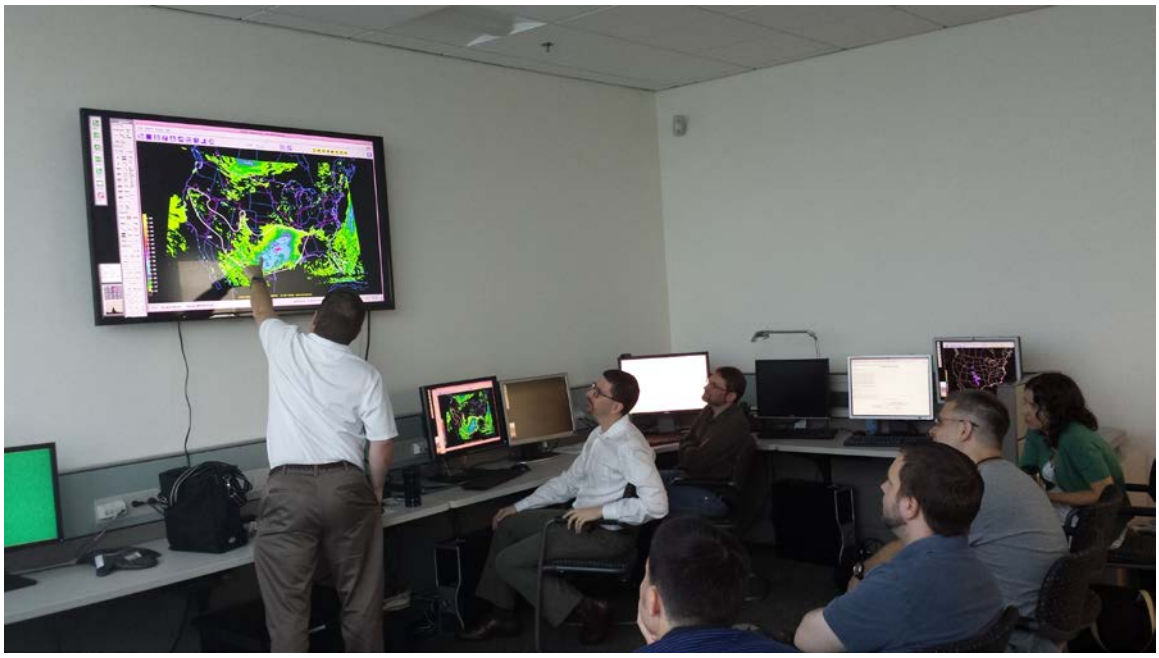
The 2014 Flash Flood and Intense Rainfall Experiment

Final Report

September 15, 2014



1. INTRODUCTION

From July 7-25, 2014, the Hydrometeorological Testbed at the Weather Prediction Center (HMT-WPC) hosted the second annual Flash Flood and Intense Rainfall Experiment (FFaIR). Developed in support of WPC's Excessive Rainfall Outlook (ERO) and Mesoscale Precipitation Discussions (MPD), FFaIR investigated methods to improve flash flood forecasting in both the immediate near term (1-6 hours) and short term (Day 1) forecast periods. The experiment brought together 23 forecasters, researchers, and model developers (Appendix A) from across the weather enterprise to explore the challenges faced by forecasters related to flash flood and quantitative precipitation forecasting (QPF) during the warm season.

In particular, the goals of the experiment were to:

- Evaluate the utility of high resolution convection-allowing models and ensembles for short-term flash flood forecasts.
- Explore proposed changes to WPC's operational Excessive Rainfall Outlook.
- Explore the utility of probabilistic flash flood forecasts at different forecast lead times.
- 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 summarizes the activities, findings, and operational impacts of the experiment.

2. EXPERIMENT DESCRIPTION

Data

In addition to the full multi-center suite of operational deterministic and ensemble guidance, the 2014 FFaIR Experiment featured several experimental ensemble systems: the Storm-Scale Ensemble of Opportunity (Jirak et al., 2012), provided by the Storm Prediction Center (SPC), a modified version of the SSEO, provided by WPC, the Experimental Regional Ensemble Forecasting System (ExREF), provided by the Earth System Research Laboratory's Global Science Division (ESRL/GSD), and the NCEP Convection Allowing Scale Ensemble (NCASE), provided by the Environmental Modeling Center (EMC). The experiment also featured two experimental high-resolution deterministic models: the High Resolution Rapid Refresh (HRRR) provided by ESRL and the parallel version (NAMP) of NCEP's North American Model (NAM) provided by EMC. Table 1 summarizes the model data that was the focus of the experiment.

Table 1. Featured 2014 FFaIR deterministic and ensemble model guidance. Experimental guidance is shaded.

Provider	Model	Resolution	Forecast Hours	Notes
RFCs	Flash Flood Guidance	5 km	01, 03, 06, 12 and 24 hour values	CONUS mosaic grid created by compiling individual RFC-domain grids
EMC	NAM	12 km (parent) 4 km (nest)	84 (parent) 60 (nest)	Operational NAM, includes 12 km parent model and 4 km CONUS nest
EMC	SREF	16 km (displayed at 32 km)	84	Operational SREF mean
EMC	NAM Parallel	12 km (parent) 4 km (nest)	84 (parent) 60 (nest)	NAMP; features differing analysis (ENKF), convective schemes, and physics from operational NAM
ESRL	HRRR	3 km	15	High resolution, hourly updated, convection allowing nest of the Rapid Refresh (RAP) model
SPC	SSEO (7 members)	4 km	36	Multi-physics, convection allowing ensemble consisting of 7 high-resolution deterministic models
SPC/ESRL/WPC	WPC-SSEO (7 members)	4 km	24	Modification of the original SSEO provided by SPC
ESRL/GSD	ExREF (8 members)	9 km	84	Multi-physics, multi-initial condition, multiple boundary condition ensemble
EMC	NCASE (15 members)	~4 km	12	Multi-model, multi-physics, time lagged ensemble consisting of 5 NAM conus-nest, 4 HRRR, 3 WRF-ARW and 3 WRF-NMM runs

Deterministic Guidance

The NAMP, provided by EMC, was the parallel version of the NAM[1] and featured both a 12 km parent and a 4 km CONUS nest. This version uses the global Ensemble Kalman Filter (EnKF) members as part of its regional GSI data assimilation system, and employs the Rapid Radiative Transfer Model (RRTM) radiation scheme (as opposed to the GFDL scheme in the operational NAM). Additionally, the parent 12 km features the Betts-Miller-Janic (BMJ) convective scheme, and its 4 km nest is fully convection allowing; this differs from the operational NAM in which the parent 12 km uses the base BMJ parameterization and the nest uses a modified version of the BMJ scheme to initially trigger convection.

[1] *The parallel version of the NAM used in FFaIR became the operational NAM starting with the 12 UTC cycle on 12 August, 2014.*

The High Resolution Rapid Refresh (HRRR, http://ruc.noaa.gov/hrrr) was also featured during the experiment. The HRRR features 3 km resolution and is hourly updating, using initial conditions from the Rapid Refresh (RAP) model. It has a WRF-ARW core, Thompson microphysics, and is fully convection allowing. The HRRR is initialized with the latest 3-D radar reflectivity using radar-DFI (digital filter initialization) technique (via the parent 13 km RAP) and provides output hourly.

Ensemble Guidance

The SSEO is a high-resolution, multi-model, multi-physics, convection-allowing ensemble system produced by SPC. 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 NMMB high resolution windows) are time-lagged by 12 hours to provide additional initial condition diversity (Jirak et al, 2012). It should be noted that the NSSL WRF-ARW and EMC WRF-NMM are non-operational and can be subject to outages, and the four high resolution window members (HRW-ARW and HRW-NMMB) are operational, but can be supplanted with other high resolution 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 6) with the latest cycle of the HRRR. This is done to mitigate the high QPF bias that has been observed with the EMC WRF-NMM. In addition to the standard 00 and 12 UTC runs, the modified SSEO is also 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 and 18 UTC cycles of the HRRR and NAM nest.

