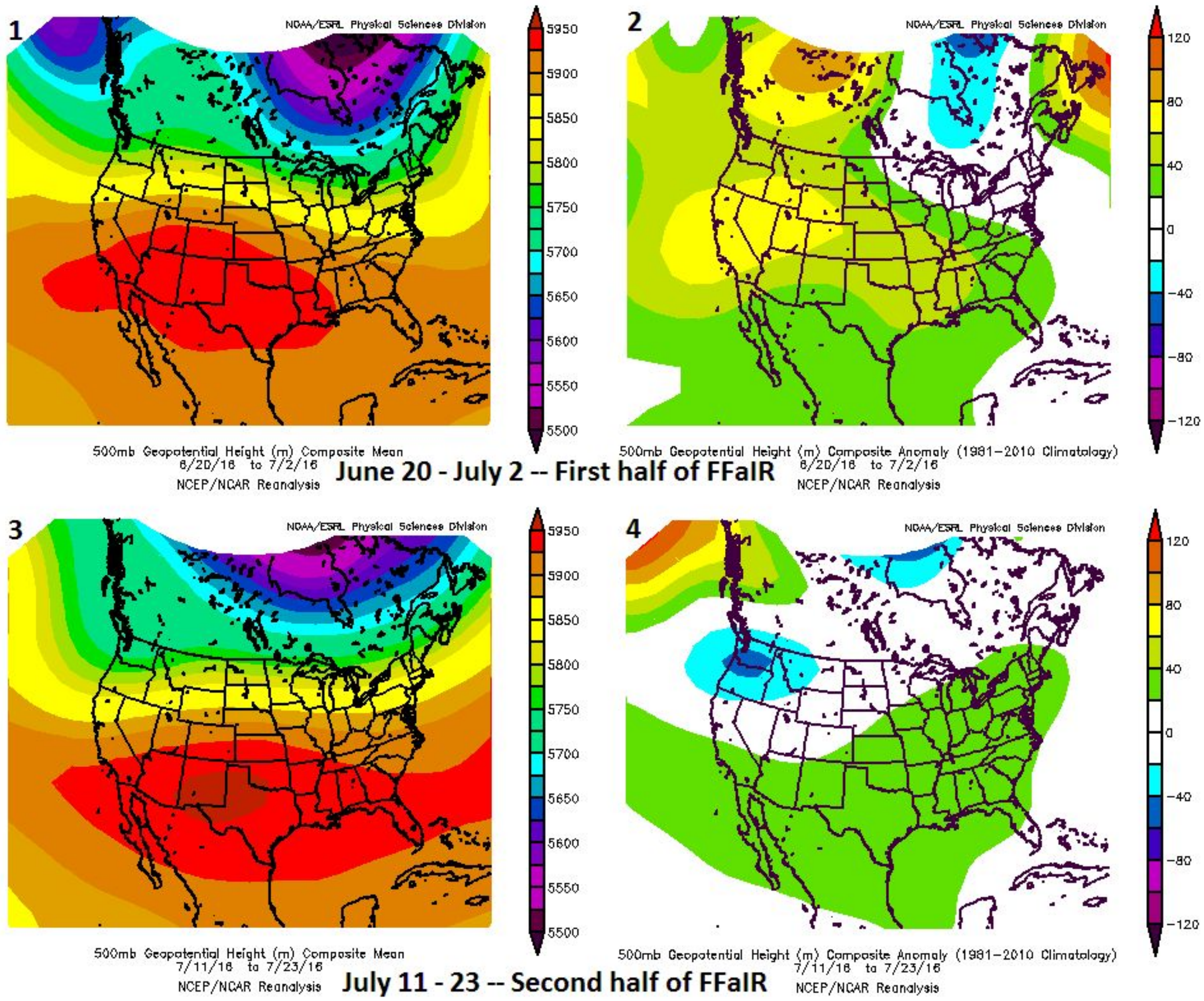


***Figure 16**. (1) 500 hPa mean geopotential height and (2) 500 hPa geopotential height composite anomalies for the first half of FFaIR covering June 20 - July 2, 2016. (3) 500 hPa mean geopotential height and (4) 500 hPa geopotential height composite anomalies for the second half of FFaIR covering July 11 - 23, 2016. Images generated from the NCEP/NCAR Reanalysis provided by NOAA/ESRL/Physical Sciences Division (http://www.esrl.noaa.gov/psd/data/composites/day/).*

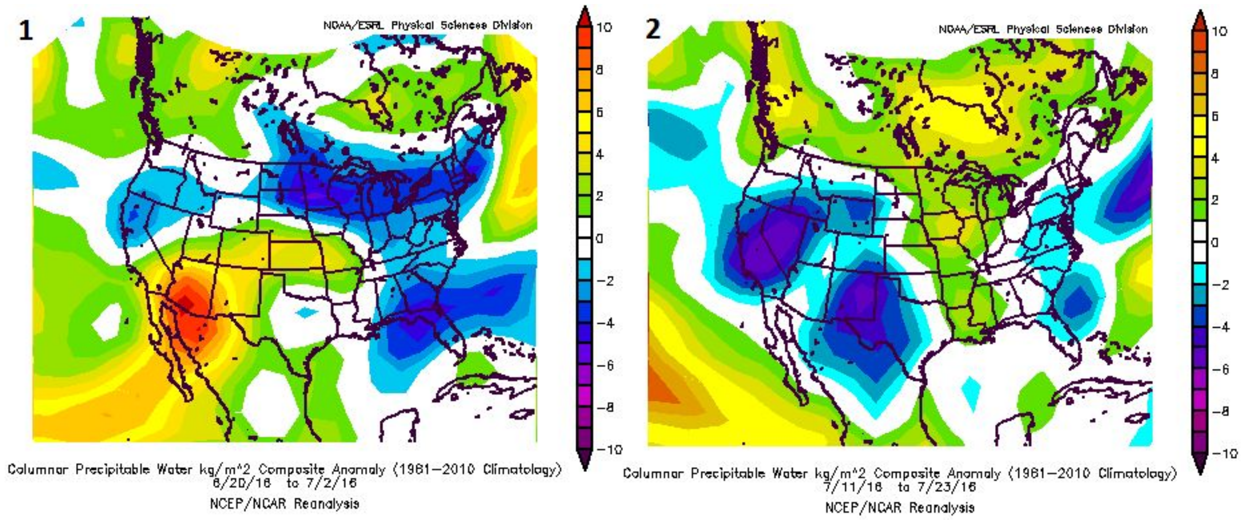


FIgure 17. *1000 - 500 hPa precipitable water composite anomalies for (1) the first half of FFaIR (June 20 - July 2) and (2) the second half of FFaIR (July 11 - 23). Images generated from the NCEP/NCAR Reanalysis provided by NOAA/ESRL/Physical Sciences Division (http://www.esrl.noaa.gov/psd/data/composites/day/).*

Table 4. *Experimental Day 1 ERO and 6-HR PFF forecasts issued during the 2016 FFaIR Experiment along with notable impacts.*

Forecast Valid End Date	Valid Time (UTC)	Forecast Area	Notes
21 June 2016	15 - 12	Central Mississippi River Valley and Colorado	
	18 - 00	No Forecast	
22 June 2016	15 - 12	Iowa/Illinois, Appalachians, Maryland/Delmarva	Flash flooding inundates a Washington Metro subway station during the evening of June 22.
	18 - 00	Maryland/Delaware/Northern Virginia and West Virginia/Kentucky	
23 June 2016	15 - 12	Lower Great Lakes region and Ohio River Valley	Numerous flash flood reports throughout Ohio including several reports of roads under water throughout Central Ohio.
	18 - 00	No Forecast	
24 June 2016	15 - 12	West Virginia and Nebraska/Eastern Colorado	Devastating flooding. **23 deaths**, over 1,200 homes destroyed
	18 - 00	West Virginia	
25 June 2016	15 - 12	West Virginia/Virginia/Northern North Carolina and Central Plains	

	18 - 00	Delmarva Peninsula	
28 June 2016	15 - 12	Southeast, Appalachians, and Western Texas	Flash flooding impacted several areas including Southwest West Virginia with reports of roads covered; downtown Birmingham, AL impacted with cars stranded, roads covered in high water; several reports also in Texas.
	18 - 00	West Virginia, Appalachians, Tennessee	
29 June 2016	15 - 12	New England/Maine and Central Plains	Multiple flash flood reports in New England including water over roadways in Maine and culverts washed out in Northern Vermont
	18 - 00	New England/Maine	
30 June 2016	15 - 12	Southwest, Central Plains, Southeast, and New England	
	18 - 00	Maine	
1 July 2016	15 - 12	Eastern Colorado/Central Plains and Arizona	Numerous flash flooding reports in and around downtown Las Vegas. **One death**, numerous swift water rescues and road closures due to high water.
	18 - 00	No Forecast	
2 July 2016	15 - 12	Southwest, Central Plains/Kansas, and New England	Tucson, AZ breaks 118-year old daily rainfall record receiving .82 inches of rain officially, while other areas saw over 2 inches. Several roads flooded and some motorists stranded.
	18 - 00	Southwest	
12 July 2016	15 - 12	Northern Plains/Minnesota/Wisconsin, Southeast and Carolinas	Up to 10 inches of rain fell. Numerous flash flood reports with roads washed out and water rescues in Minnesota and Wisconsin. **Two reported deaths**.
	18 - 00	Eastern Dakotas and Minnesota	
13 July 2016	15 - 12	Central Plains, Great Lakes Region, Appalachians	
	18 - 00	West Virginia/Virginia/North Carolina and Indiana	
14 July 2016	15 - 12	Missouri/Illinois and Eastern Pennsylvania/Maryland	Rt. 30 in Lancaster, PA shut down due to high water, numerous water rescues. Several campers had to be evacuated from flooded campsite in SW Missouri
	18 - 00	Missouri/Illinois/Iowa	
15 July 2016	15 - 12	South central Plains, Ohio River Valley, New England	State of Emergency due to flooding declared in Huntington, WV. Numerous homes flooded. **One death** reported when car washed away.
	18 - 00	New England	
16 July 2016	15 - 12	South central Plains and Southeast	
	18 - 00	Southwestern Kansas	
19 July 2016	15 - 12	New England, Ohio River Valley, Iowa, and Southwest	

	18 - 00	Missouri/Iowa/Nebraska border and Southeastern Illinois	
20 July 2016	15 - 12	Northern Plains/Iowa and Southwest	Flash flooding reported in numerous parks and roads throughout the Des Moines, IA metro area
	18 - 00	Iowa	
21 July 2016	15 - 12	Southwest and Northern/Central Mississippi River Valley	
	18 - 00	Arkansas, Illinois, Kentucky	
22 July 2016	15 - 12	Southwest, Northern Great Lakes/Wisconsin, and Georgia	Numerous flash flood reports in Madison, WI. Several intersections in downtown Madison impassable.
	18 - 00	No Forecast	
23 July 2016	15 - 12	Four Corners region, Northern Plains, and Ohio River Valley	
	18 - 00	Southeast	

4. ATMOSPHERIC GUIDANCE RESULTS

Performance of the Deterministic Convection-Allowing Models (CAMs)

Convective Initiation

Using event-driven archived radar data over a limited domain of the CONUS, participants were asked to subjectively rate the proficiency of convective initiation and storm evolution/structure shown in the composite reflectivity field from the ESRL HRRRv2, NAMRR, High-Res Window NMMB and ARW, and the NSSL-WRF models. The evaluation was to focus on the ability of the model to efficiently "spin up," or match the radar progression of the convective event. Results were mixed, with participants noting that the CAMs still need development work to be fully trusted by forecasters, and that performance was often better when synoptic forcings were stronger. These deterministic high-resolution models struggled more with marginal events and scattered,weakly-forced convection. Figure 18 shows the average scores for the CAM convective initiation over all 4 weeks of the experiment. The ESRL HRRRv2 had a slight edge over the rest of the CAMs with an average score of 5.89 out of 10. The NAMRR and the NSSL-WRF were nearly tied with scores of 5.19 and 5.12 out of 10, respectively. The high-res window ARW (4.34 out of 10) and NMMB (4.39 out of 10) scored the lowest overall.

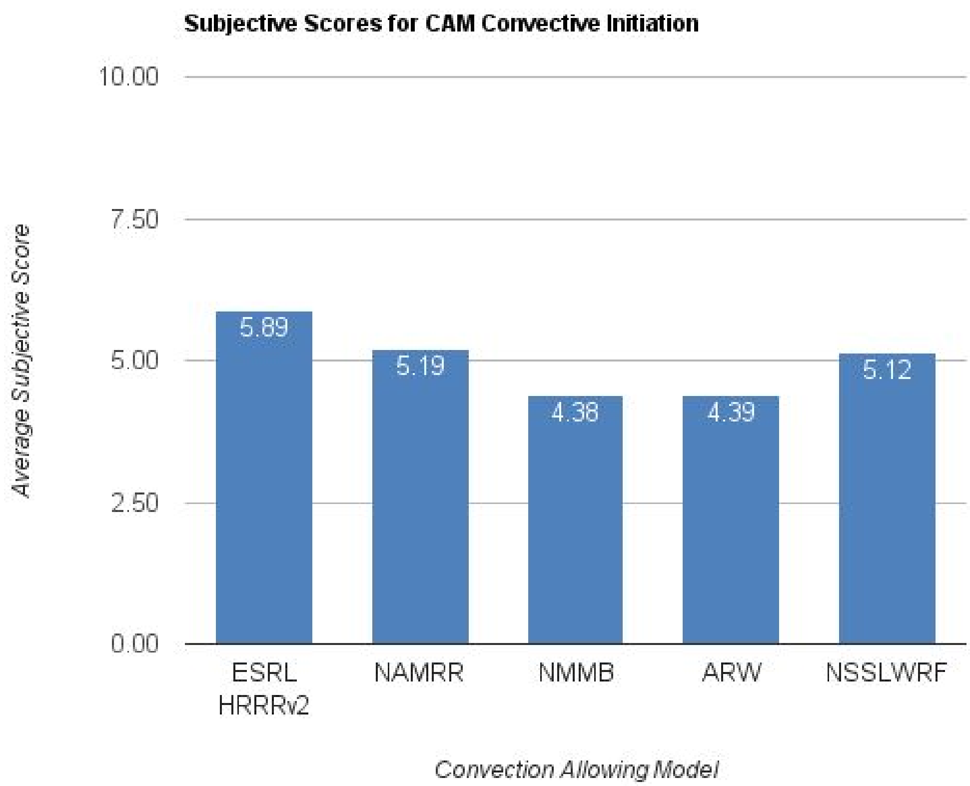


Figure 18. *Average subjective scores by participants when evaluating the proficiency of CAM convective initiation.*

The CAMs universally struggled with producing secondary lines, or “re-firing” of convection that occurred after the main event. In theory, a factor may be cold pools that dominate too much in the CAMs after the initial convection, which results in a more stable air mass and inhibits re-firing. Each CAM traded instances in which they were underdone, too aggressive, too slow to spin up, had incorrect placement and structure, or dissipated the convection too quickly. Differences in the initializations and data assimilations made systematic biases difficult to identify. Given the variability in performance from day to day, it was hard to trust any one solution. The general forecaster distrust in the guidance tended to have a negative impact on the resulting probabilistic forecast.

Although the CAMs scored similarly within a small margin on average, it can be noted in the individual scores per day (Figure 19) that 60% (12 out of 20) of the NAMRR scores were 6 and 7. The ESRL HRRRv2 achieved a score of 8 three different times, which was the highest overall, and scored at least a 6 60% (9 out of 15) of the time.