Table 2. Membership characteristics of the SSEO and modified SSEO. Members denoted by the asterisk (*) are time lagged by 12 hours. For the modified 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[2]	EMC	4 km	MYJ	Ferrier

[2] *During the 8-25 July period, the parallel version of the NAM (NAMX) was used in the modified SSEO.*

NCASE is a convection allowing, multi-model, time-lagged (Fig. 1) ensemble produced by EMC that contains 15 members: the 5 most recent NAM nest runs, the 3 most recent WRF-ARW and WRF-NMMB runs, and the 4 most recent HRRR runs. The ensemble employs a linear weighting scheme in which each member is weighted by its age (e.g. older model runs have less weight in the ensemble solution). The system is run every three hours, when available, and provides hourly output out to 12 hours.

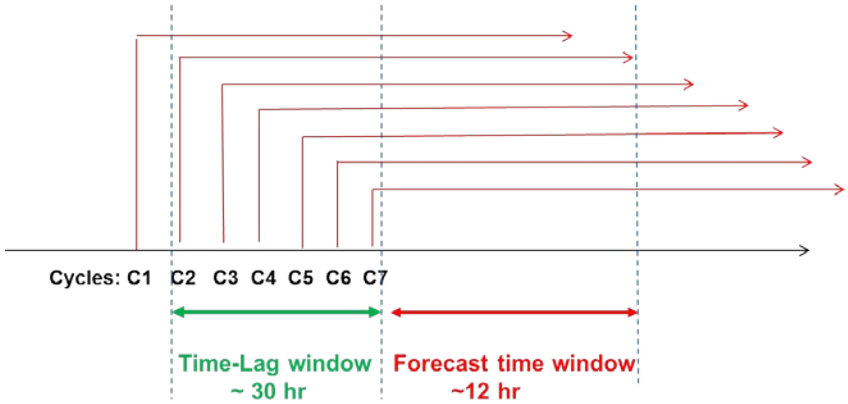


Figure 1. Showing the time-lagged structure for the NCASE ensemble. Image adapted from Zhou et al. (2014).

Finally, ExREF is a multi-physics, multi-initial condition, multiple boundary condition ensemble (Table 3). 7 of its 8 members feature use of the Local Analysis and Prediction System (LAPS; laps.noaa.gov) for their initial conditions, with the first member using the GFS analysis. The system is displayed at 9 km resolution, and available at 00 and 12 UTC.

Table 3. Membership characteristics of the ExREF. Member denoted by asterisk (*) denotes use of the 'variational' version of the LAPS analysis; all others use the 'traditional' version.

Member	Initial Conditions	Boundary Conditions	Microphysics
m00	GFS	GFS	Thompson
m01	LAPS	GFS	Thompson
m02	LAPS	GEFS 01	Ferrier
m03	LAPS	GEFS 02	WSM6
m04	LAPS	GEFS 03	Thompson
m05	LAPS	GEFS 04	Ferrier
m06	LAPS	GEFS 05	WSM6
m07	LAPS*	GFS	Thompson

Ensemble Forecast Tools

The modified SSEO and ExREF were both used to create experimental probabilistic flash flood forecast tools. Neighborhood probabilities (e.g., Schwartz et al. 2009, Schwartz et al. 2010, Ebert 2008) were generated for the two systems based on the 'neighborhood maximum value'; this technique accounts for spatial uncertainty in high-resolution model forecasts by conducting a search within a certain radius (e.g. 40 km) of each grid point to locate the maximum value of a parameter (e.g. precipitation) within that radius. The *value of the original grid point is then replaced with this maximum value,* and ensemble probabilities of exceedance are calculated.

Neighborhood probabilities based on a 40 km search radius were created for two threshold concepts:

- QPF exceeding a certain amount (e.g. 1 inch; QPF > 1")
- QPF exceeding flash flood guidance (QPF > FFG)

The QPF > FFG neighborhood probabilities were created using 3 and 6 hour flash flood guidance values (see below), and displayed at three ratios: QPF > FFG, QPF > 90% of FFG, and QPF > 75% of FFG. The 90% and 75% FFG ratios were developed in order to evaluate ways to further account for the known uncertainty in the RFC flash flood guidance.

In addition to the neighborhood probabilities from the ExREF and modified SSEO, probabilities for several flash-flood and heavy rain parameters were provided from the NCASE. These include probabilities of precipitable water exceeding certain thresholds (1", 1.5", etc.), mean layer winds less than 5 m s^{-1}, and simulated radar reflectivities exceeding certain thresholds (e.g. 40 dBZ).

Additional Guidance

The CIMSS Nearcast model uses information from the GOES-13 water vapor channels to identify areas that might be susceptible to convection and heavy rainfall. The model uses winds from the Rapid Refresh model (RAP) to transport temperature and moisture that is measured from the GOES-13 sounder; this provides information on areas with changing values of precipitable water, as well as areas of increasing/decreasing theta-e with height (e.g. atmospheric stability). The system is updated hourly, and provides data in half-hour intervals out to nine hours.

Lastly, gridded flash flood guidance (FFG) for the CONUS was available, containing flash flood guidance values for 1, 3 and 6 hour timeframes. Local FFG is produced by individual NWS River Forecast Centers (RFC), and WPC compiles the guidance to create a 5 km CONUS mosaic FFG grid. Since RFCs can update FFG at their discretion, WPC checks hourly for any new guidance and recompiles the mosaic. There are several methods currently employed to create FFG; therefore, the method and timing of the production of FFG is inconsistent across RFCs.

Daily Activities

Each week, participants were grouped with a different WPC MetWatch forecaster to form a collaborative forecast team. Each day, the team was tasked with completing several different experimental forecast exercises, which aimed to simulate the timeframe, workload and thought processes associated with creating WPC's MPD and Day 1 Excessive Rainfall products. Unlike their operational counterparts, all experimental forecasts in FFaIR employed the neighborhood approach and were defined as the probability of flash flooding occurring with 40 km of a point.

To start each day, the team generated a CONUS-wide 18 – 12 UTC 'flash flood outlook,' which asked forecasters to assign probabilistic risks of flash flooding, keeping in the spirit of WPC's ERO. In the afternoon, the team was tasked with creating two smaller-scale probabilistic flash flood forecasts similar to WPC's MPDs. For these forecasts, the team was asked to identify a multi-state region where the risk of flash flooding was assessed to be the highest, then create a probabilistic flash flood forecast for the given 6 hour period. A detailed version of the daily schedule can be found in Appendix B.

18-hour (18 – 12 UTC) flash flood outlook (FFO), due at 1330 UTC. Participants were asked to draw contours of 2%, 10% and 30% probability of flash flooding within 40 km of a point, when applicable, over the entire CONUS. This forecast was similar in scope to WPC's ERO, but examined the applicability of the neighborhood approach and used different probability thresholds for each risk category.

6-hour (18 – 00 UTC) probability of flash flooding (PFF1), due at 1800 UTC. This forecast mirrored elements of WPC's MPD, except participants were instructed to draw contours of a 10%, 30% and 50% probability of flash flooding occurring within 40 km of a point, when applicable, over their chosen area of interest (Fig. 2b). This required the forecast team to consider both hydrologic and meteorological information to assess the flash flood threat to issue a forecast for the likelihood of flash flooding.

6-hour (00 – 06 UTC) probability of flash flooding (PFF2), due at 2000 UTC. This forecast also mirrored elements of WPC's MPD, except the forecast team was required to submit their forecast with a longer lead time (4 hours). Participants were again instructed to draw contours of a 10%, 30% and 50% probability of flash flooding within 40 km of a point, when applicable, over their chosen area of interest (Fig. 2c).