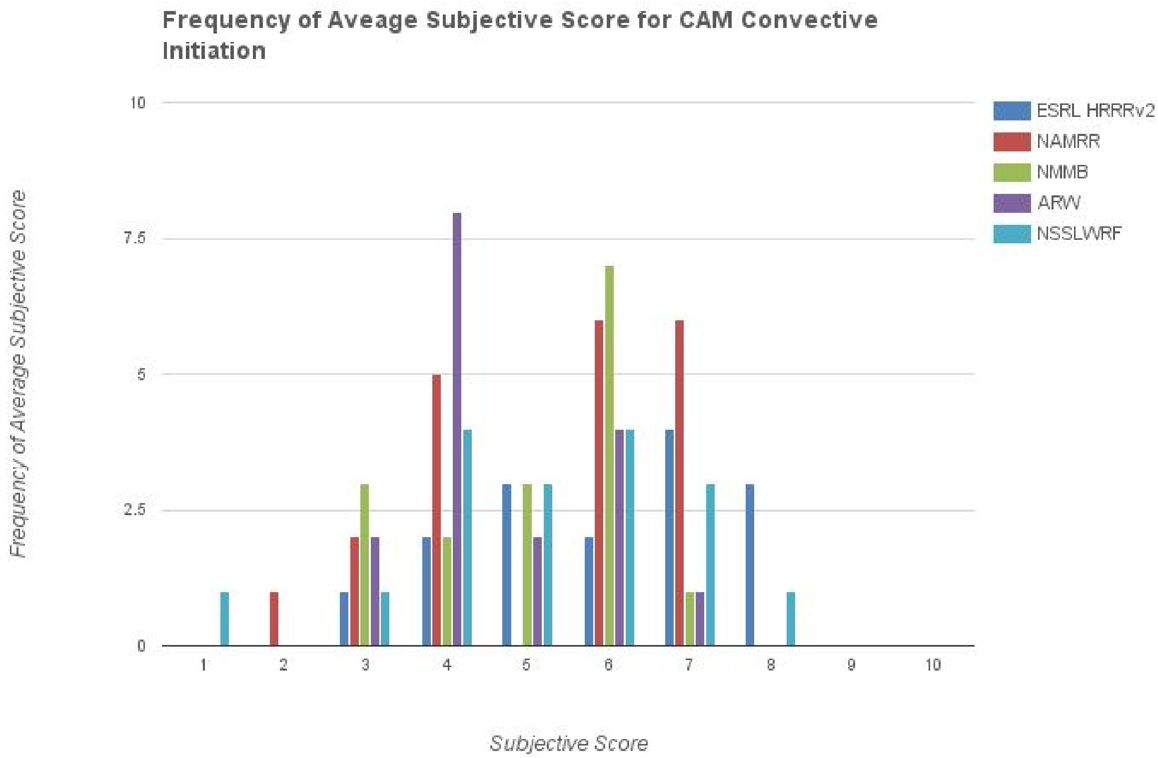


Figure 19. *The frequency of the averaged subjective scores during the evaluation of the efficiency of CAM convective initiation or "spin-up."*

6-hour QPF

The skill of 6-hour QPF from deterministic CAMs was also evaluated during the FFaIR Experiment. Using MRMS QPE as verification, participants were asked to perform a visual comparison and subjectively rate the 6-hour (18Z - 00Z) QPF forecasts from the ESRL HRRRv2, NCEP HRRRv2, WRF-NSSL, NAMRR and NAM CONUS-Nest (CONEST) (Figure 20) on a scale from 1 to 10.

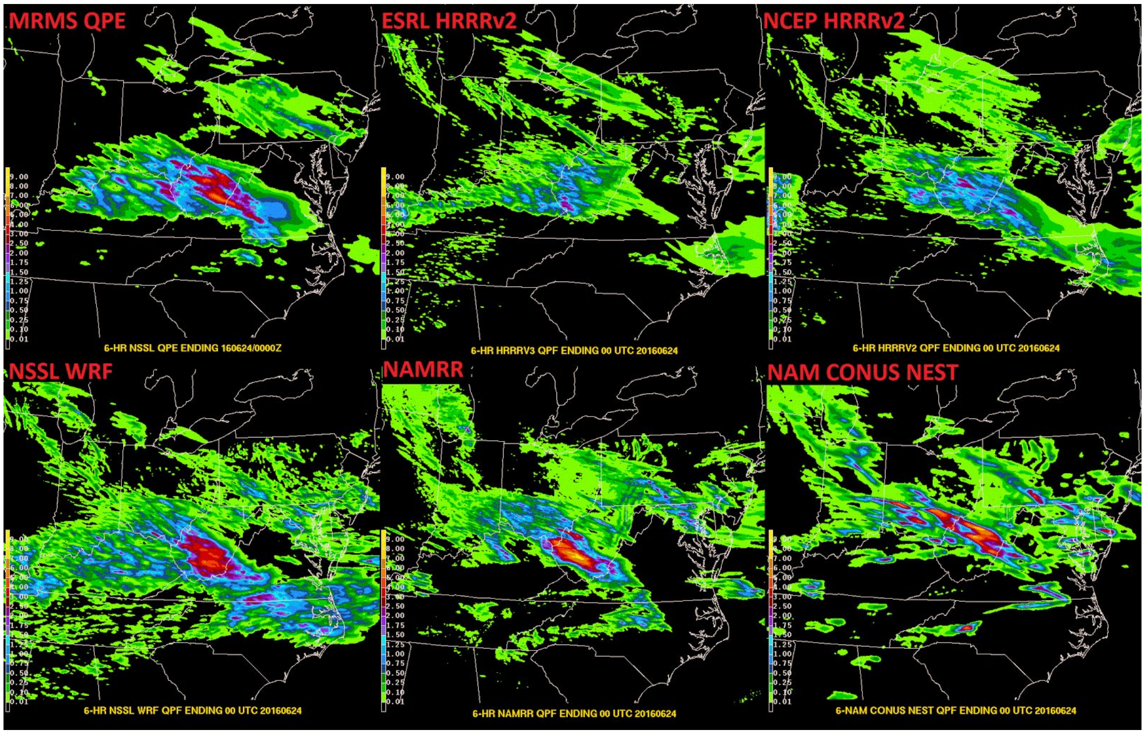


Figure 20*. An example of the 6-hr MRMS QPE (top left) presented alongside the deterministic CAM 6-hr QPF for subjective evaluation.*

Similar to the convective initiation, the QPF from the CAMs most often captured the general spatial extent and magnitude of a precipitation event, but differed in the details. The participants attempted to capture these differences in the average overall scores of the 6-hour QPF shown in Figure 21.

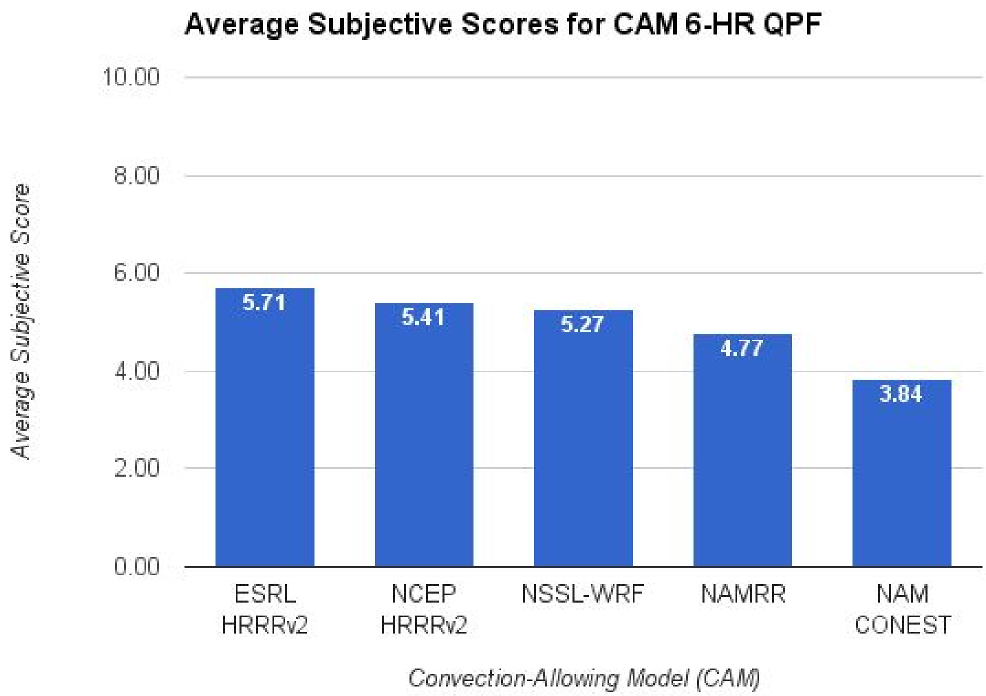


Figure 21. *Average subjective scores by participants when evaluating the 6-hour QPF.*

The ESRL HRRRv2 rated the highest for areal extent and magnitude of precipitation with an average score of 5.71 out of 10. The ESRL HRRRv2 demonstrated marginal improvement over the NCEP HRRRv2, which scored an average of 5.41 out of 10. Participants commented that the ESRL HRRRv2 and the NCEP HRRRv2 often produced too low a magnitude of precipitation, but best captured the spatial coverage and main axis of QPF events.

The NAMRR was much improved over the NAM CONEST, but conversely, often produced too high a magnitude of precipitation. The participants still appreciated the general spatial coverage and structure of the NAMRR QPF, which received an overall score of 4.77 out of 10, and would “mentally bring down” the magnitude of the precipitation, as it was described as a “known wet bias.” The CONEST was the least favored model with an average score of 3.84 out of 10, due to its significant wet bias and struggle to capture the spatial extent and structure needed to be useful guidance.

The NSSL-WRF performed consistently well in both magnitude and spatial extent. The participants found the NSSL-WRF QPF to be useful guidance and gave it an average subjective rating of 5.27 out of 10.

Ensemble Performance During 2016 FFaIR Experiment

Comparing Conventional Ensemble Means and Probability Matched Means

With the emergence of numerous different ensemble systems, determining the best method for displaying the myriad of data they provide is essential. During the 2016 FFaIR Experiment, the conventional ensemble mean (CM) and probability matched mean (PM) from two high resolution, convection allowing ensembles were examined. Conventional ensemble mean is simply averaging all the ensemble members' solutions over a point. Probability matched mean combines the spatial pattern of the ensemble mean QPF with the frequency distribution of the rainfall rates to provide a more realistic ensemble rainfall intensity forecast (Ebert 2001).

To evaluate these two different mean calculations, participants were shown 6-hour CM and PM output valid from 18-00Z over a limited domain from the HREFX and SSEFX during verification. MRMS QPE was used as observation. Participants provided a subjective ranking from 1 (very poor) to 10 (very good) as well as commented on how the two means differed from each other and if they thought one was more useful than the other. Figure 22 shows an example of how the evaluation was presented to the participants.

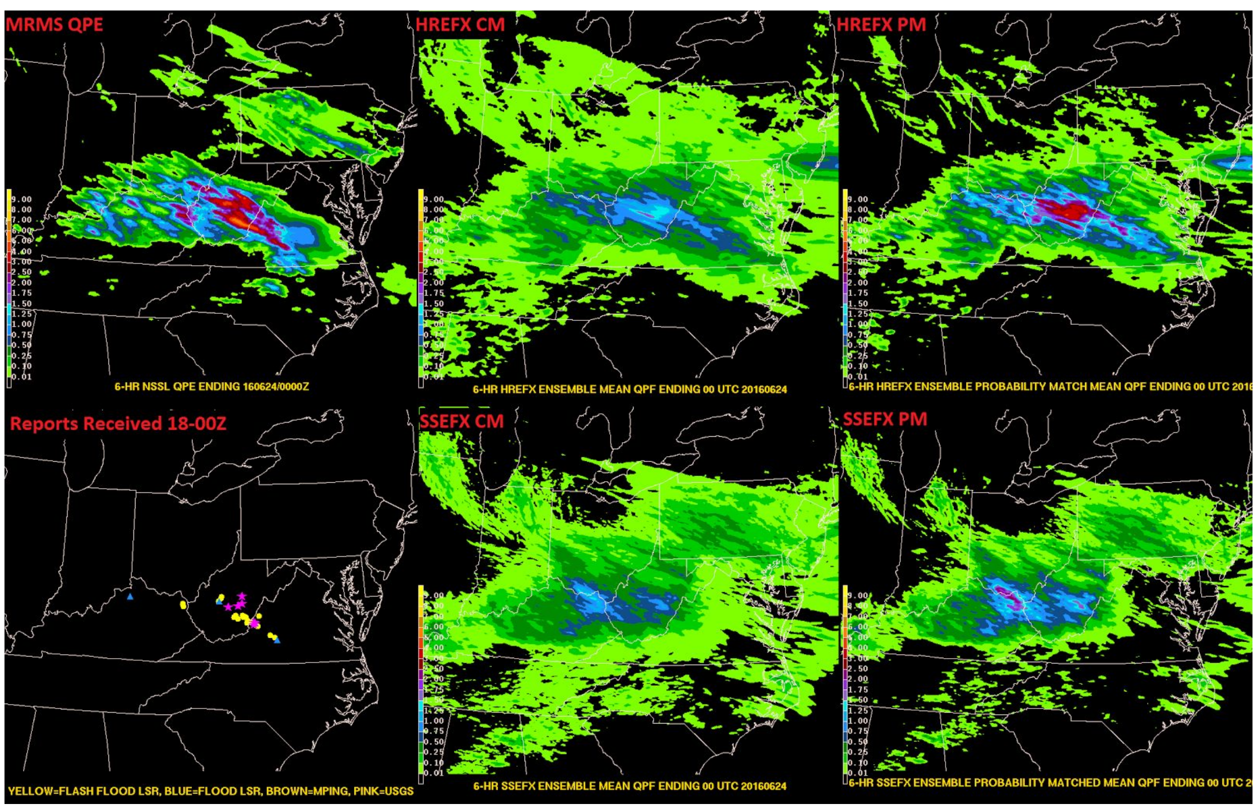


Figure 22. *Six panel display of MRMS QPE, HREFX CM, HREFX PM, Reports, SSEFX CM, and SSEFX PM for the 6-hour period valid 18Z June 23 - 00Z June 24.*

Figure 23 shows the average subjective scores each model mean collected over the course of the experiment and Figure 24 shows the distribution of the individual scores. When comparing the two individual ensemble systems, the SSEFX scored just slightly higher overall than the HREFX in average subjective scores. In examination of the mean calculation scores, both ensembles PM scored higher than the CM. The graph depicting the distribution of all the scores over the course of the experiment show the large clustering of scores around 5 and 6. The SSEFX PM and HREFX PM each scored the highest marks with two 8s and one 8 respectively during the experiment. Both also scored a seven a few times, along with the SSEFX CM. On the other end of the spectrum, the SSEFX PM and SSEFX CM were each subjectively scored a one once, the two lowest scores during the experiment.

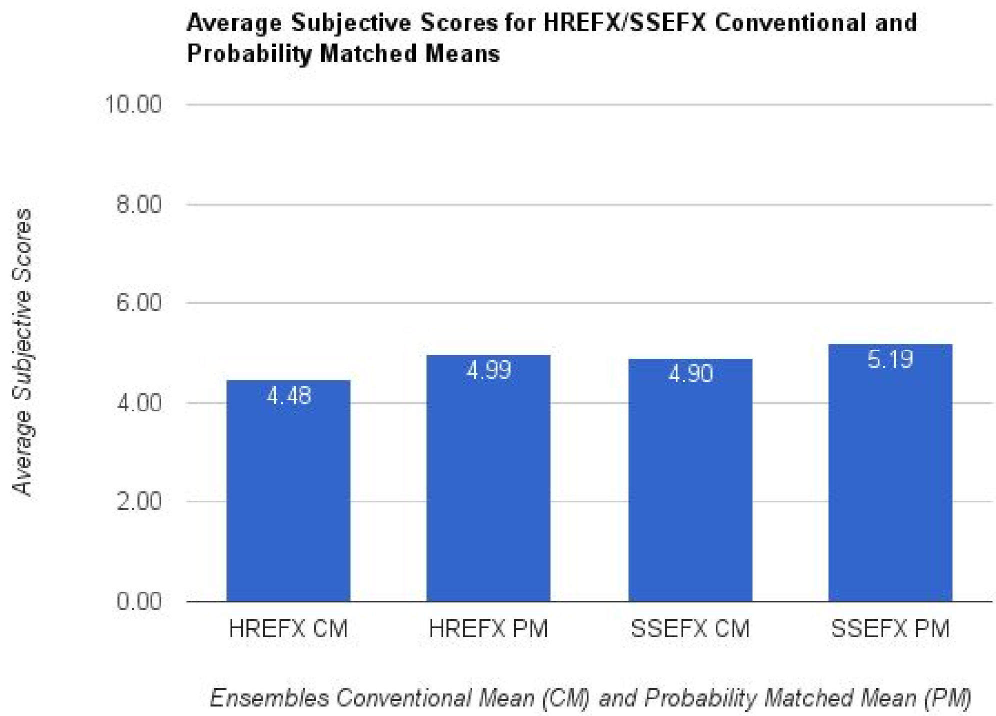


Figure 23. *Average subjective score of each model and mean calculation as determined by the participants in FFaIR on a scale of 1 (worst) to 10 (best).*