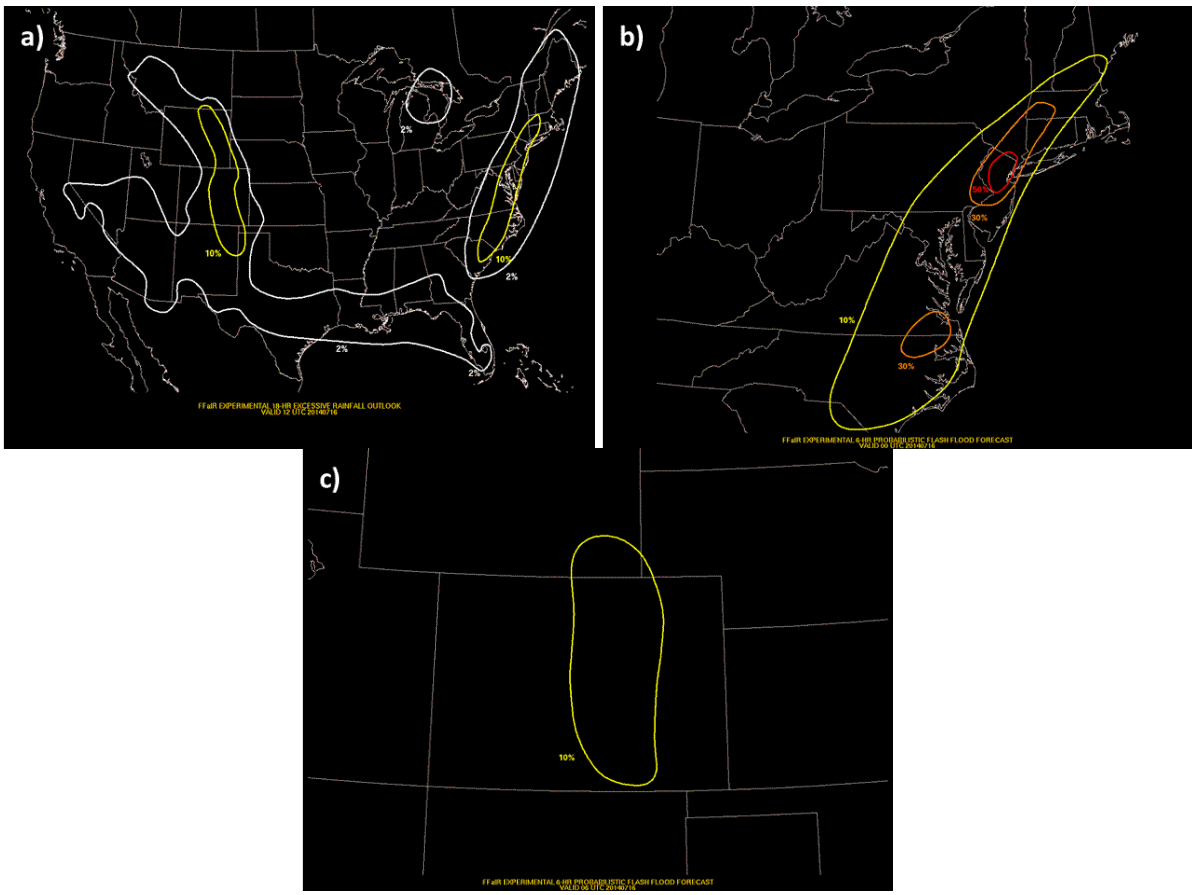


Figure 2. Displaying the (a) 18-hour flash flood outlook forecast, (b) the 18-00 UTC probability of flash flooding forecast, and (c) the 00-06 UTC probability of flash flood forecast for exercises from July 15, 2014.

In addition to issuing these experimental forecasts, participants were also asked to prepare a forecast discussion explaining their forecast rationale and highlighting the major areas of concern across the country. This discussion was then used to provide a daily forecast briefing to participants in the Hazardous Weather Testbed's (HWT) Experimental Warning Program (EWP), which was conducting a concurrent flash flood forecasting exercise.

Finally, participants were also asked to subjectively evaluate their experimental forecasts and the corresponding experimental model guidance and forecast tools. The subjective evaluations graded the relative accuracy and effectiveness of the experimental forecasts and model guidance by using a combination of flash flood indicators, including radar-estimated QPE from the Multi-Radar/Multi-Sensor (MRMS) system, flash flood warnings (FFWs), areas of QPE-to-FFG exceedance, local storm reports (LSRs) of flooding and/or flash flooding, stream flow data from the U.S. Geological Survey (USGS), and National Severe Storms Laboratory (NSSL) Meteorological Phenomena Identification Near the Ground mPING reports.

3. CASES

Like much of the summer of 2014, the experiment period was characterized by a pronounced ridge over the western U.S. and trough over the eastern U.S. (Fig. 3a). This pattern allowed monsoonal moisture to advance well into central and northern California, while precipitable water values remained anomalously low across the eastern two thirds of the country (Fig. 3b). This resulted in an abundance of flash flood events across the desert southwest and intermountain west during the first two weeks of the experiment. In addition, a slow moving cold front during the second week of the experiment led to a multi-day flash flood event stretching from Texas to the Northeast. A complete list of the events that were investigated during this year's experiment can be found in Table 4.

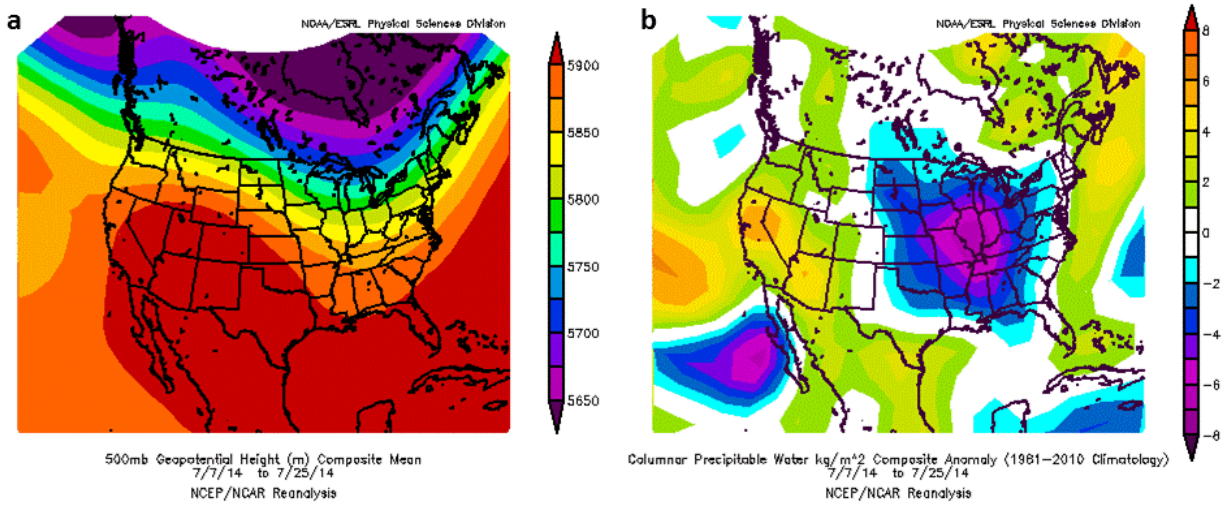


Figure 3. (a) Composite mean 500 hPa heights and (b) precipitable water anomalies during the 7-25 July 2015 period. Images generated from the NCEP/NCAR Reanalysis provided by NOAA/ESRL/Physical Sciences Division (http://www.esrl.noaa.gov/psd/data/composites/day/).

Table 4. Experimental forecasts issued during the 2014 Flash Flood and Intense Rainfall Experiment.