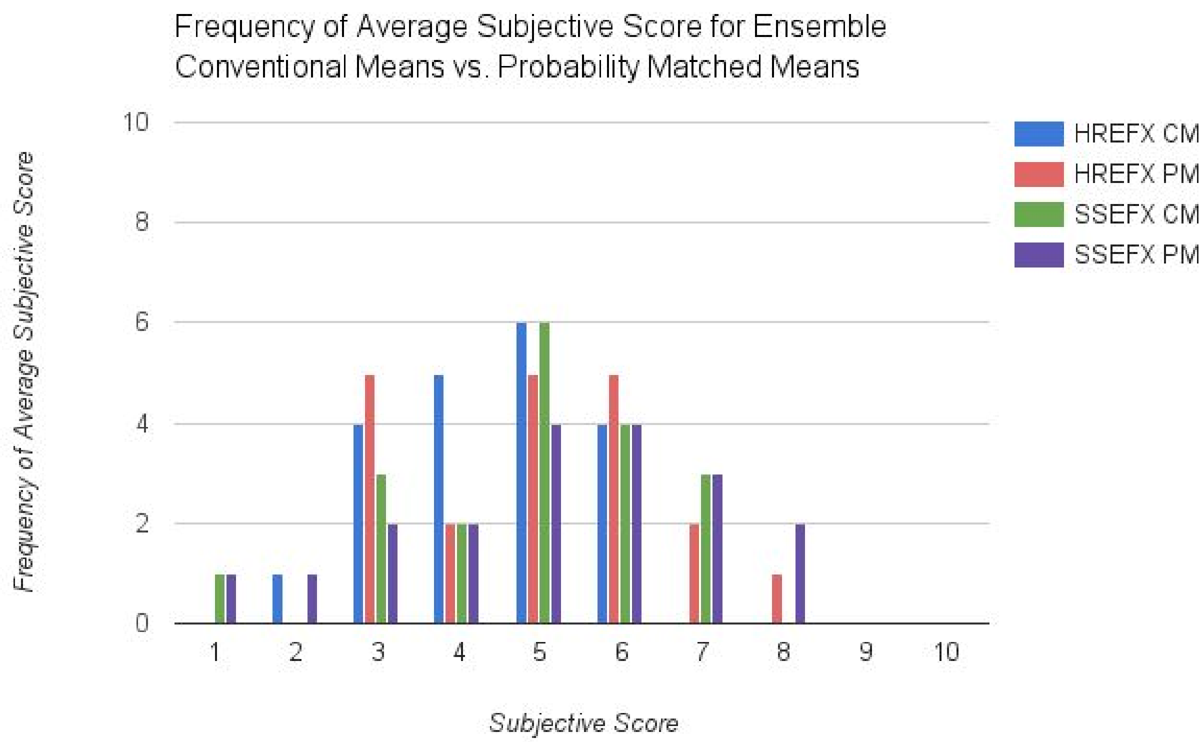


***Figure 24**. Frequency of all subjective scores (1-10) for both models as determined by the participants over the course of the 2016 FFaIR experiment.*

As the scores reflect, feedback from the participants indicate that the probability matched mean method was preferred over the conventional ensemble mean. Figure 25 shows an example in which the HREFX PM performed exceptionally by capturing the heavy rainfall event in West Virginia, whereas the HREFX CM was much too light. Frequent commentary over the weeks stated that the PM versions of both models did well with increasing the magnitude of the precipitation and scouring out lighter precipitation that did not verify. Conversely, Figure 26 shows an example where the PM versions of the models increased the amounts too much and distracted the participants. In these cases, several lighter or more marginal events, participants did choose the CM version over the PM version. Some participants noted, however, if the PM versions had a higher bias than the CM versions consistently and still did well with placement, it was still preferred over a consistently underdone solution. Overall feedback was very positive in regards to both high-resolution ensemble systems and the probability matched mean method was preferred over the conventional mean when calculating QPF in most cases.

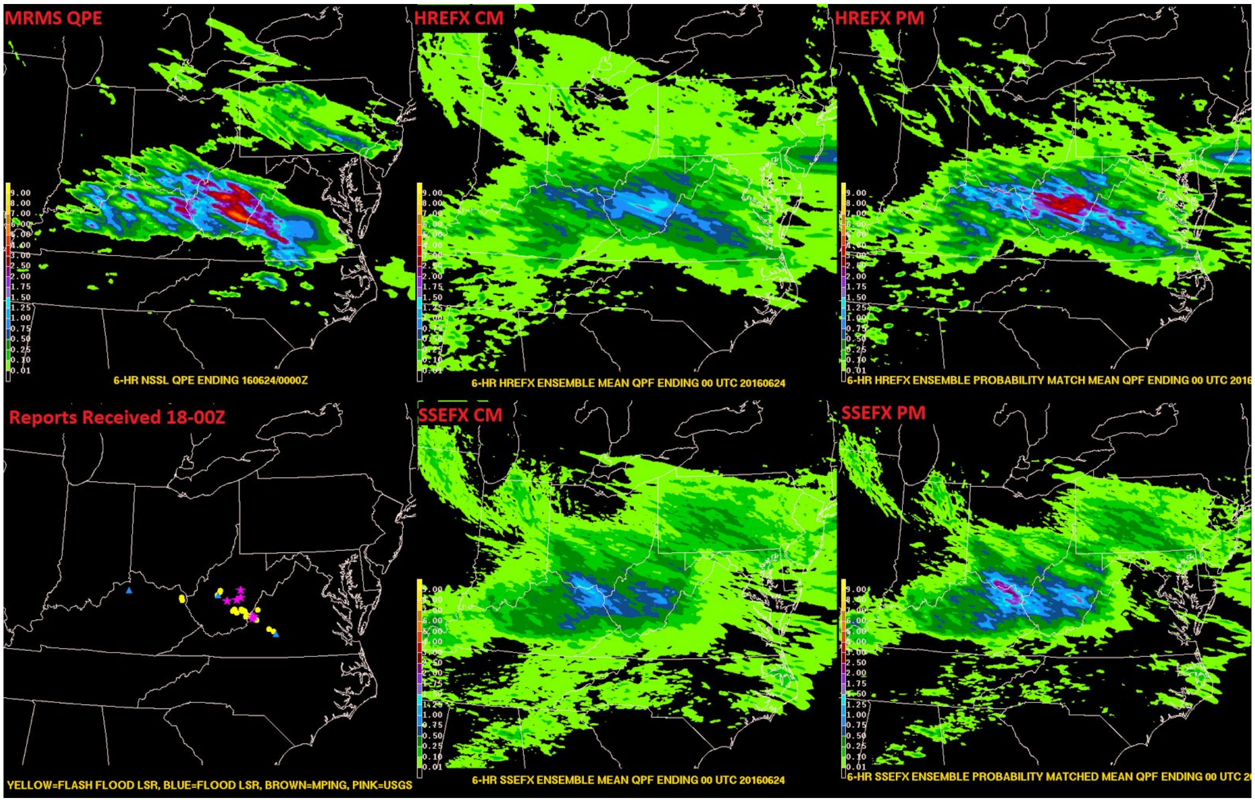


***Figure 25**. Example of MRMS QPE, HREFX CM, HREFX PM, Reports, SSEFX CM, and SSEFX PM for the 6-hour period valid 18Z June 23 - 00Z June 24.*

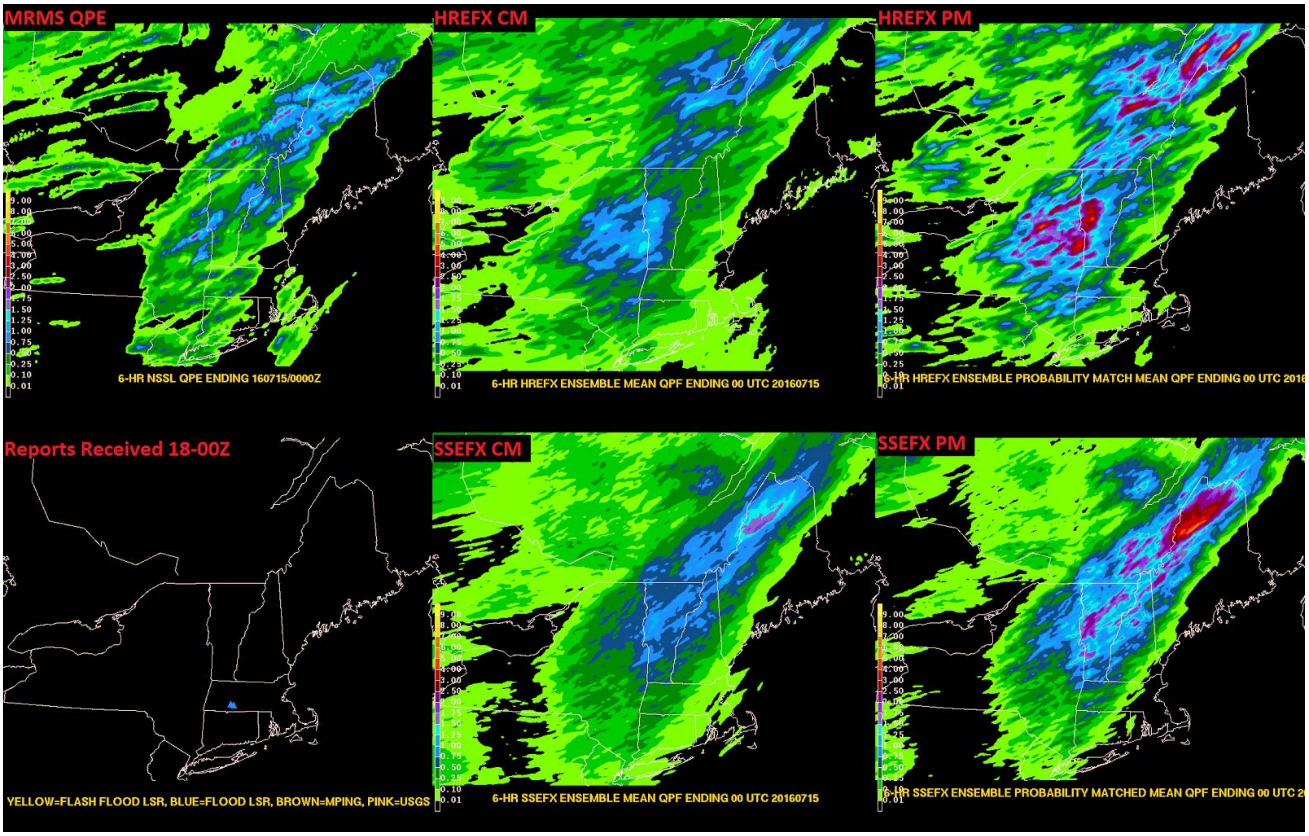


***Figure 26**. Example of MRMS QPE, HREFX CM, HREFX PM, Reports, SSEFX CM, and SSEFX PM for the 6-hour period valid 18Z July 14 - 00Z July 15.*

Calibrated One-Degree GEFSX

The calibrated one-degree GEFSX was subjectively evaluated for its utility as probabilistic QPF guidance, specifically the probability of equalling or exceeding one and two inches of QPF in 24 hours. Participant comments were collected each day as the GEFSX probabilities were typically used daily as guidance, especially for the Day 2 ERO. Figure 27 displays the distribution of the subjective rankings for each probabilistic threshold that was scored through the experiment. The probability of one inch of QPF consistently scored better than the two inch probability. The average subjective scores for the probability of one inch and two inches of QPF were 4.53 and 3.60 out of 10 respectively. The coarse one-degree ensemble struggled with mesoscale, convective events often missing areas of precipitation or underperforming. However, in larger-scale, strongly forced events, the calibrated GEFSX performed quite well. Figure 28 is an example of a good case where there was a strongly-forced MCS that moved through the Ohio River Valley over the period of 12Z June 22 - 12Z June 23. This run of the GEFSX suffered similar problems as the operational GEFS, such as a dry bias in some areas. It was also noted that in several cases throughout the experiment, the GEFSX often displaced the heaviest precipitation to the north or northeast of where it actually occurred. Finally, participants felt that the calibration technique employed in this experimental version did well with removing false high-end amounts as well as eliminating over-estimated precipitation around the periphery of the main area of precipitation.

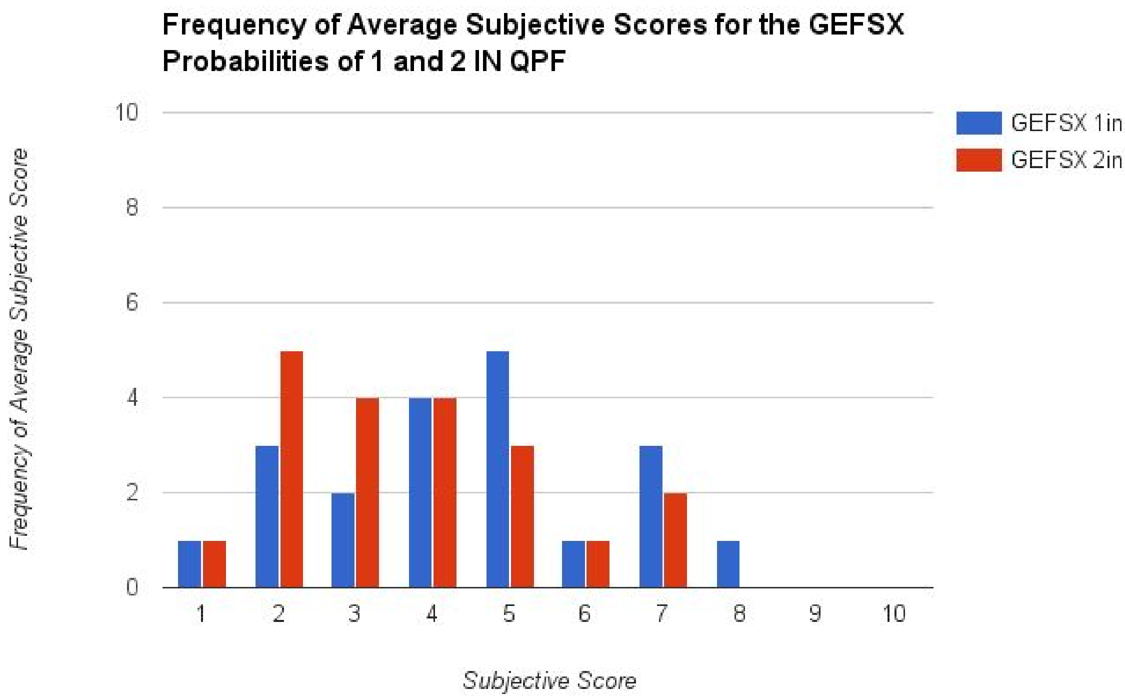


***Figure 27**. Frequency of all subjective scores (1-10) for the GEFSX probability of 1 in of QPF (blue) and probability of 2 inches of QPF (red) as determined by the participants over the course of the 2016 FFaIR Experiment.*

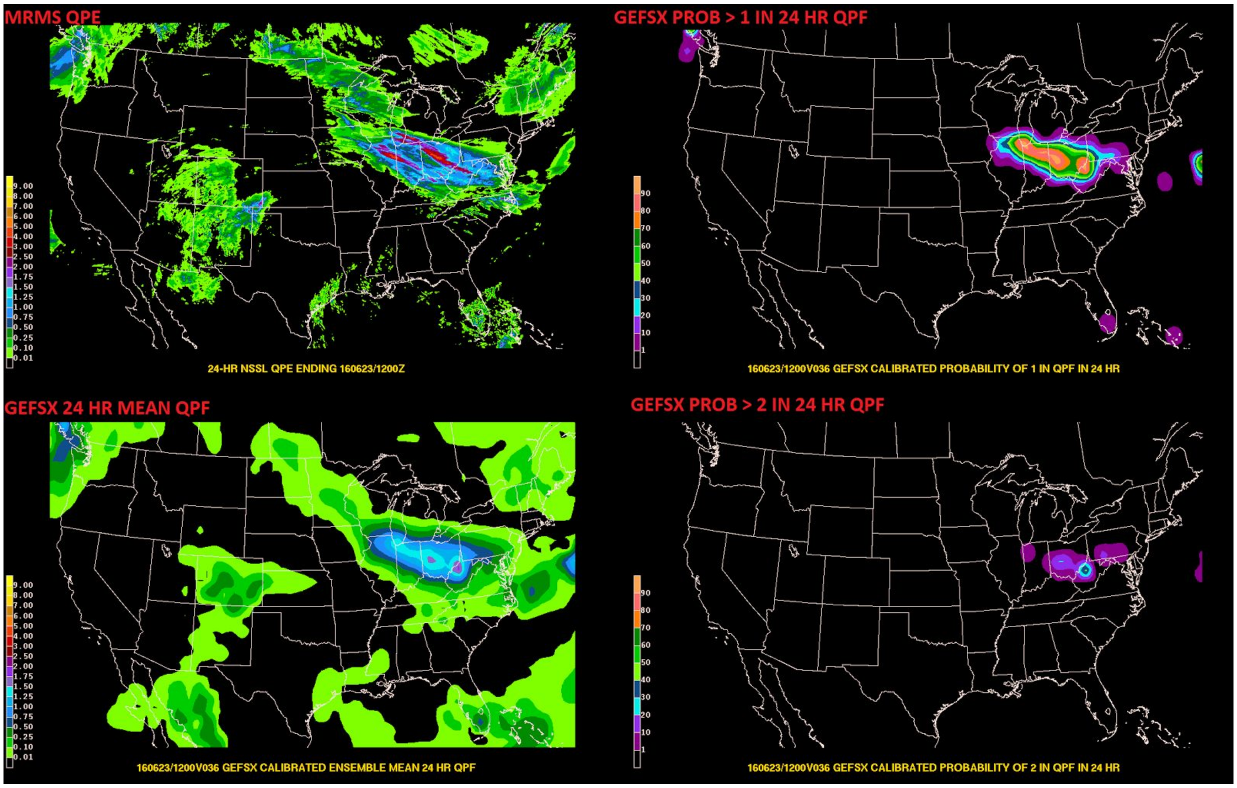


Figure 28. *MRMS QPE, GEFSX probability of 1 inch of 24 HR QPF, GEFSX 24 HR ensemble mean QPF, and GEFSX probability of 2 inches of 24 HR QPF valid from 12Z June 22 - 12Z June 23, 2016.*

Day 2 Model Guidance Performance

During the 2016 FFaIR Experiment, the generation of an experimental Day 2 Excessive Rainfall Outlook (ERO) was introduced. The experimental Day 2 ERO in FFaIR was defined as the same neighborhood probability of a flash flood/flood report as the Day 1 ERO while exploring the use of experimental and operational guidance in this longer-range forecasting process.

Guidance included a variety of ensembles including the operational GEFS and SREF, the Ensemble Bias Corrected (ENSBC, an in-house WPC developed ensemble that uses a large number of operational models for members and performs bias-correction) as well as the high resolution, convection-allowing SSEFX ensemble and the NAMRR. Lastly, the medium- range forecast from the National Water Model (NWM) was evaluated in which both the streamflow anomaly and soil moisture products are forced by the QPF from the operational GFS. Figure 29 shows the average subjective score of each model received throughout FFaIR and Figure 30 displays the frequency of the subjective scores assigned to each model over the four week experiment by the participants.

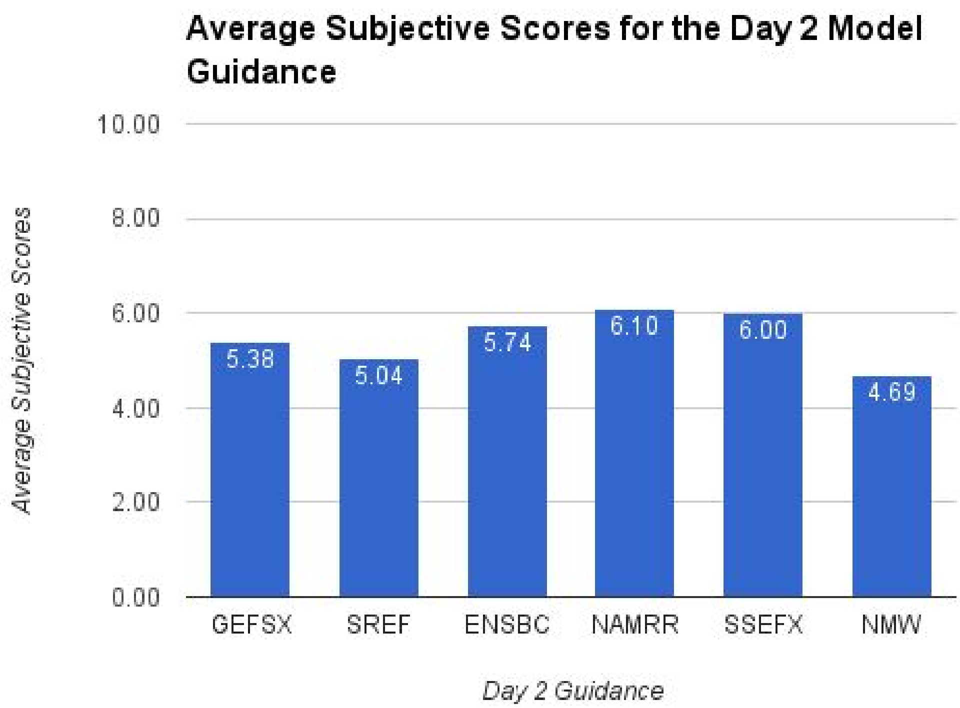


Figure 29. *Average subjective score of each model and mean calculation as determined by the participants in FFaIR on a scale of 1 (worst) to 10 (best).*

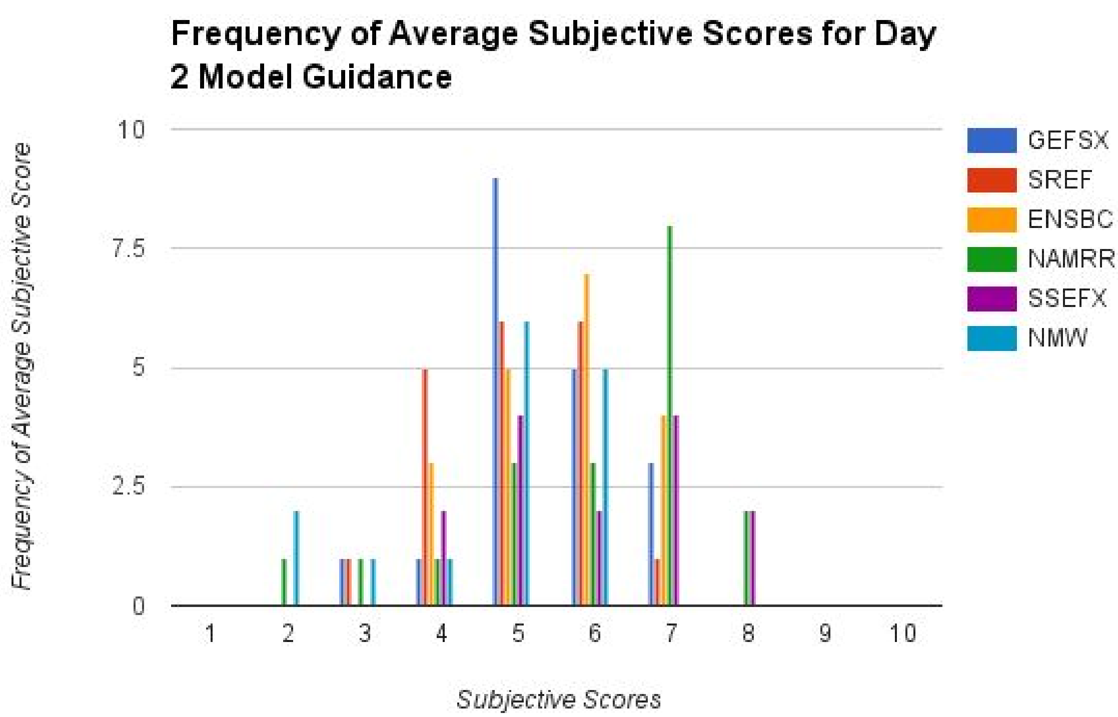


Figure 30*. Frequency of all subjective scores (1-10) for both models as determined by the participants over the course of the whole experiment.*

NAMRR

The NAMRR was the only high resolution, deterministic CAM available as Day 2 guidance. It performed the best according to the subjective score averages with a score of 6.10 out of 10. In fact, it scored either a 7 or an 8 for 52% of the days. Participants were often enthusiastic about the performance of a high resolution CAM that went out 60 hours and could be used for a Day 2 forecast. Figure 31 shows a case of the NAMRR subjectively scoring quite well. For a 60-hour forecast, it did well in capturing the areas of precipitation in the Southwest, Midwest, and Southeast regions of the U.S. There were some cases where smaller scale events were underdone or completely missed by the NAMRR. Figure 32 shows one of these examples from June 20-21 where it failed to pick up on the small scale convective system that formed over the Central Plains. Despite this, most participants agreed that while the magnitude and areal coverage of the QPF was not always perfect, the model guidance provided great value when developing a Day 2 ERO forecast.

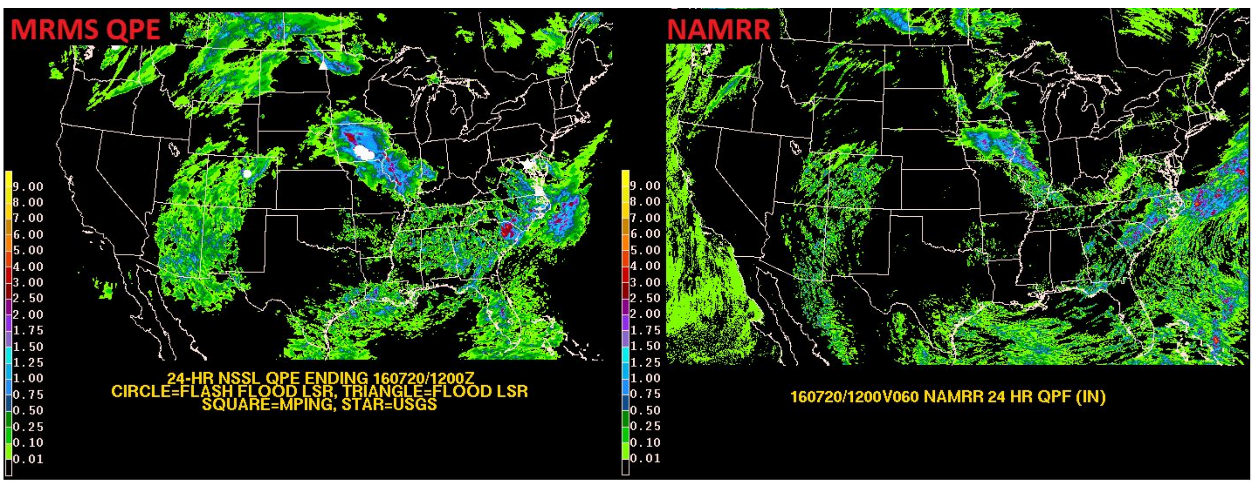


***Figure 31**. 24 HR MRMS QPE valid 12Z July 19 - 12Z July 20 and the 60 hour forecast of 24 HR QPF valid over the same time period from the NAMRR. Flooding reports appear as white symbols in the MRMS QPE plot.*

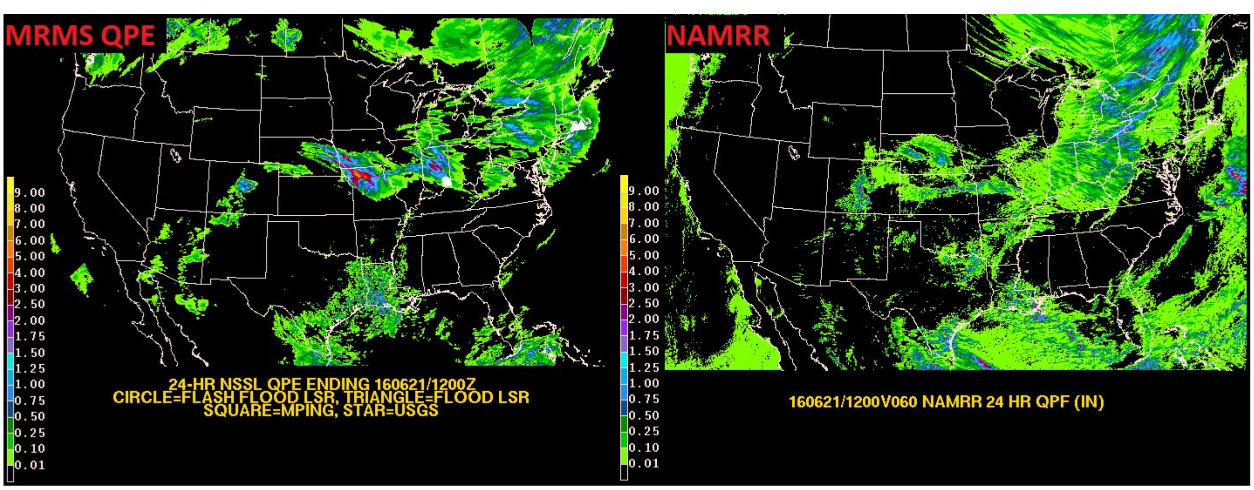


Figure 32. 24 HR MRMS QPE valid 12Z June 20 - 12Z June 21 and the 60 hour forecast of 24 HR QPF valid over the same time period from the NAMRR. Flooding reports appear as white symbols in the MRMS QPE plot.

SSEFX

The SSEFX provided by the OU/CAPS team and is a high-resolution, convection allowing ensemble with a forecast projection through 60 hours. In this evaluation, the 24-hour probability matched mean QPF was examined. Out of the six models, the SSEFX had an average subjective evaluation score of 6.00 out of 10. The frequency of the subjective scores of the SSEFX Day 2 forecast utility throughout the experiment are represented in Figure 30. It received a subjective score of a 7 or an 8 for 43% of the 14 scores. A majority of the comments were positive and participants preferred having a high resolution, convection allowing ensemble at their disposal when creating a Day 2 forecast. It received high praise for magnitude and areal coverage, however some days, perhaps due to the probability matched mean solution, the QPF magnitude was too high. The issues were relatively minor and overall participants felt like this model had great utility as Day 2 guidance. Figures 33 and 34 show examples of forecasts rated highly and less favorably, respectively.

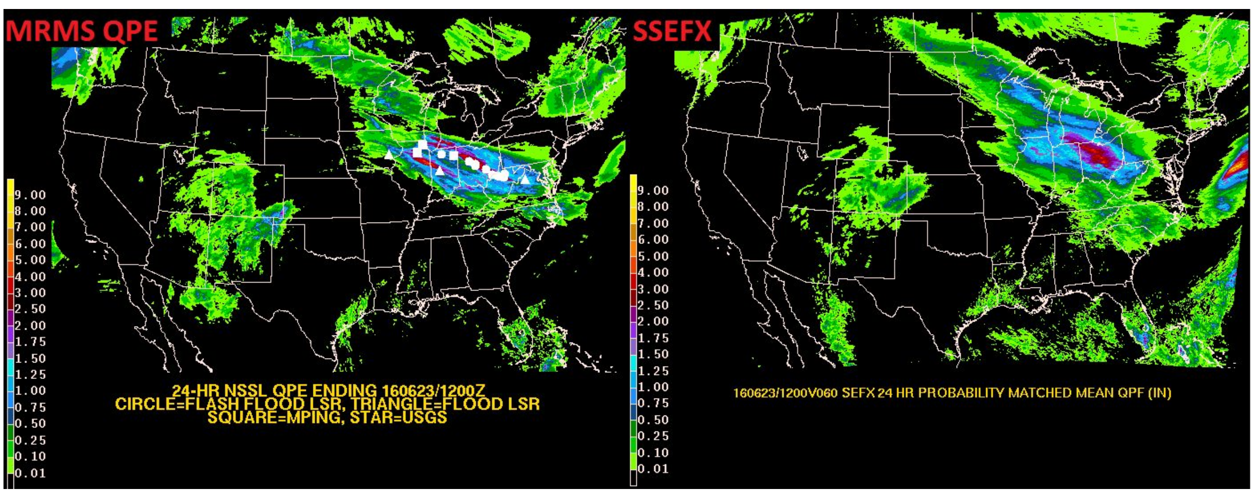


Figure 33. *24 HR MRMS QPE valid 12Z June 22 - 12Z June 23 and the 60 hour forecast of 24 HR probability matched mean QPF valid over the same time period from the SSEFX. Flooding reports appear as white symbols in the MRMS QPE plot.*

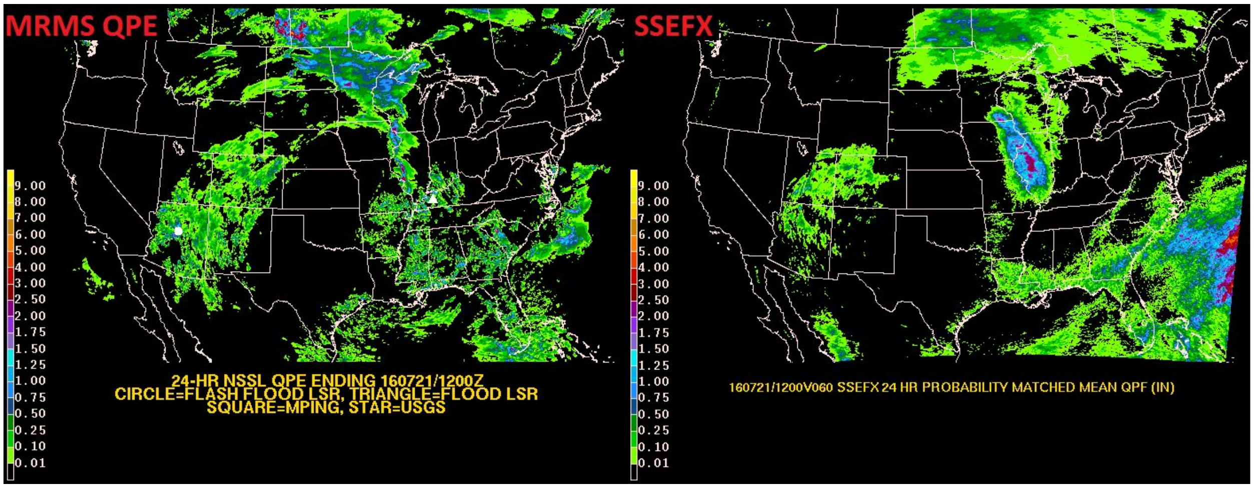


Figure 34. *24 HR MRMS QPE valid 12Z July 20 - 12Z July 21 and the 60 hour forecast of 24 HR probability matched mean QPF valid over the same time period from the SSEFX. Flooding reports appear as white symbols in the MRMS QPE plot.*

Other Ensemble Systems

The ENSBC was subjectively rated the highest at 5.74 out of 10, with 42% of the scores ranking 5 or below. It was followed by the GEFSX at 5.38 out of 10 with 58% of the scores ranking 5 or below. The operational SREF ranked the lowest at 5.04 out of 10, with 63% of the scores ranking 5 or below. All three ensemble means often had precipitation that was either too light or deficient in areal coverage. A few cases presented means that were too heavy in spots. The ENSBC did the best job of increasing some of the lighter amounts in its mean to make it more similar to verification. Despite some of these common problems, the participants stated most days that these systems would have some use as Day 2 guidance as a way of getting a more general overview before bringing in other, perhaps higher resolution guidance to continue investigating. Figure 35 shows an example of the three lower resolution ensemble systems.