Forecast Valid Date	Valid Time	Forecast Area	Notes
8 July 2014	00 UTC	Southwest	
	06 UTC	Central Plains to Mid-Mississippi Valley	
9 July 2014	00 UTC	Tennessee Valley	Widespread flash flood reports across western AZ and southwest TN
	06UTC	Tennessee Valley	
10 July 2014	00 UTC	Intermountain west to Southwest	
	06 UTC	Intermountain west to Southwest	
11 July 2014	00 UTC	Mid Atlantic	
	06 UTC	Northern Plains to Upper Mississippi Valley	
12 July 2014	00 UTC	N/A	
	06 UTC	Central Plains	
15 July 2014	00 UTC	Mid Atlantic to Northeast	Water rescues from cars and homes in eastern PA
	06 UTC	Southern Rockies	
16 July 2014	00 UTC	Mid Atlantic to Northeast	Urban street flooding along east coast; flash flooding of canyons with debris across roads in southwest
	06 UTC	Central Rockies to Central Plains	
17 July 2014	00 UTC	Central and Southern Rockies to Southwest	
	06 UTC	Central Rockies to Southern Plains	
18 July 2014	00 UTC	Southern Plains	Widespread flash flooding across south-central TX
	06 UTC	Southern Plains	
19 July 2014	00 UTC	N/A	
	06 UTC	Lower Mississippi Valley to Southeast	
22 July 2014	00 UTC	Southeast to Mid Atlantic	Cars trapped in flood waters in eastern NC
	06 UTC	Northern Plains to Upper Mississippi Valley	
23 July 2014	00 UTC	Southeast	
	06 UTC	Southwest and Southern Rockies	
24 July 2014	00 UTC	Northeast	
	06 UTC	Ohio Valley	
25 July 2014	00 UTC	Intermountain West and Southwest	
	06 UTC	Lower Mississippi Valley and Southeast	
26 July 2014	00 UTC	N/A	
	06 UTC	Southwest	

4. DETERMINISTIC HIGH RESOLUTION MODEL PERFORMANCE

As part of the subjective evaluation process, participants were asked to rate the QPF guidance provided by each of the deterministic convection-allowing models on a scale of 1 (very poor) to 5 (very good) based on the observed precipitation during the 18 – 00 UTC period. Each model solution was evaluated independently; the models were not ranked from best to worst. In addition, participants were also asked to compare the QPF guidance provided by three consecutive runs of the HRRR. The results below are based primarily on these subjective responses.

Overall, the majority of the high resolution convection-allowing models examined during the experiment provided useful guidance for flash flood forecasts. While the NAMP Nest received the highest subjective evaluation ratings, both the HRRR and the NCEP high resolution window runs (HRW-ARW and HRW-NMMB) also typically provided useful forecast information (Fig. 5). The only model that routinely provided poor forecast guidance was the EMC WRF-NMM, and this was largely due to its known high precipitation bias.

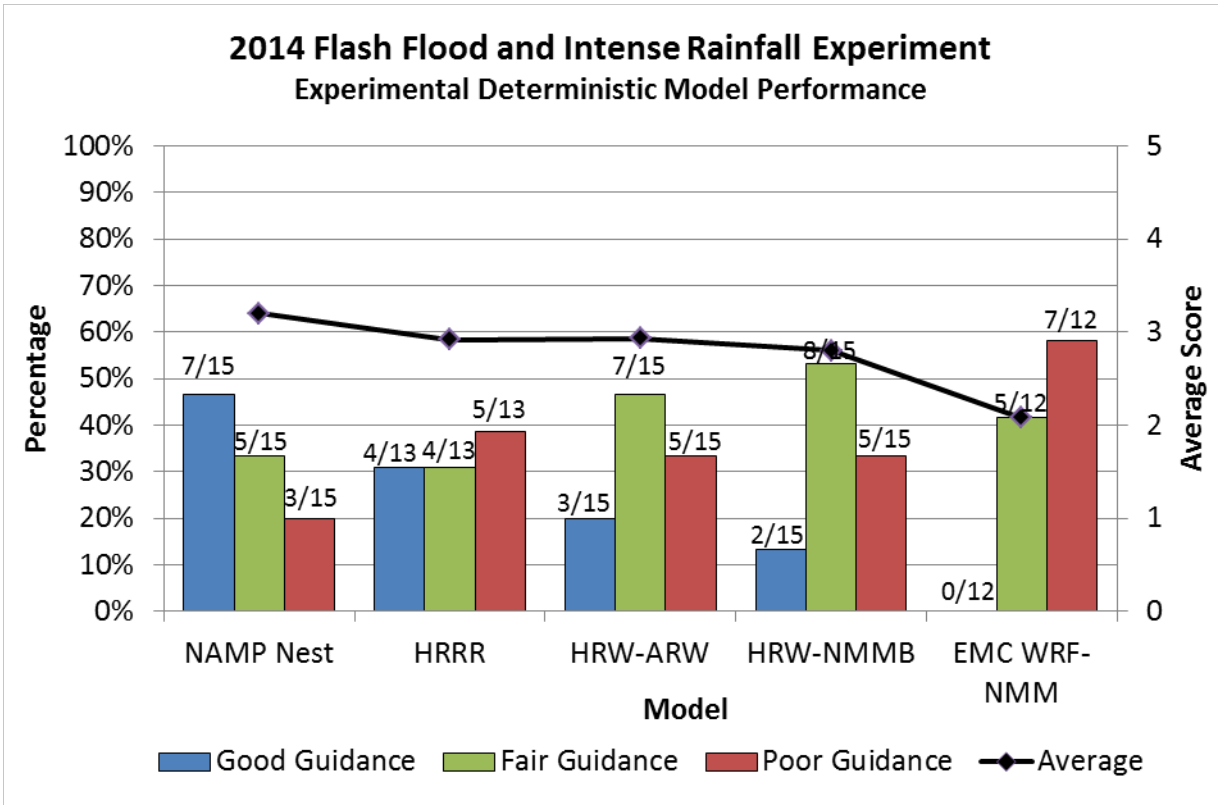


Figure 5. Experimental deterministic model performance based on feedback from subjective model evaluations conducted during the 2014 Flash Flood and Intense Rainfall Experiment. Participants were asked to rate the performance of each model on a scale of 1 (very poor) to 5 (very good).

Although the NAMP Nest received the highest overall ratings, on several occasions participants noted that it had a tendency to produce very high precipitation maxima. For example, during the 6 hour period ending 00 UTC 17 July 2014, the NAMP Nest generated numerous precipitation maxima in excess of five inches across eastern Arizona and western New Mexico (Fig. 6). In contrast, the highest value produced by any of the other high resolution models was 3.30 in in the HRW-ARW, with most values falling in the 1.5 – 2.5 in range. The maximum radar-