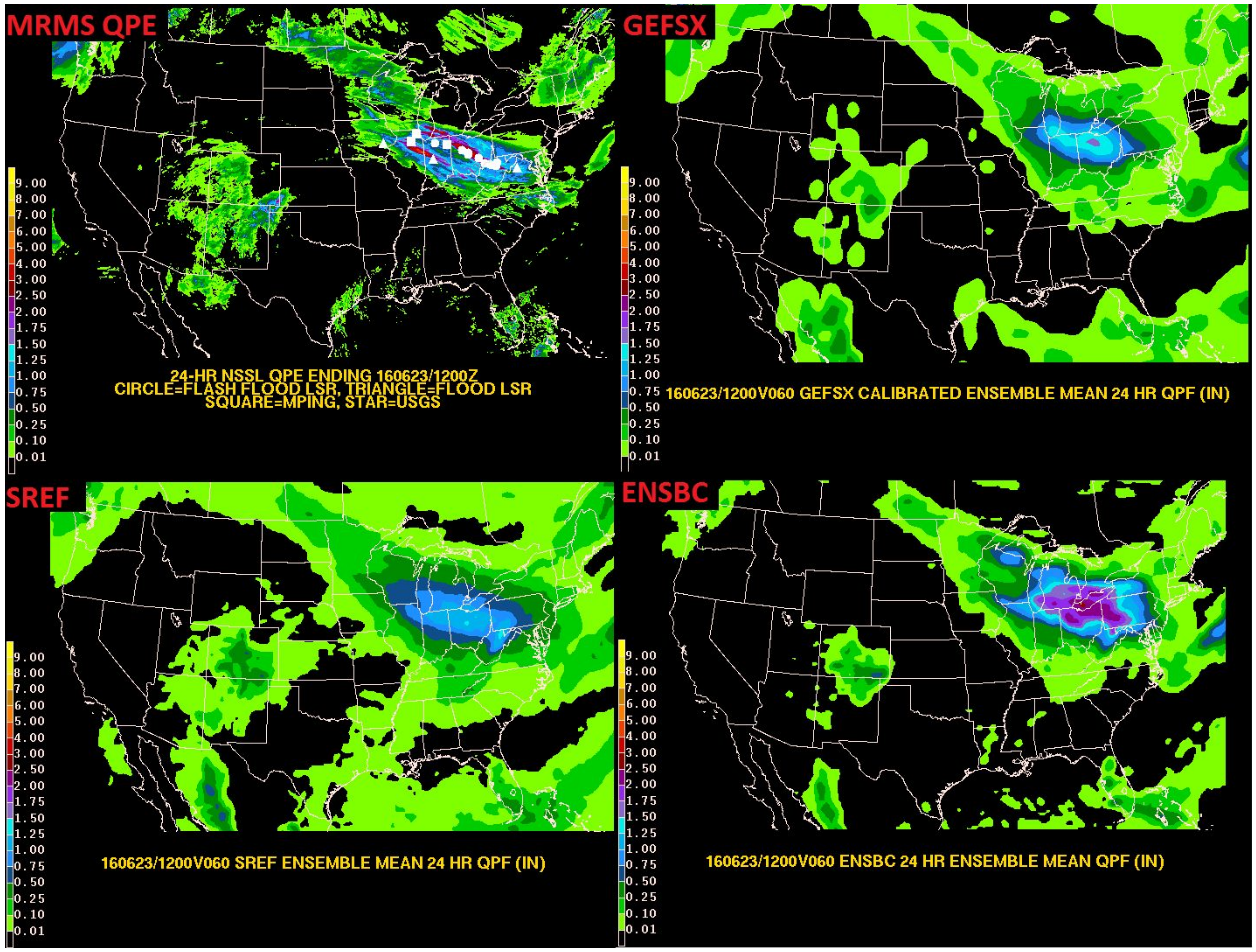


Figure 35. *24 HR MRMS QPE valid 12Z June 22 - 12Z June 23 and the 60 hour forecast of 24 HR ensemble mean QPF valid over the same time period from the GEFSX, SREF, and ENSBC. Flooding reports appear as white symbols in the MRMS QPE plot.*

National Water Model

The operational web-based National Water Model fields of streamflow anomalies and soil saturation provided by the Office of Water Prediction (OWP) were evaluated in the 2016 FFaIR Experiment. Streamflow anomalies are based on anomalies from monthly average flow, where the values are taken from the NHDPlus Enhanced Runoff Method (EROM) table. The NWM soil moisture product depicts fraction of soil saturation in the upper soils of the NWM (i.e. 1-40 cm). During subjective evaluation, participants looked at these two products from the medium-range version of the NWM (QPF forced by the GFS) in comparison to the MRMS QPE and flash flood reports to try to begin associating streamflow anomalies and soil moisture to flash flood events and develop patterns. The short-range NWM forecast is forced by the operational HRRR QPF and utilized during forecasting exercises but not formally evaluated in verification.

Participants noted that the data provided by the streamflow anomaly graphics were very raw and lacking context. Whereas potential for this product to be useful was recognized, forecasters desired overlays such as QPF, streamflow gauge data, inundation or other visual to show that a stream was expected to overflow its banks. Many times the forecast called for streamflow anomalies of “high flow,” which is the highest threshold, but uncertainty remained for whether or not the river or stream would actually leave its banks and flood or remain within its banks. Figure 36 shows streamflow anomaly output from the short-range NWM forecast for the West Virginia flood event on June 23. In the circled region a large area of rivers and streams are at high flow or much above normal flow in the affected areas. Without this distinction, the NWM forecasts served more as a situational awareness tool rather than being effectively leveraged as data applicable to a flash flood forecast. Trust in the medium-range NWM guidance was noted with the prior knowledge that the GFS QPF forcings may not have performed well over a particular region. Participants also found it difficult to focus in on areas due to the expansive network of rivers and streams that were represented on the CONUS wide map. Some days very tiny streams regionally would briefly flash to high flow, but this would be very difficult to see if trying to forecast in real time.

The NWM products were rated 4.69 out of 10, the lowest rating for potential Day 2 ERO guidance, reiterating that forecasters feel the NWM products are not ready to be used for flash flood forecasting. In spite of all this, participants still were highly supportive of viewing this type of data for the first time and enthusiastic about the future evolution of these products.

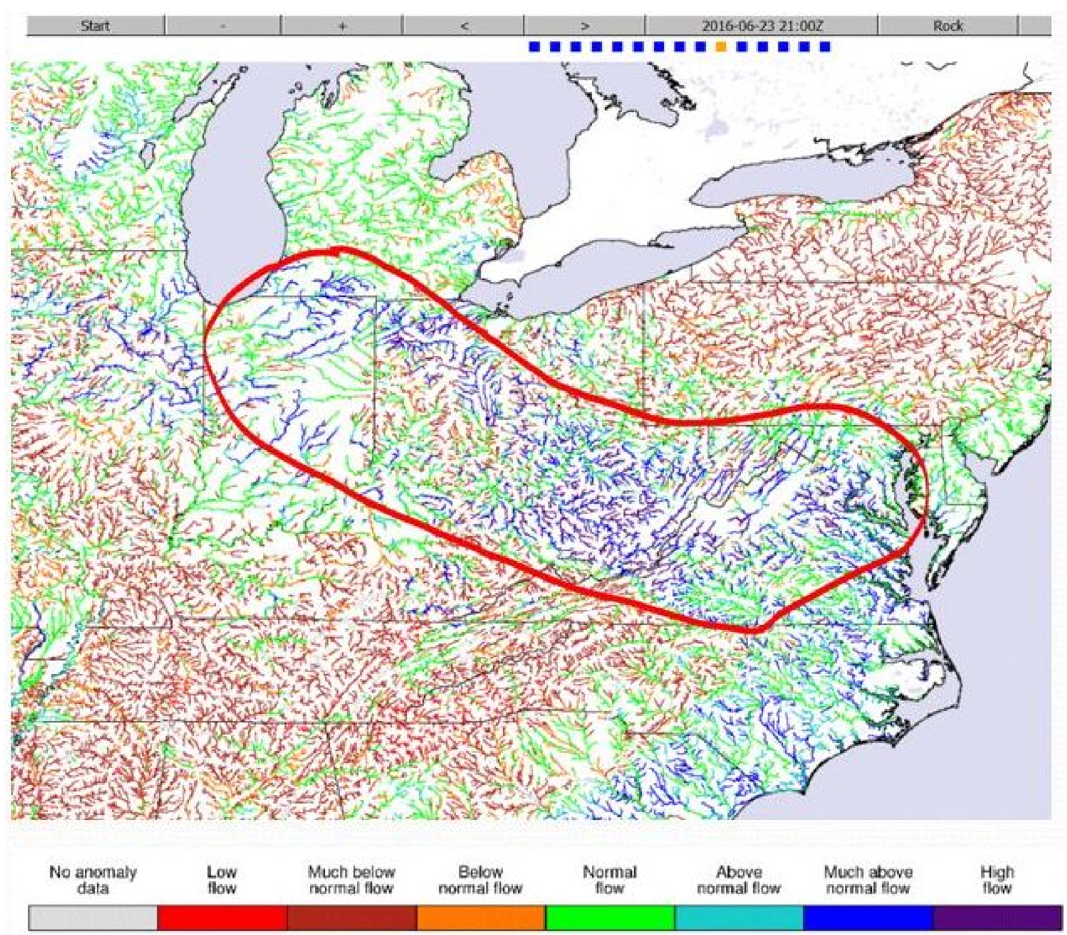


***Figure 36**. Short-range National Water Model streamflow anomaly output valid on 21Z June 23, 2016.*

4. HYDROLOGIC GUIDANCE RESULTS

Storm Surface Runoff

For the 2016 FFaIR Experiment, we examined the soil moisture availability and storm surface runoff fields available from the Land Surface Model (LSM) and QPF of the ESRL HRRRv2, NAMRR and National Water models. Probabilistic storm surface runoff probabilities were also available from the HRRR time-lagged ensemble (HRRR-TLE). These fields were explored as a supplement to Flash Flood Guidance for assessment of antecedent conditions and local flood vulnerabilities.

The 3-hour storm surface runoff from the ESRL HRRRv2 and the NAMRR are instantaneous fields derived from the LSM and QPF. The probability of 3-hour runoff exceeding a half inch, one inch or two inches from the HRRR-TLE draw upon three previous runs of the ESRL HRRRv2. All three runoff products were evaluated against the 3-hour MRMS QPE and reports of flash flooding and flooding (Figure 37), since there were no true runoff verification sources.

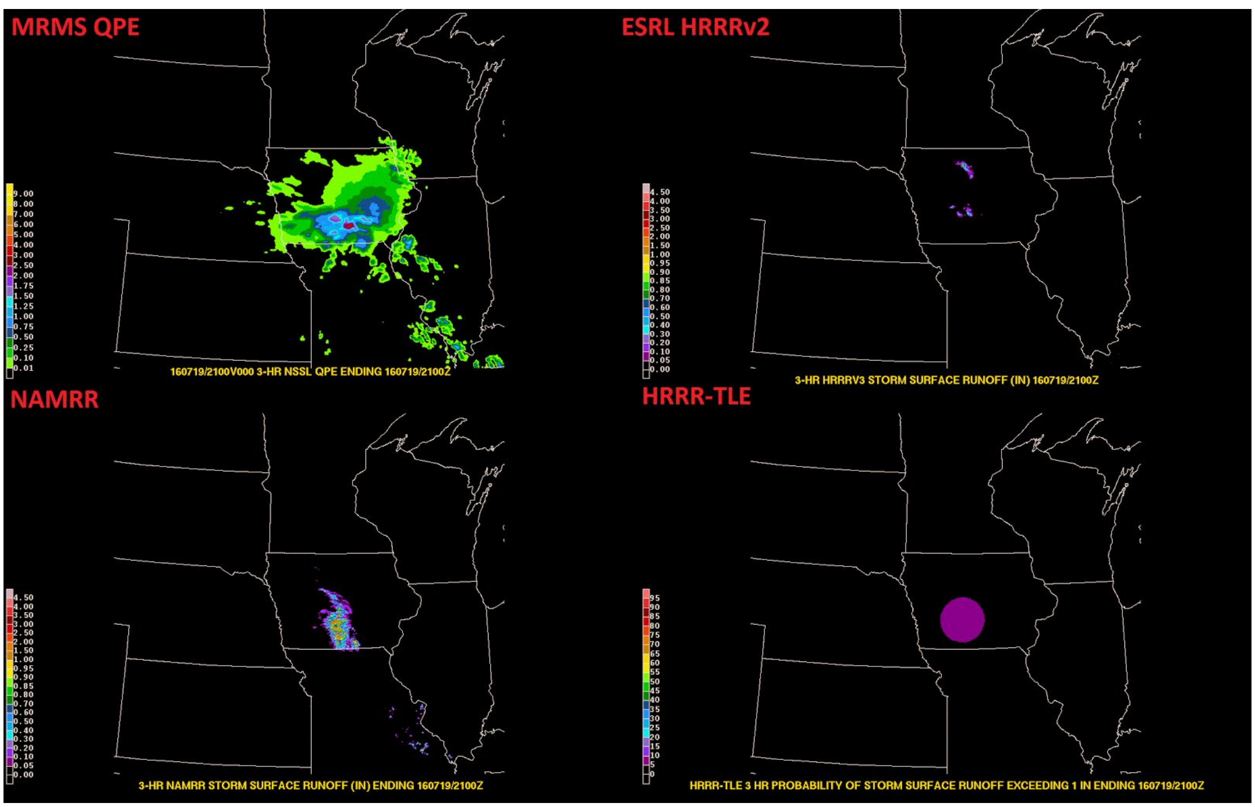


***Figure 37.** An example of the subjective evaluation question, valid July 19, 2016, which*

compares both deterministic and ensemble storm ***surface runoff forecasts*** *to assess individual utility for improving flash flood prediction.*

Participants noted each day that, although calculated the same, consistently the ESRL HRRRv2 produced lighter runoff amounts when compared to the NAMRR which consistently produced high runoff amounts. The dry bias of the ESRL HRRRv2 impacted the HRRR-TLE as well, causing very low probabilities of the "one inch of storm surface runoff over three hours" field that was evaluated for this question. This is attributed to the different LSMs used in each model and as well as QPF differences.

Despite this, there was value in the products by making mental adjustments: the NAMRR was best at showing where the QPF would be most impactful, and if the ESRL HRRRv2 and/or HRRR-TLE had any signal at all, it pointed to a potentially significant event. Although the participants overall were enthusiastic about looking at the storm surface runoff fields for the first time, more work and testing would need to be done for forecasters to trust the products and develop patterns in runoff signals in association with verifiable flooding to have them be contributors to flash flood prediction. Figure 38 shows the average scores of the runoff products, showing that the wet bias of the NAMRR increased its value for the participants over the drier ESRL HRRRv2 products.

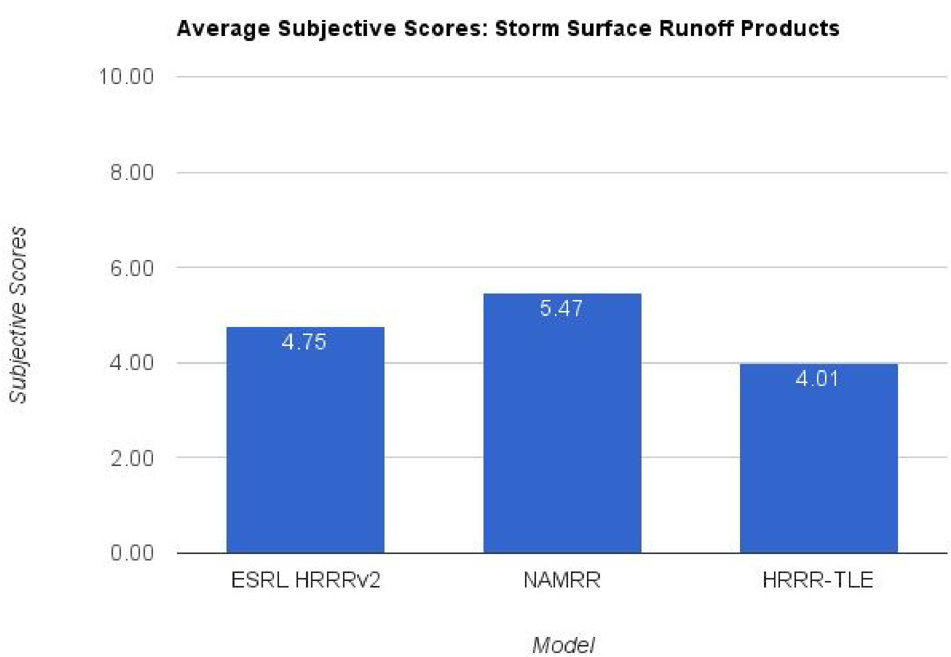


Figure 38*. Average subjective scores, on a scale from 1 to 10, of the storm surface runoff products from the ESRL HRRRv2, NAMRR, and HRRR time-lagged ensemble*.