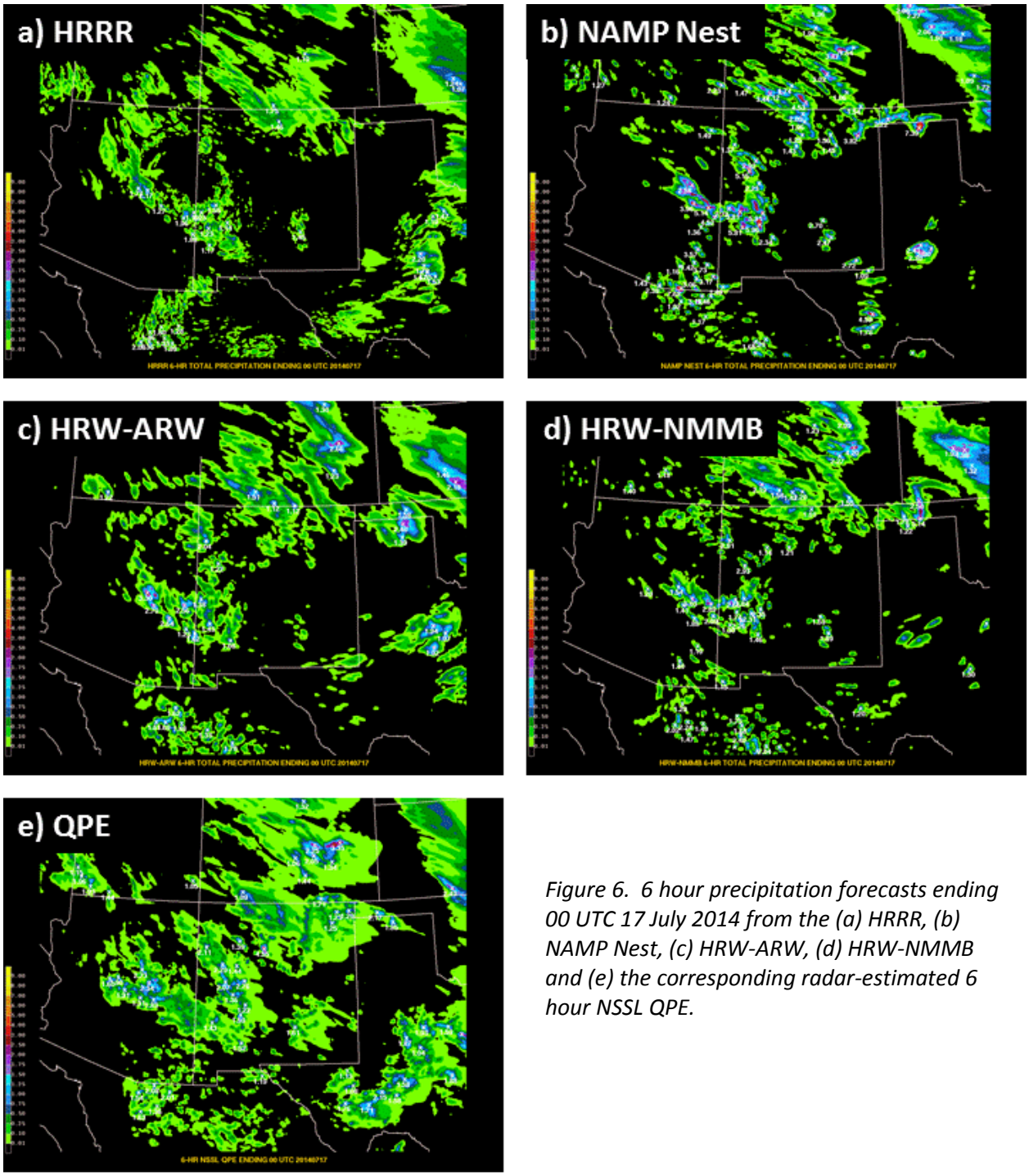


Figure 6. 6 hour precipitation forecasts ending 00 UTC 17 July 2014 from the (a) HRRR, (b) NAMP Nest, (c) HRW-ARW, (d) HRW-NMMB and (e) the corresponding radar-estimated 6 hour NSSL QPE.

estimated QPE across the region was similar, with values of 2 – 3 in. Other than the high point maximum precipitation values, the overall pattern of the NAMP Nest QPF guidance was in line with the other models.

While the HRRR was found to be the most useful high resolution deterministic model during the 2013 experiment, it was less effective during this year's experiment. Even though its overall precipitation pattern was still generally realistic, participants consistently noted that the precipitation amounts from the HRRR appeared to be too light, which often led to the HRRR receiving slightly lower scores in the subjective evaluation process. This apparent dry bias was particularly pronounced in the western U.S., where the HRRR struggled to produce isolated precipitation maxima with the same magnitude as the other models.

Beyond the emergence of a potential dry bias, one of the persistent questions about the HRRR in the operational meteorology community has been the utility of its hourly updates. In order to help address this, participants were asked to compare the QPF guidance provided by consecutive runs of the HRRR (13 UTC compared to 12 UTC) as well as runs separated by two hours (14 UTC compared to 12 UTC). Overall, the later runs of the HRRR were generally found to be quite similar to the initial 12 UTC run, with the 14 UTC run tending to provide slightly more improvement than the 13 UTC run (Fig. 7). Most participants had previous experience using the HRRR during developing weather events in their local offices, and many suggested that identifying trends in the HRRR guidance was more valuable than focusing on the specifics of each new model solution.

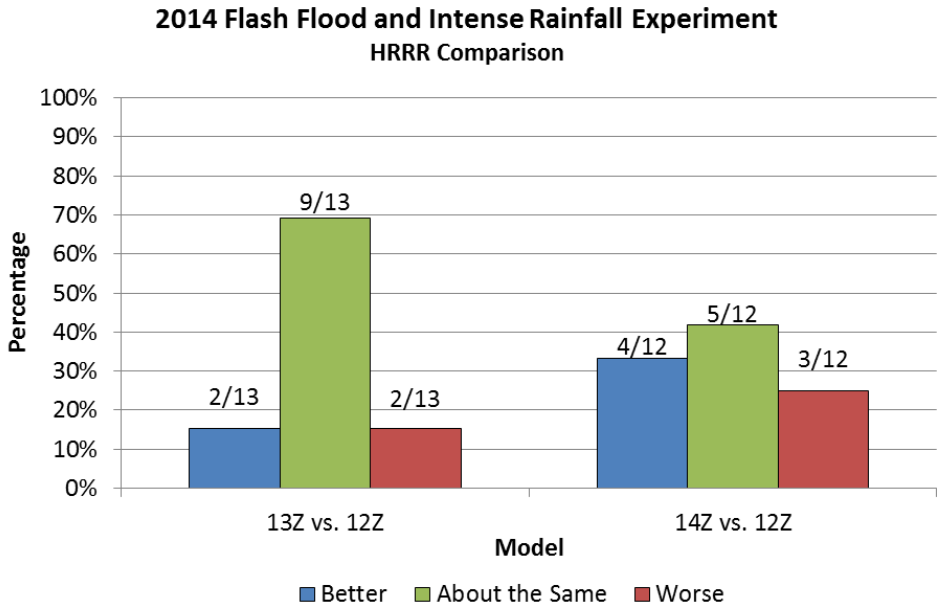


Figure 7. Comparison of HRRR model performance based on feedback from subjective evaluations conducted during the 2014 Flash Flood and Intense Rainfall Experiment. Participants were asked to determine whether the 13 UTC (14 UTC) run of the HRRR provided guidance that was better, worse, or about the same than the guidance provided by the 12 UTC run for 6 hour QPF during the 18 – 00 UTC period.

Finally, there were several events during the experiment in the western U.S. that provided an opportunity for a more focused examination of numerical model performance across this part of the country. This examination suggests that high resolution guidance does not necessarily provide an advantage in the western U.S. In fact, while the sample size is small, there was a distinct difference in the subjective evaluations of model performance for cases in the western U.S. when compared with those east of the Rockies. For the four western cases where the high resolution data was specifically examined, the NAMP provided the best guidance, as it did not receive a 'poor' rating for any of the cases. The largest drop-off in performance in the west was by the HRRR, which did not receive a 'poor' rating in the 9 evaluations that took place east of Rockies, but provided a 'poor' forecast in 3 of the 4 (75%) of the evaluations that took place in the west.

Given the highly variable topography that characterizes this part of the country, a 4 – 5 km resolution model should be better able to represent these differences and the resulting impacts on atmospheric flow than the standard 20 – 30 km scale of most operational models. Instead, high resolution model QPF can appear very scattered, even over time scales of 12 hours or more. This often has the effect of minimizing the perceived threat of heavy precipitation. For example, Figure 8 shows a comparison of the 18 hour QPF from the NAMP, NAMP Nest, and HRW-ARW. While the NAMP shows the potential for widespread light to moderate precipitation over much of the desert southwest, both the HRW-ARW and the NAMP Nest depict a much more scattered precipitation event. While some of the detail provided by these high resolution runs is likely helpful, including the higher point maxima, the character of the observed precipitation ultimately appears to be much more similar to the broader-scale guidance provided by the NAMP.