Soil Moisture

In addition to runoff fields, soil moisture from the NAMRR, ESRL HRRRv2 and the National Water Model (NWM) were also made available to participants as forecast guidance and discussion. Soil moisture and soil moisture availability were available at varying depths of 0 to 1, 10, and 40 cm from the different models (Figure 39). The participants were in overall agreement that the soil moisture products were very useful in the flood forecasting process as a supplement to flash flood guidance. Most favored were the top levels of 0 to 1 and 10 cm with QPF overlaid, as these levels proved to be the most reactive to past and ongoing precipitation in the models. It was noted that the visualization of these fields were beneficial and drew forecaster attention to the areas most prone to flood risk.

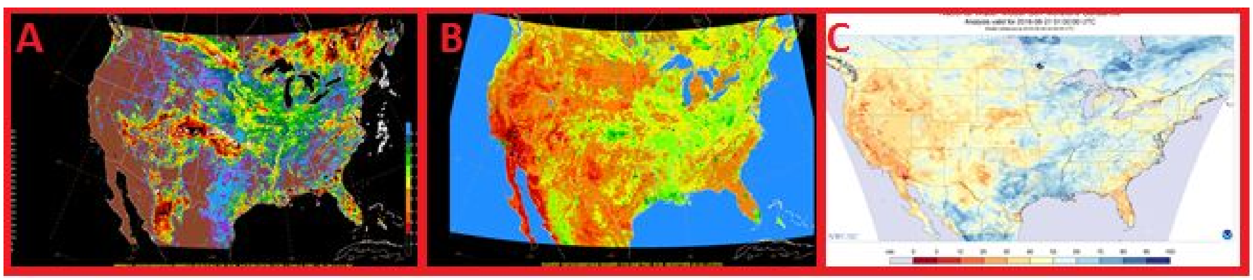


***Figure 39**. Examples of soil moisture products offered, including (a) Percent Soil Saturation for 0-1 cm from the ESRL HRRRv2, (b) the Volumetric Soil Moisture for 0-10 cm from the NAMRR, and (c) the Soil Moisture for 0-40 cm from the National Water Model.*

However, not all reviews were favorable. As with the runoff fields, participants felt these fields needed more testing in order to establish patterns in the data that led to actual flooding events. It was also noted that the soil moisture alone was not as useful as fusing the field with QPF, and possibly other data such as streamflow. Percentages of soil saturation and changes in soil moisture were requested. Deeper soil depths were deemed more useful for longer-lived rain events that may cause areal flooding, and thus were not as useful for the shorter temporal scale of a flash flood. Participants also felt that forecasters would need extensive training to understand the differences between the products, the soil reaction at the different depths, and the varying impacts of heavy QPF with saturated soil.

Probabilistic Tools

During the 2016 FFaIR experiment, participants utilized several high-resolution ensemble neighborhood probabilistic exceedance tools including thresholds for QPF, Flash Flood Guidance (FFG), and Recurrence Intervals (RIs). As shown in Figure 40, varying probabilities were provided from the experimental HREFX, SSEFX, HRRR-TLE and GEFSX ensembles. These tools were designed to provide forecasters with probabilistic "first-guess" information and situational awareness for assessing flash flood risk exacerbated by heavy precipitation, antecedent conditions, and climatology.

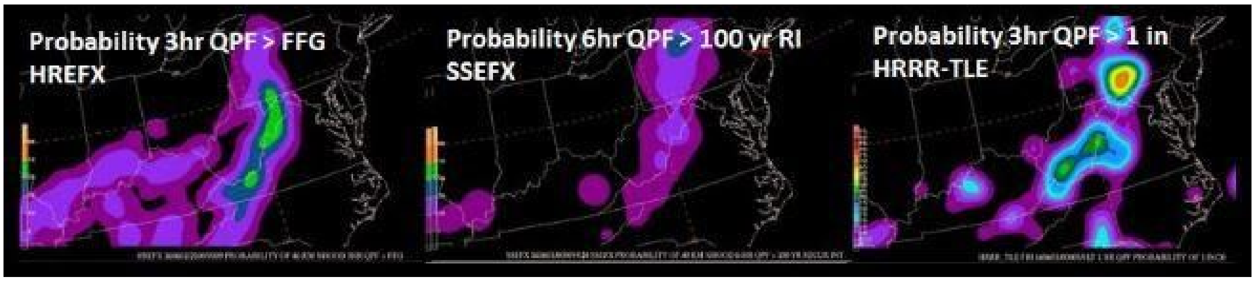


***Figure 40**. Examples of exceedance probabilistic tools provided from several experimental ensembles during the 2016 FFaIR experiment.*

The participants were asked to comment on the utility and value of these neighborhood exceedance probabilities derived from ensemble QPF, how they could be used in the field for flood risk assessment and forecasting, and the different types of verification or verification statistics that might make forecasters more comfortable with these probabilistic fields. All remarked that the ensemble probabilities of exceedance were very valuable tools to the flash flood forecast process. A majority preferred forecasts of QPF exceeding FFG to forecasts of QPF exceeding RIs. It was noted that field forecasters have a deep knowledge of the local soil types, vegetation, and urban development contributing to the FFG, and despite its known shortcomings, still find FFG useful. Comments also included that ARI information is best used as situational awareness rather than applicable flood forecast guidance. Though most were in favor, it was suggested that the neighborhood technique causes too much spread in the ensemble guidance at times making it difficult to accurately assess smaller areas of highest flash flood risk, whereas point probabilities may offer an advantage. Some participants also felt that shorter time ranges, such as 1-hour and 3-hour forecasts, were more valuable and provided more flash flood information than longer-range forecasts (6-hour, 12-hour, 24-hour).

5. OTHER EXPERIMENTAL TOOLS RESULTS

Excessive Precipitation with Elevated Convection (EPEC) Index

In addition to the experimental models and ensemble systems used in the 2016 FFaIR Experiment, other tools like the EPEC Index were used as well. The EPEC Index combines the K-index, precipitable water, and divergence at 250 hPa to try and diagnose likely areas of elevated convection, which could lead to areas of heavy rainfall. The index was applied to three models during FFaIR: the operational NAM, GFS, and European model. Participants used the index throughout the four weeks in various situations and weather patterns and asked to comment on the tool at the end of the week.

The EPEC Index was used during the experiment to identify a focus area of heavy precipitation among other model guidance near the end of the forecast process to double-check that the index was in agreement. Figure 41 is an example of a day where the EPEC Index showed very high values indicating areas where elevated convection were likely to occur valid at 21Z on July 11, 2016. In Figure 41, very heavy rain associated with convection did occur in Minnesota over the 18-00Z period of the same day, albeit with a somewhat different orientation/location than what EPEC predicted.

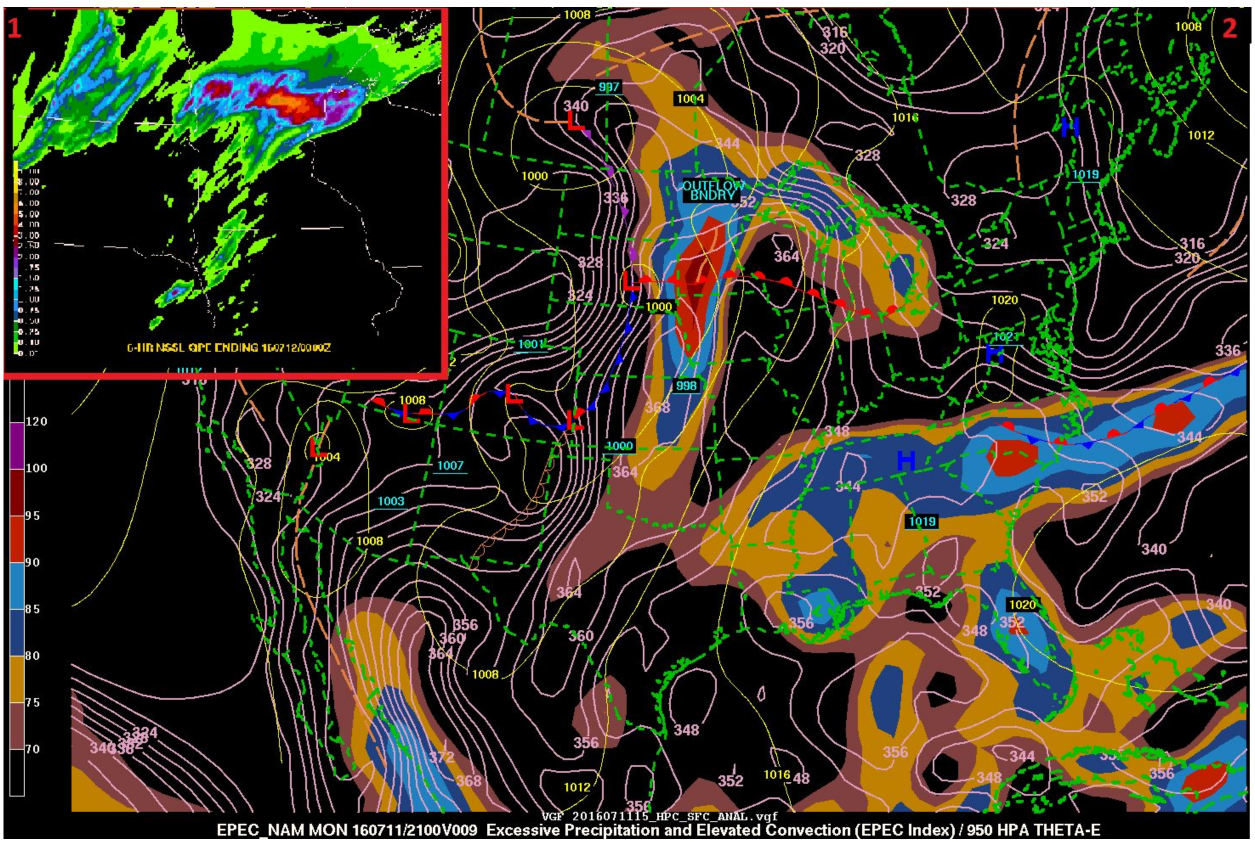


***Figure 41**. (1) 18Z July 11 - 00Z July 12, 2016 MRMS QPE. (2) EPEC Index from the operational 12Z NAM model valid at 21Z July 11, 2016. Values of red indicate high likelihood of elevated convection. WPC surface fronts valid at 15Z July 11 are overlaid.*

Survey feedback on the EPEC Index was mixed from participants. Some did not see a value in using an ingredients based index like EPEC when compared to available high-resolution, convection allowing model guidance. Some did not feel like there were enough cases using it to really determine its usefulness, while a couple participants questioned the connection between heavy precipitation or elevated convection and whether it has utility for flash flood forecasting. More explanation and training might also be needed to when conditions favor using the index.

There were times that the index showed high values in locations that, based on the synoptic setup, would not normally be favored for elevated convection. Several participants did find the index useful, especially in terms of a situational awareness tool for getting an idea of areas to pay closer attention to. Others found it to be a nice, quick way to evaluate elevated convection potential rather than looking at multiple soundings and other information. Overall, the EPEC Index, used during FFaIR as a way to diagnose elevated convection that may lead to heavy rainfall and flash flooding, was positively reviewed as a good situational awareness tool. Again, there were some participants that questioned the utility of an ingredients based tool with the availability of high-resolution CAMs.

GOES-R Advanced Baseline Imager (ABI) Algorithm for Rainfall Rate/QPE

The GOES-R ABI QPE algorithm was evaluated during the subjective science evaluation portion of the 2016 FFaIR Experiment. The algorithm uses infrared and microwave, for calibration, satellite inputs to estimate the amount of precipitation. To evaluate this product during FFaIR, 24 HR QPE from GOES-R was compared with 24 HR Stage VI precipitation data.

The average subjective score for this product, which is based on the usefulness of the 24 HR GOES-R QPE as a QPE product, was 4.1 out of 10. Figure 42 shows an example of the GOES-R QPE product compared to the Stage IV data valid from 12Z July 19 - 12Z July 20. The majority of participants noted that over mountainous regions in the West where traditional gauge data is sparser, the GOES-R QPE did not perform any better and often-times missed precipitation in those regions and often had trouble handling precipitation in the Southeast and along the Gulf coast. Additionally, for large convective events, the GOES-R QPE was often too heavy or covered too large of an area. The strengths of the GOES-R algorithm were the estimates of QPE from convective events. The algorithm is being tested as a proxy for GOES-R using GOES-13/15 and will improve with the new satellite series. Although it is not really meant for current GOES, it will be much-improved for GOES-R, including a 2 km resolution. Other applications for this tool were identified, such as coverage for areas off the coast of the United States where traditional Stage IV data and MRMS QPE do not provide data. The GOES-R QPE could help forecasters evaluate atmospheric river events that are still off-shore in the Eastern Pacific Ocean. With continual improvements to the algorithm and the launch of GOES-R itself, this product will likely improve and need to be re-evaluated.

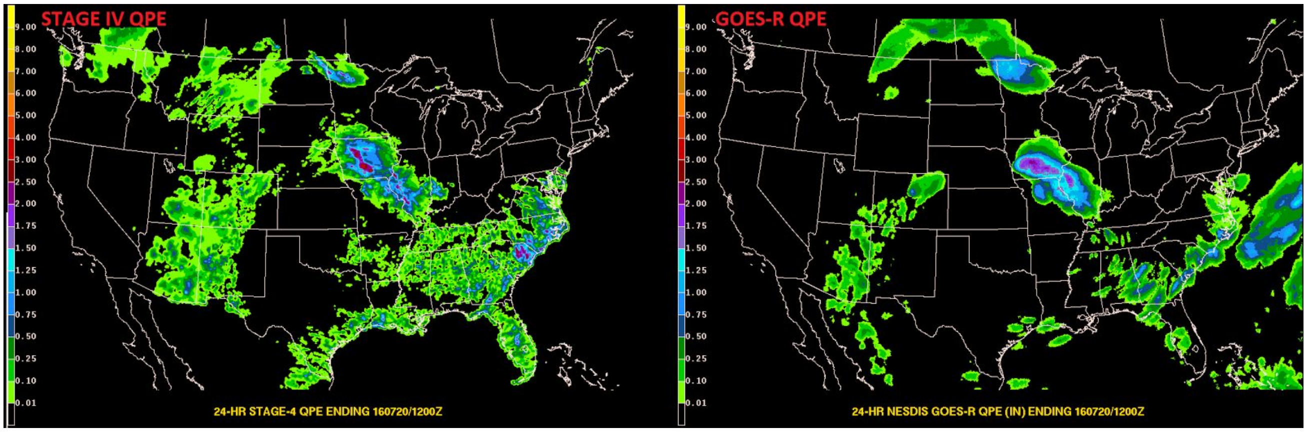
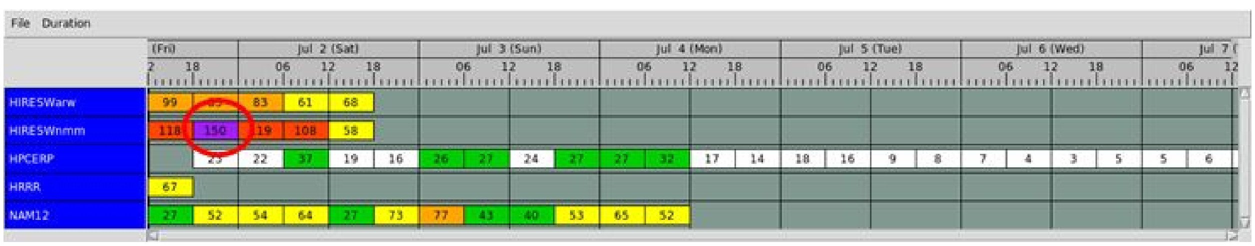


Figure 42. *Stage IV QPE (left) and the GOES-R QPE algorithm (right) valid from 12Z July 19 - 12Z July 20.*

Excessive Precipitation Forecasting Table (EPFT)

The Extreme Precipitation Forecasting Table (EPFT) (Fig. 43) was used during FFaIR as a situational awareness tool on the AWIPS II workstation for displaying the ratio of the deterministic model QPF to the 6-hour 100-year annual recurrence interval. Participants were asked if they found the EPFT to be effective for alerting forecasters of possible heavy rain which may lead to flash flooding and to comment on the usability of the interface.



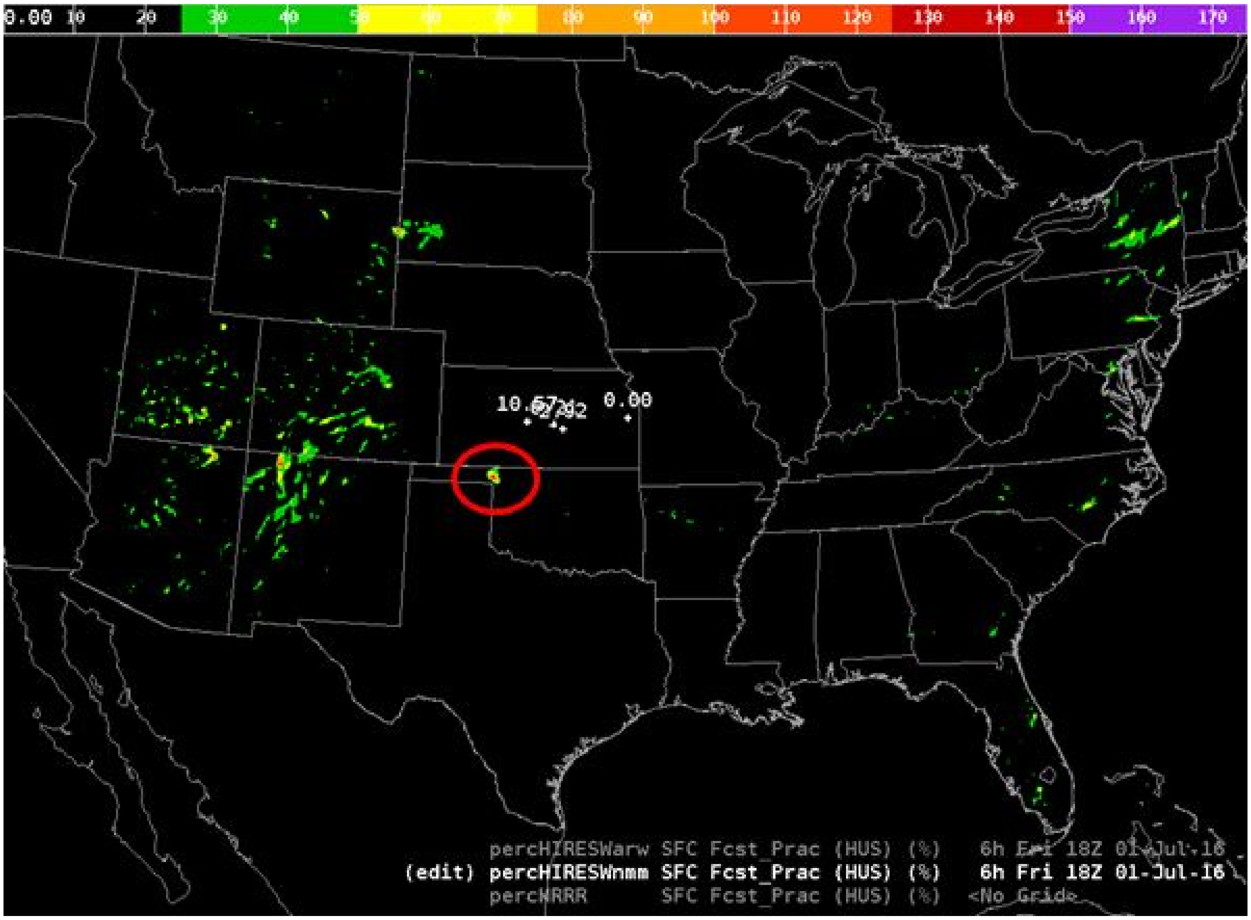


Figure 43. *An example of the EPFT showing data in the table for 5 models: HIRES ARW, HIRES NMMB, 5-km WPC QPF (HPCERP), HRRR, and the NAM12. Map display shows the graphical results of the QPF/ARI ratio from the HIRES NMMB. Red circles indicate the corresponding high value in the table to its location on the graphical map display.*

Overall, the participants found the EPFT to be a valuable situational awareness tool for alerting forecasters to the potential of flooding. However, there was confusion when attempting to understand the QPF/ARI ratio concept as opposed to a more direct exceedance representation or probability. Frustration was noted when forecasters had to pan around on the map to find the location of the highest ratio value stated in the table because at times it could be a very small area. Also noted was the dependency on the skill of the model QPF and the desire to also see the ratio against QPE and/or gauge observations.

Other remarks included the desire to have more high-resolution models available, and more importantly, ensembles and ensemble probabilities of the model QPF to ARI ratio. Participants also wanted more ARI intervals in addition to the 6-hour, 100-year. Field forecasters liked the idea of having this available in GFE, but would like to see a more user-friendly interface. Many feel that forecaster training is imperative to maximize the utility of the EPFT.

6. EXPERIMENTAL PROBABILISTIC FORECAST PERFORMANCE

The participants scored the experimental Day 1 and Day 2 Excessive Rainfall Outlooks (ERO) and 6-hour Probability of Flash Flooding (PFF) forecasts that were produced each day. They were asked to subjectively rate the skill of areal extent and probabilistic value of the forecasts on a scale of 1 to 10 using verification resources such as MRMS QPE, USGS rain gauges, NWS local storm reports, mPING reports and NWS-issued flash flood warnings.

Increased skill in the forecasts was apparent with larger scale, synoptically-forced events, for which the CAMs and other models better performed. Spatial uncertainty was introduced with more generalized and small scale convection which led to flash flooding not easily captured by the CAMs or ensembles. Skill and confidence in the forecast ensemble probabilities is still growing but not yet fully trusted among forecasters, as several forecast evaluations noted. It was observed throughout the experiment that at times forecasters hesitated to over-emphasize common summer patterns such as daily convection along the Gulf Coast and Florida and the monsoonal flow over the southwestern states. Unfortunately, this resulted in flash flooding that was missed to the detriment of the experimental ERO.

Subjective Results

The participants rated the Day 1 ERO more harshly than the Day 2 ERO. Given the temporal range of the guidance and the forecast, they were more lenient on the Day 2 ERO results. Although the overall average scores were close (6.4 out of 10 for the Day 1 ERO and 6.7 out of 10 for the Day 2 ERO), it is shown in Figure 44 that the Day 2 ERO had higher subjective scores (6 or above) more frequently than the Day 1.

The Probability of Flash Flooding (PFF) forecasts were rated the highest out of the three daily forecasts, averaging 6.8 out of 10. Some felt this forecast had a significant advantage because of the newly-available high-resolution guidance and current radar from which to create the 6-hour forecast that was valid 18-00 UTC each day and had to be issued by 18Z. CAM reflectivity was often used more than probabilistic atmospheric or hydrologic guidance due to the short-term nature of the product, although, on several days participants commented that the hydrologic guidance was key to the final decision on the magnitude of the contours. On three of the experiment days, the participants decided there was not enough flash flood risk to issue a product and instead selected “probabilities too low,” however, one of the those days included significant flash floods in Las Vegas, NV which resulted in water rescues. It was noted that none of the experimental guidance picked up on this event in advance.

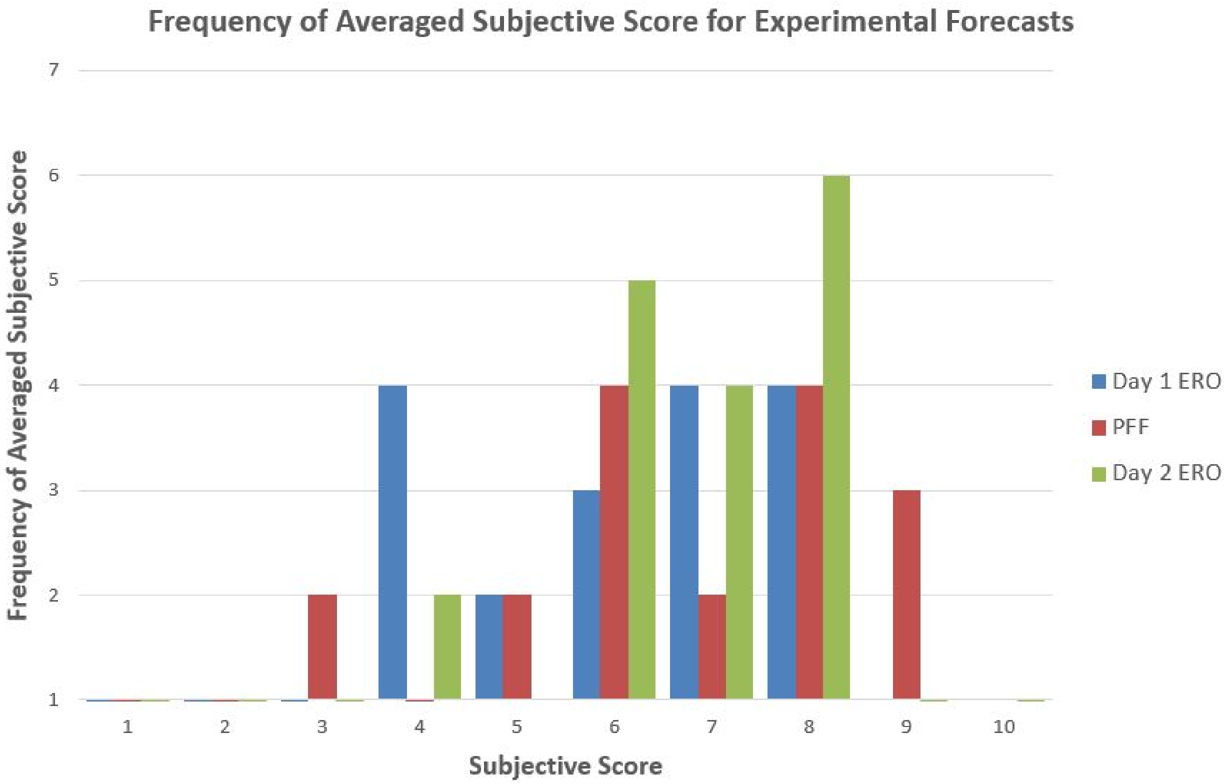


Figure 44. *The frequency of the subjective scores for the daily experimental 2016 FFaIR forecasts rated on a scale from 1 to 10.*

Objective Results

Objective statistics were calculated on both the ERO and Day 2 ERO showing the percent of area within the probabilistic contours that is within 40 km of a flash flood or flood report. Ideally, this percentage would equal the probability of the contour. Figure 45 shows the results for the last three years with 2016 in green, 2015 in red, and 2014 in blue. The percent area within 40 km of a point for the 2% line in 2016 was 2.00%, which was an improvement over the past two years. For the five percent contour, the percent area within 40 km was 3.49%, slightly lower than 2015. Note, there was no 5% contour in the 2014 experiment. The 10% contour has an percent area within 40 km of a report of 11.49% in 2016, which was an improvement over the previous two years. Finally, only one 30% contour was forecast over the entire 2016 experiment, which had a percent area within 40 km of a report of 66.67%. The previous years had at least six 30% forecasts. Figure 46 shows a comparison from the last three years of the average areal coverage of the probabilistic ERO forecasts. Every year the average areal coverage of the probabilistic forecasts have decreased while maintaining similar or better skill in forecasting. Finally, Figure 47 and Figure 48 compare the percent area within 40 km of a report and average areal coverage of the 2016 ERO and Day 2 ERO (ERO2). The ERO2 probability contours were closer to perfect calibration for the 2, 5, and 10% probabilities. There were no 30% forecasts for the ERO2. The average areal extents of the probability contours for both forecasts were very similar, with the ERO2 probabilities having slightly smaller average areal extents.

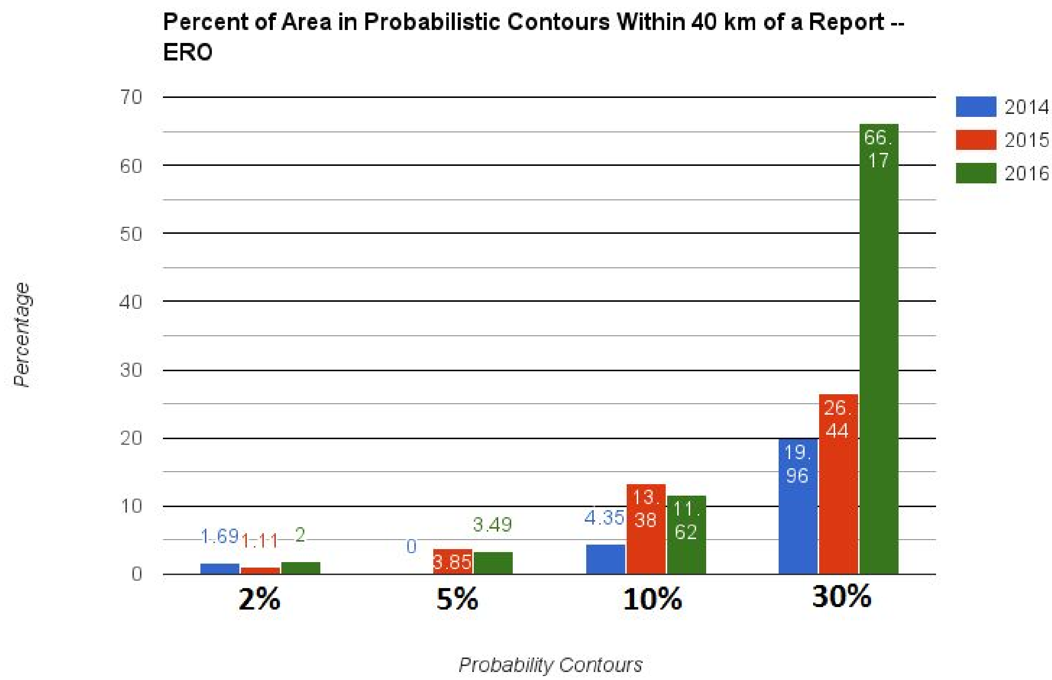


Figure 45. *Percent of area in probabilistic contours within 40 km of a flash flood/flood report for the 2014 (blue), 2015 (red), and 2016 (green) Excessive Rainfall Outlooks.*

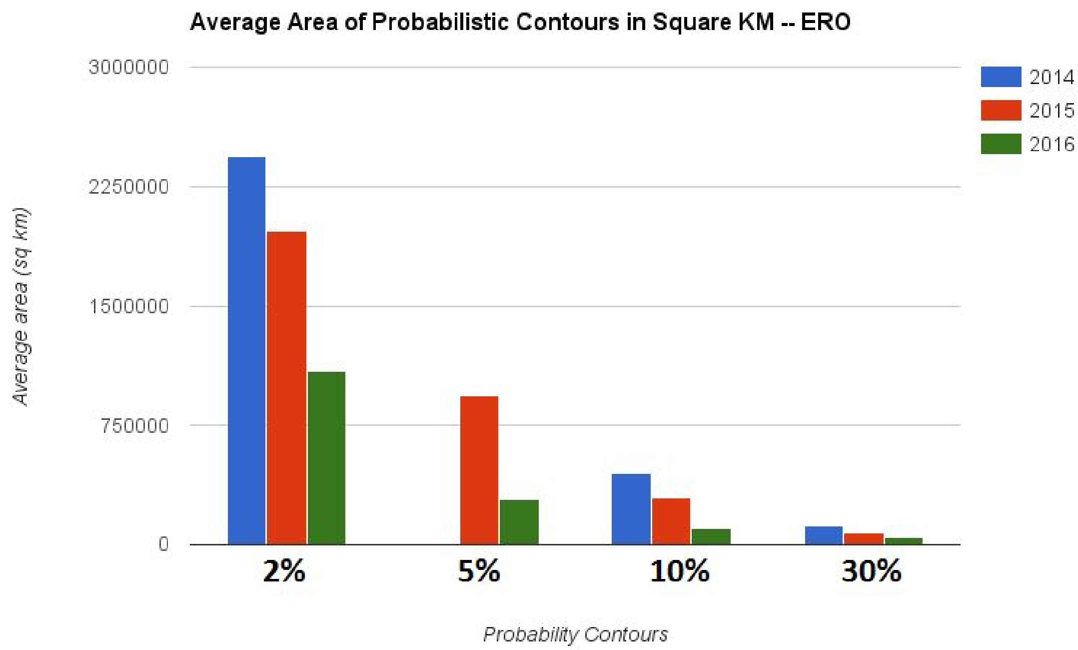


Figure 46. *Average areal coverage (sq km) of the 2014 (blue), 2015 (red), and 2016 (green) Excessive Rainfall Outlook probabilistic contours.*