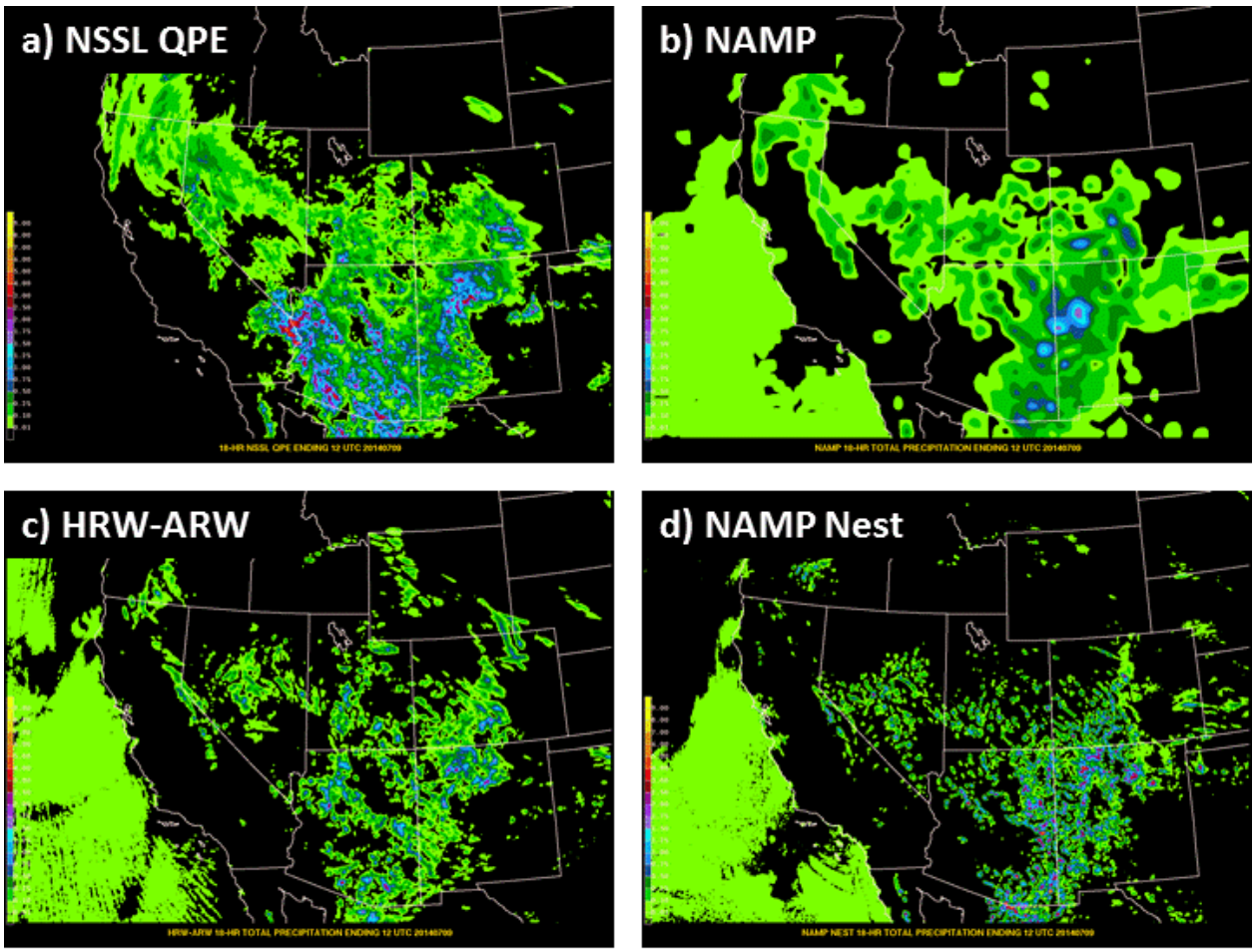


Figure 8. 18 hour (a) radar-estimated NSSL QPE, (b) NAMP forecast, (c) HRW-ARW forecast, and (d) NAMP Nest forecast for the southwestern U.S. valid 12 UTC 9 July 2014.

5. EXPERIMENTAL ENSEMBLE PERFORMANCE

The ensemble guidance systems and corresponding experimental ensemble probabilistic forecast tools were also subjectively evaluated during experiment operations. When assessing the performance of each system's mean 6 hour QPF during the 18 – 00 UTC forecast period, the SSEO, NCASE, ExREF and modified SSEO ensemble mean were subjectively rated as *very poor* (1), *poor* (2), *fair* (3), *good* (4) and *very good* (5). Participants were asked to score each ensemble mean QPF (1-5) on the quality of guidance they gave the forecaster, independent of the other three models (e.g. not ranking the models from best to worst).

As seen in Figure 9, the SSEO and modified SSEO consistently provided the best mean QPF guidance to forecasters. These two ensembles were able to reliably identify the areas where heaviest precipitation would fall, and the 4 km resolution often provided helpful details to forecasters, despite being a mean QPF. While the location of the precipitation was generally reasonable, participants often noticed that the NCASE produced a mean QPF field that was too light overall and did not provide a helpful signal of the details of the precipitation field; this is

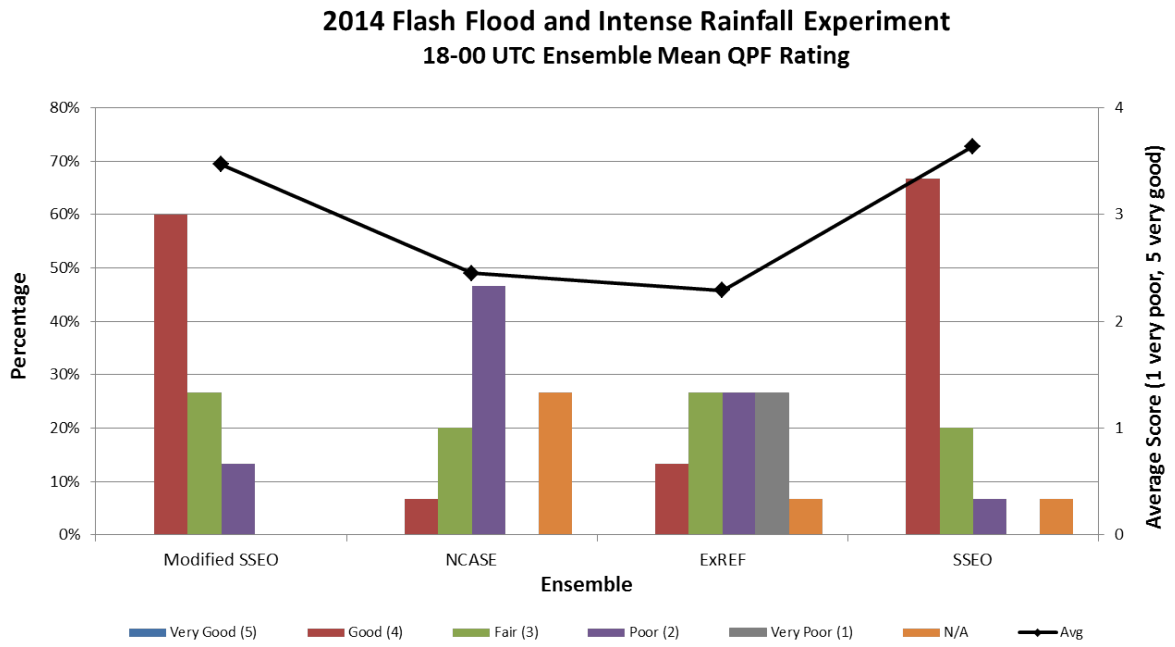


Figure 9. Showing the percentage of 6 hour (18-00 UTC) QPF from the modified SSEO, NCASE, ExREF and SSEO ensembles that were subjectively rated as very poor (gray), poor (purple), fair (green), good (red), very good (blue) or not available (orange) during the 2014 FFaIR experiment. The average score (out of 5, 1 = very poor and 5 = very good) for each ensemble system for the experiment is denoted by the black line.

likely due to it containing 15 members, many of them time-lagged, which worked to 'smooth out' the resulting QPF field. In regards to the ExREF, participants noted that its decreased resolution (9 km) was a detriment, and combining that with its parameterized convection scheme, it struggled to generate small scale or weakly forced convection (e.g. away from surface boundaries, no mid-level trough support, etc.). This tended to result in low QPF amounts in areas of weakly convection and erroneous spatial distribution of the heaviest precipitation.