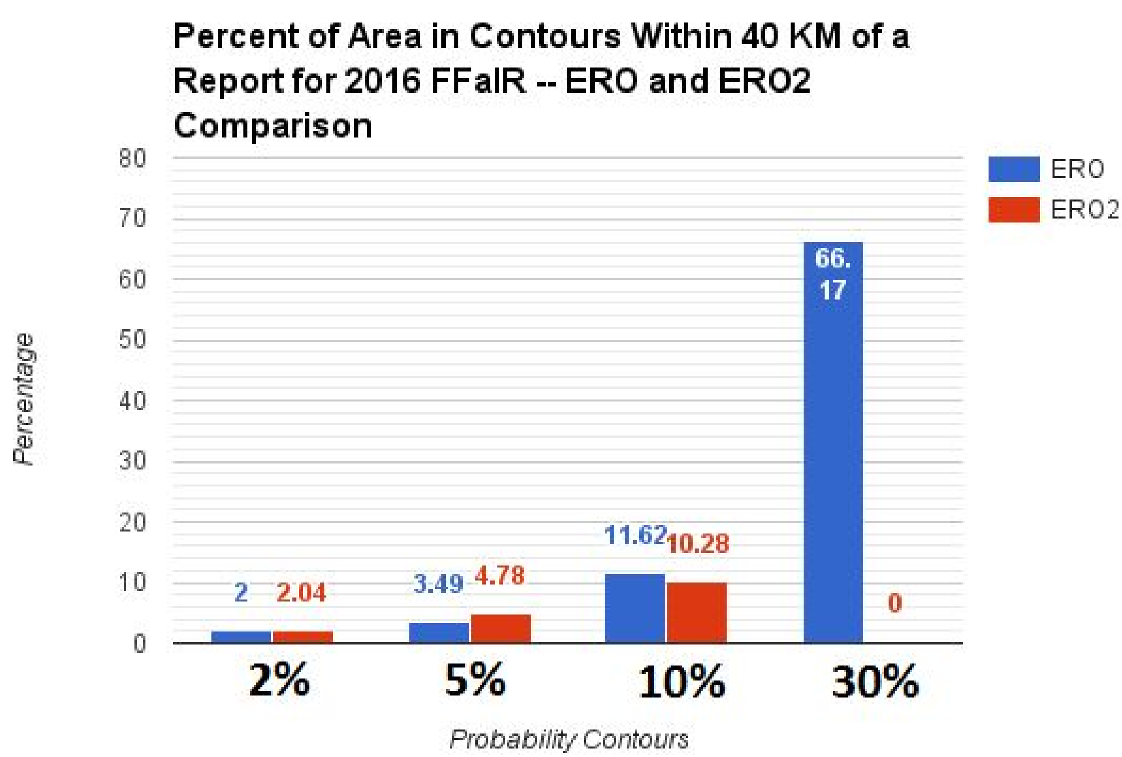


Figure 47. *Percent of area in probabilistic contours within 40 km of a flash flood/flood report for the 2016 Day 1 Excessive Rainfall Outlook (blue) and Day 2 Excessive Rainfall Outlook (red).*

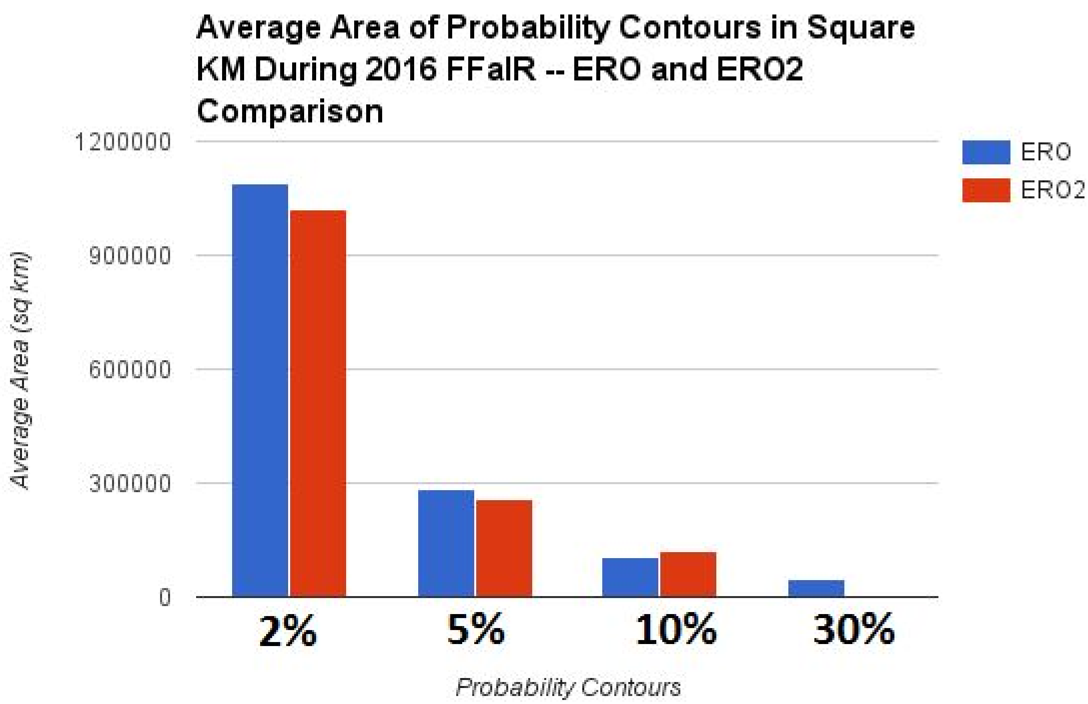


Figure 48. *Average areal coverage (sq km) of the 2016 Day 1 Excessive Rainfall Outlook (blue) and Day 2 Excessive Rainfall Outlook (red) probability contours.*

Similar to the ERO, objective statistics were also computed for the 18-00Z PFF forecast issued each day. Figure 49 shows the percent of area in the probabilistic contours within 40 km of a flash flood/flood report. The probabilities for the PFF for perfect calibration are 10, 30, and 50%. The percentages for the 10 and 30% contours are quite below perfect calibration for all three years. The 50% probability contour is more varied over the three years and it should be noted these are based on just two forecasts of 50% in 2014 and 2015 and just one forecast of 50% in 2016. Therefore, there is not a large enough sample size from which to draw conclusions. Figure 50 shows the average areal coverage of the PFF probabilistic contours over the past three years. This shows that the average areal coverage in 2016 followed the trend of the EROs above and decreases while maintaining similar skill.

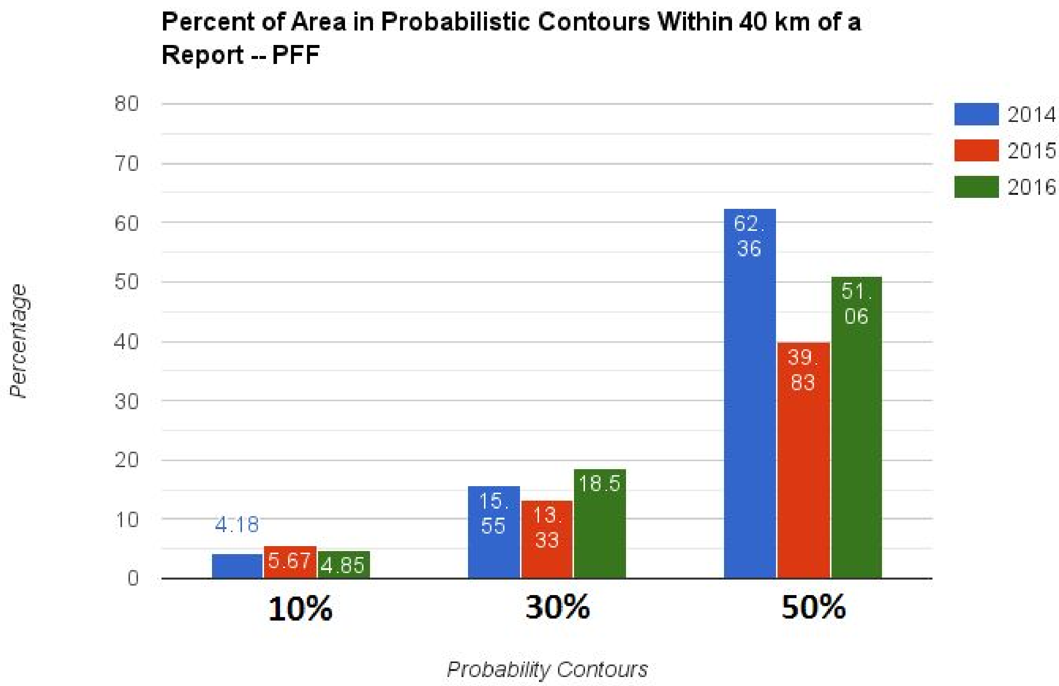


Figure 49. *Percent of area in probabilistic contours within 40 km of a flash flood/flood report for the 2014 (blue), 2015 (red), and 2016 (green) Probability of Flash Flooding Forecast valid from 18-00Z each day.*

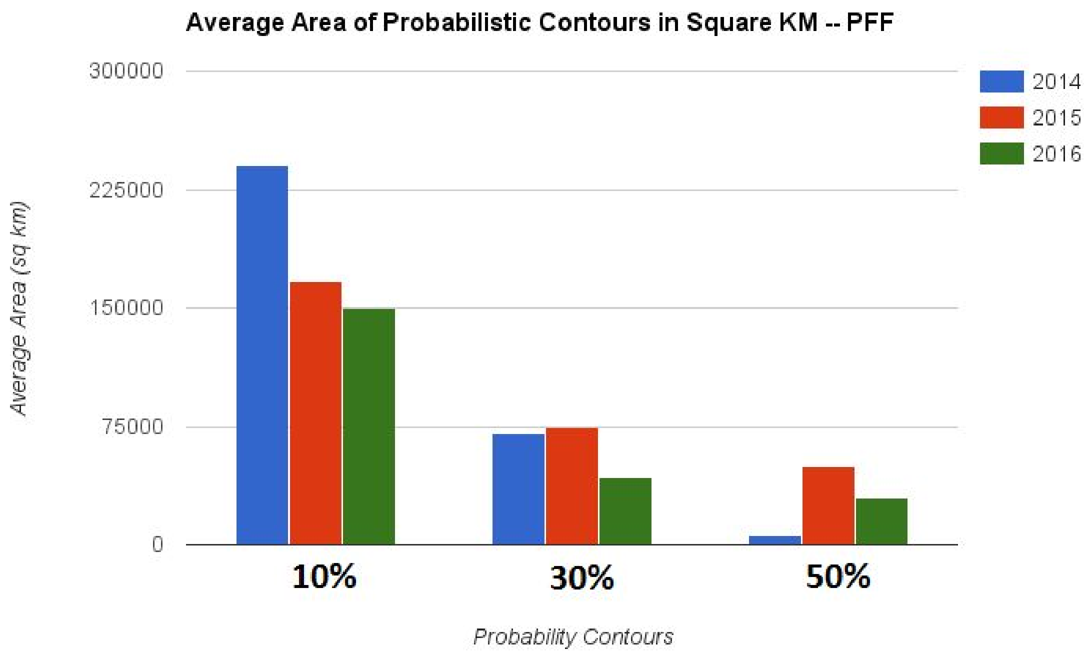


Figure 50. *Average areal coverage (sq km) of the 2014 (blue), 2015 (red), and 2016 (green) Probability of Flash Flooding forecast probabilistic contours.*

7. SUMMARY AND CONCLUSIONS

The 2016 FFaIR Experiment successfully broadened the suite of both operational and experimental hydrologic forecasting tools available to the 30 participants from across the forecasting and flash flood communities. These emerging hydrologic products were utilized in tandem with both operational and experimental atmospheric QPF guidance and extreme precipitation awareness tools to develop flash flood outlooks and forecasts which were evaluated for utility and skill.

The experiment activities included the daily issuance of experimental Day 1 and Day 2 EROs, a 6-hour probabilistic flash flood forecast valid 18-00 UTC, and subjective evaluation of the forecasts and experimental data. Statistical evaluation of the experimental forecasts shows overall reduced areal coverage and similar or increased skill of the drawn probabilistic contours which demonstrates improved forecaster confidence and applications of new and better guidance. Participants and testbed staff agree that objective verification methods would enhance the experimental evaluations and should be phased into future experiments. This may include:

- Brier Skill Scores (BSS)
- Relative Operating Characteristics (ROC) curves
- Fraction Skill Scores (FSS)

In addition to the evaluation of the experimental guidance and forecasts, the participants were polled regarding the applicability of new hydrologic and atmospheric guidance to improvement of flash flood prediction and public preparation. The overall consensus taken from the 2016 FFaIR experiment is below, with clearly denoted key points:

1. **High-resolution convection-allowing models, both deterministic and ensemble, are essential and improving each year**. There is value to running them out further in time to cover longer forecast periods.
2. **Probabilistic exceedance tools are very useful to probabilistic flood forecasting and should continue to be developed further**. However, debate still exists among forecasters regarding neighborhood probabilities versus point probabilities.
3. **The hydrologic and atmospheric data must be used together to effectively predict flash flooding** and convey impact messaging to the public, therefore more and better tools should be developed that bring this guidance together efficiently.
4. **More testing is needed over an entire warm season to fully assess the utility and identify patterns of new guidance** that is available such as runoff fields, shallow-depth soil reaction, and streamflow anomaly forecasts which may indicate flash flood risk.
5. New tools that can accompany and support Flash Flood Guidance were liked and desired, however, FFG is still the primarily favored resource for antecedent conditions and flood vulnerability assessments. **Climatological and extreme precipitation information such as ARIs are beneficial, but mostly supported as glancing situational awareness rather than a primary flood forecasting tool**.
6. **No tools or pieces of guidance were perfect solutions to the flash flood problem.** All tools, guidance, and applications should be used together by the forecaster to help determine the likelihood of flash flooding. Future experiments will further explore the strengths and weaknesses of the tools and guidance.
7. **The need continues for an agency-wide approach to the flash flood forecast problem, ranging from definition, to warning practices, to reporting requirements.** Subjective opinion and geographic inconsistencies with regard to the definition of a flash flood, combined with inconsistent reporting practices from WFO to WFO, leads to a severe lack of database verification when it comes to identifying where flash flooding has occurred.

Acknowledgments

The authors of this report, Ben Albright and Sarah Perfater, would like to sincerely recognize all of those who offered their support and skills to the execution and success of the 2016 FFaIR Experiment. This includes, but is not limited to: Mike Bodner, Mark Klein, Kathy Gilbert, Jim Nelson, David Novak, Steve Martinaitis, JJ Gourley, Arlene Laing, Greg Carbin, and the WPC Senior Branch Forecasters.

8. REFERENCES

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.

Foscato, A. and P. Market, 2015: An Index for Anticipating Excessive Precipitation with Elevated Thunderstorms, National Weather Association 40th Annual Meeting, Oklahoma City, OK. Available at: http://www.nwas.org/meetings/nwas15/presents/2571.pptx

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

Jirak, I. L., S. J. Weiss and C. J. Melick, 2012: The SPC storm-scale ensemble of opportunity: overview and results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. *Preprints*, 26th Conf. Severe Local Storms, Nashville, TN. Amer. Meteor. Soc. P9.137.

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 20 – 24	Robert Oravec Bruce Sullivan	Rich Grumm (WFO CTP) Paul Fitzsimmons (WFO BGM) Dan Valle (WFO PHI)	Curtis Alexander (ESRL-GSD) Patrick Market (University of Missouri) Kelly Mahoney* (ESRL-PSD)	Roshan Shrestha (EMC-G) Shun Liu (EMC-M) Matt Pyle (EMC-M)
June 27 – July 1	Brendon Rubin-Oster Mike Musher Robert Oravec	Jane Marie Wix (WFO JKL) David Marsalek (WFO RLX) Katherine Rowden (WFO OTX) Aaron Gleason (SPC-meso)	Tamarah Curtis (MDL) Jill Hardy* (WDTD)	Hua-Ya Chuang (EMC-G) Tracey Dorian (EMC-G) Geoff Manikin (EMC-M) Jacob Carley (EMC-M)
July 11 – 15	Patrick Burke Greg Carbin	Mason Rowell (ABRFC) Pete Geogerian (WFO JKL)	Isidora Jankov (ESRL-GSD) Keith Brewster (OU-CAPS) Jess Erlingis Lamers* (OU) Tressa Fowler* (NCAR)	Glenn White (EMC-G) Binbin Zhou (EMC-M)
July 18 – 22	Robert Oravec	Jessica Winton (WFO MRX) Peter Corrigan (WFO RNK)	Wanru Wu (OWP) Trevor Alcott (ESRL-GSD) Greg Herman (Colorado State Univ.) Diana Stovern* (WPC) Julie Demuth* (UCAR)	Brad Ferrier (EMC-M) Eric Aligo (EMC-M)

APPENDIX B

Sample Daily Schedule

8:00 am – 9:00 am	Team Hydro and Team Atmos work separately to assess assigned fields
9:00 am - 10:15 am	Collaborative Excessive Rainfall Outlook creation, valid 15 – 12 UTC Prepare discussion/PPT
10:15 am – 10:30 am	Break
10:30 am – 11:45am	Subjective model evaluation (WPC Forecaster will exit to brief at Map Discussion, return for voting)
11:45 am – 12:45 pm	Lunch
12:45 pm – 1:45 pm	Probabilistic flash flood forecast, valid 18 – 00 UTC
1:45 pm - 2:00 pm	Finish discussion/PPT
2:00 pm – 2:40 pm	HMT-Norman forecast briefing
2:40 pm - 2:50 pm	Break
2:50 pm – 4:00 pm	Collaborative Day 2 ERO Creation, Valid 12 - 12 UTC