An example of this can be seen in Figure 10, which shows the 6 hour (18-00 UTC) observed precipitation from NSSL QPE, as well as 6 hour mean QPFs from the SSEO, NCASE, ExREF and modified SSEO valid at 00 UTC on 22 July. While both versions of the SSEO (Fig. 10b,e) were too heavy with precipitation in North Carolina and did not extend heavier precipitation farther west into central South Carolina , participants did like that they both identified a signal for heavy rainfall in northeastern SC and picked up on the rainfall in eastern Georgia. This differs from the performance of the NCASE (Fig. 10c), which was too light overall, largely missing most heavy rainfall in SC. While the light precipitation in the NCASE is thought to be due to the number of members used in the ensemble, the light precipitation in the ExREF (Fig. 10d), is thought to be due primarily to the weakly forced nature of the convection.

Participants were also asked to evaluate the performance of a variety of ensemble neighborhood probability forecast tools in regards to their ability to help forecasters identify

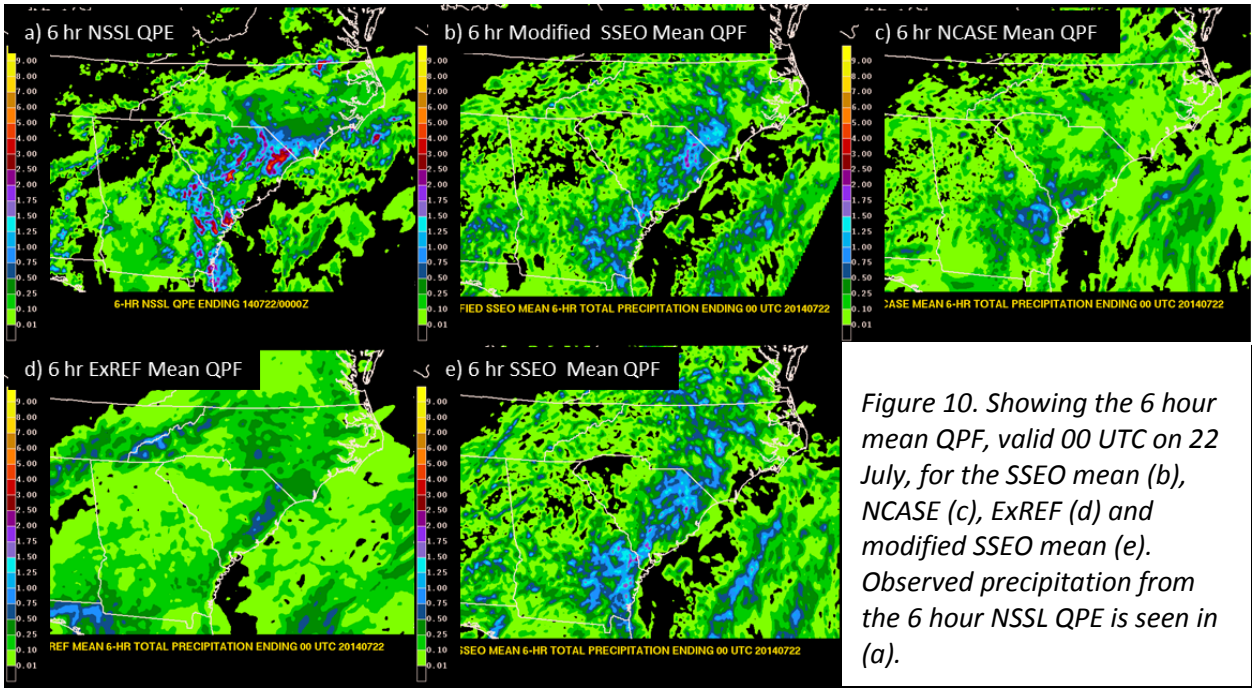


Figure 10. Showing the 6 hour mean QPF, valid 00 UTC on 22 July, for the SSEO mean (b), NCASE (c), ExREF (d) and modified SSEO mean (e). Observed precipitation from the 6 hour NSSL QPE is seen in (a).

areas at risk for flash flooding. Figure 11 shows an example of the neighborhood probabilities of 6 hour QPF exceeding three 6 hour FFG thresholds: QPF > 100% of FFG (i.e. QPF > FFG, Fig. 11a), QPF > 90% of FFG (Fig. 11b), and QPF > 75% of FFG (Fig. 11c) from the modified SSEO, valid at 06 UTC on 19 July. Output featuring these three different FFG thresholds from both the modified SSEO and the ExREF was compared to plots of flash flood observations/warnings (Fig. 11d), as well as 6 hour QPE (Fig. 11e). This helped evaluators get a sense for not only where flash flooding was observed, but also where heavy rain fell and flash flooding may have occurred but not been reported. The probabilities for each FFG threshold for were then scored, for both the modified SSEO and the ExREF separately, on a scale of 1-3, with *'1' representing the worst guidance, and '3' representing the best guidance*. The same score could be given to multiple probability products.

The results of this analysis are shown in Figure 12. Overall, participants felt that the neighborhood probability of exceeding FFG was a useful product, and was helpful in identifying areas at risk for flash flooding. For the modified SSEO, the QPF > 75% of FFG received the highest percentage of "worst" ratings (Fig. 12a). The lower threshold often resulted in a significant expansion in spatial coverage of lower (<20%) probabilities, which could remove critical small scale details and make it harder to identify exact areas that may be threatened. Additionally, participants noted that the 75% threshold had a tendency to falsely inflate the magnitude of the probabilities, and created higher probabilities in areas that did not receive flooding, such as southeast Colorado in Figure 11c. However, the 75% threshold also tied for the most amount of 'best' ratings; in these cases evaluators liked that the reduced threshold helped identify borderline or isolated risk areas, and the increase in spatial coverage could also help to correct the guidance in cases where the numerical models had errors in QPF location. Subsequently, the QPF > 100% FFG was deemed the best overall product for the modified SSEO

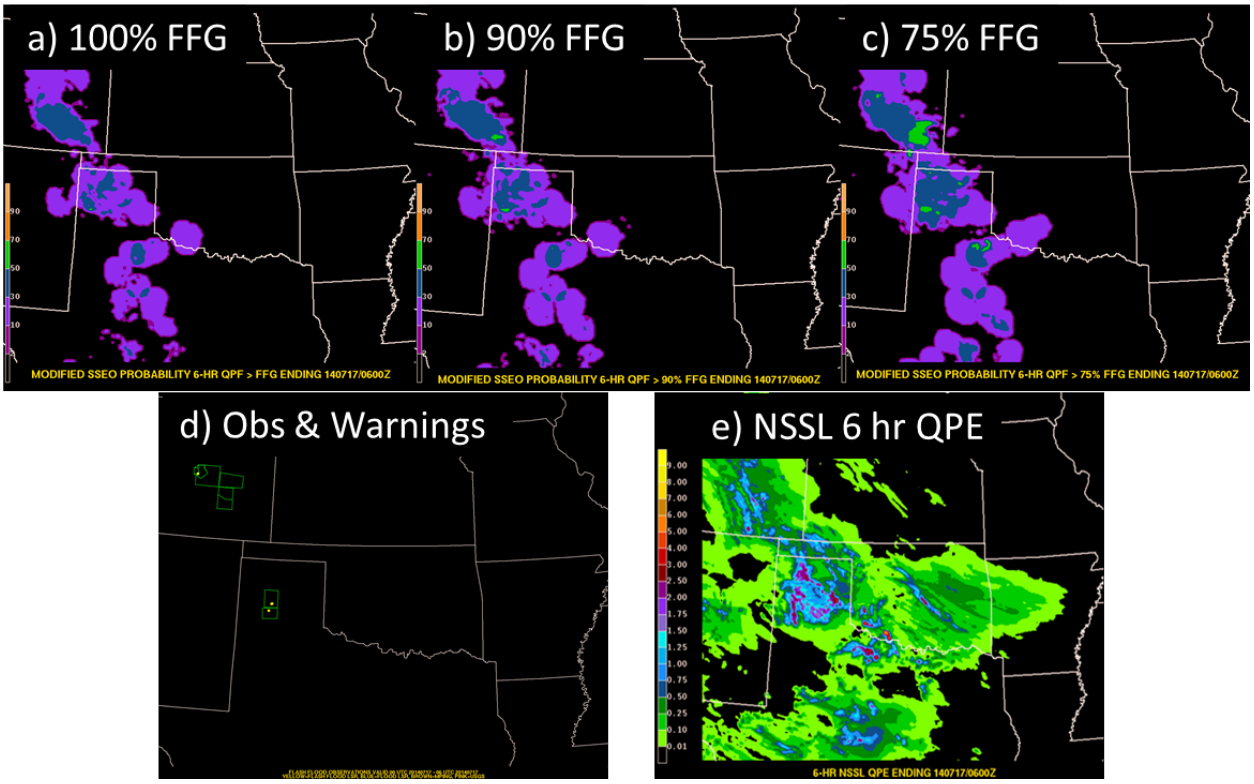


Figure 11. Neighborhood probabilities of 6 hour QPF exceeding three 6 hour FFG thresholds: QPF > 100% of FFG (a), QPF > 90% of FFG (b), and QPF > 75% of FFG (c) from the modified SSEO, valid at 06 UTC on 19 July. Also shown are flash flood reports (yellow dots) and flash flood warnings (green box) in (d), and 6 hour observed precipitation from the NSSL QPE (e).

(Fig. 12a), scoring the highest combination of '2' and '3' ratings, and sharing the highest percentage of 'best' scores. Evaluators thought using FFG as the threshold provided the best balance of spatial coverage and false alarms, and that higher magnitude probabilities tended to match best with areas of highest flash flood risk. Results for the ExREF (Fig. 12b) are less clear. The general perspective was that using the lower thresholds (75% and 90%) provided the best guidance, as the ExREF tended to produce lower QPF amounts overall and therefore had a hard time exceeding FFG values.

Participants were also asked to compare the probability of QPF > FFG for the 00-06 UTC period from both the 18 and 12 UTC cycles of the modified SSEO, to evaluate if generating probabilities every six hours (instead of the default 12 hours) provides additional value. The 18 UTC probabilities were evaluated to give better guidance than the 12 UTC cycle in 40% of the cases, and only evaluated to give worse guidance than the 12 UTC cycle in 13% of cases, supporting the hypothesis that providing updated exceedance probabilities every 6 hours that incorporate both the latest model guidance and more recent flash flood guidance provides value to forecasters.

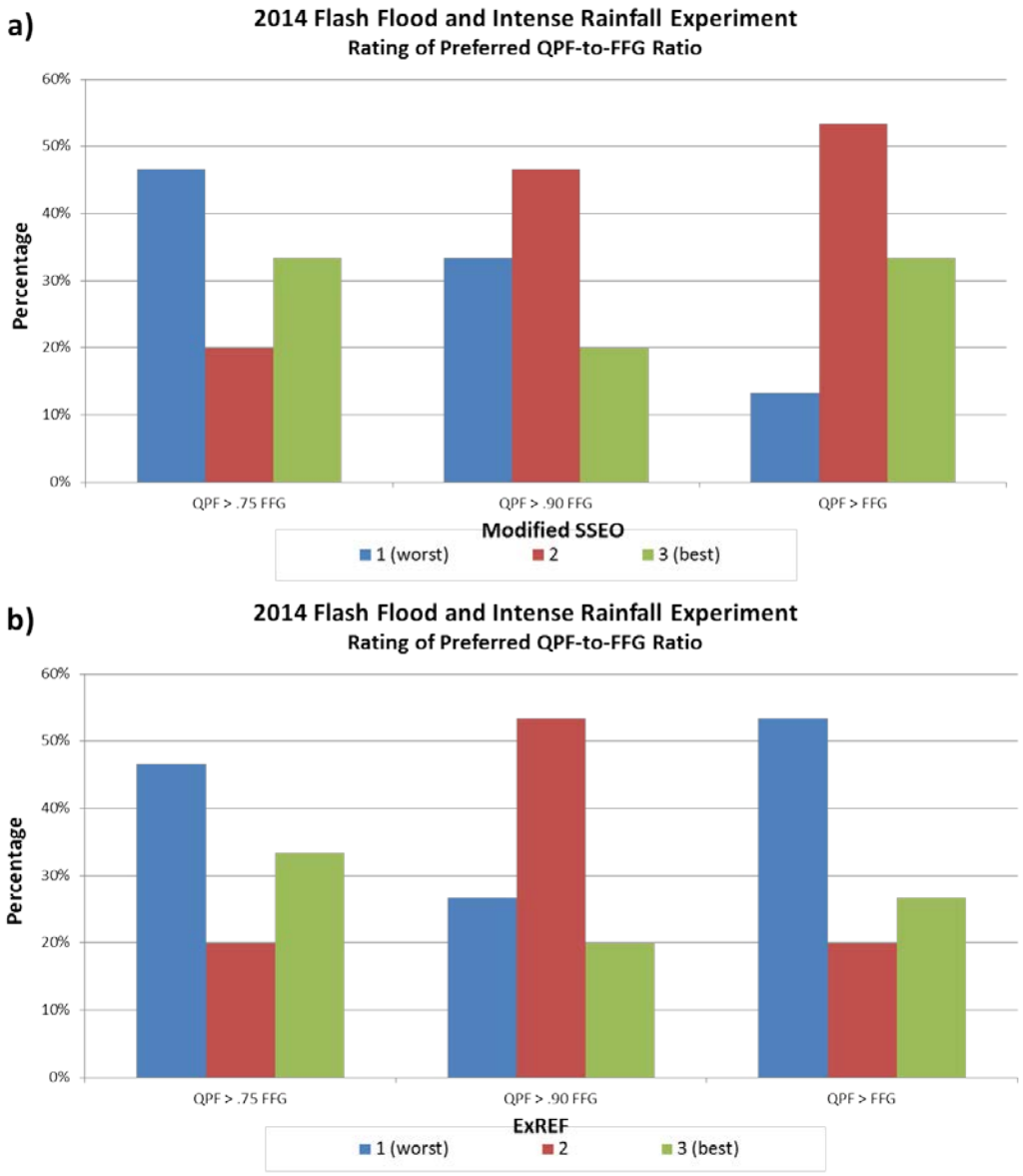


Figure 12. Showing the ranking of the three FFG thresholds used in the creation of 6 hour neighborhood QPF > 6 hour FFG probabilities (18-00 UTC) for both the modified SSEO (a) and ExREF (b). Participants were asked to rate each probabilistic product on a scale of 1-3, with one providing the worst guidance, and three providing the best guidance.

6. THE FLASH FLOOD VERIFICATION DATABASE AND PRACTICALLY PERFECT VERIFICATION TECHNIQUE

As noted during the 2013 experiment, verification presents a significant hurdle to the continued improvement of flash flood forecasts. Without being able to develop a complete understanding of when, where, and how frequently flash flooding is occurring, it is difficult to identify which aspects of the forecast process are in need of improvement. Flash flood forecast verification is made even more challenging by the lack of a comprehensive verification dataset. Observations

of flash flooding are available from a variety of different sources in a variety of different forms, each with a varying level of accuracy.

To address this issue, WPC developed a verification database that collects and archives flash flood observations from several different sources. The database currently includes National Weather Service (NWS) Local Storm Reports (LSRs) of floods and flash floods, NSSL Meteorological Phenomena Identification Near the Ground (mPING) reports, and stream gauge measurements from the U.S. Geological Survey (USGS). LSRs are compiled from reports received by local NWS offices from both the public and trained spotters during and just after significant weather events. While this is typically a high quality dataset, varying definitions of flooding versus flash flooding, differences in reporting approaches, population, event timing, etc. can result in an incomplete depiction of events. mPING reports are collected through a freely available mobile application. The crowd-sourced nature of these reports makes them somewhat less reliable since there is no quality control of the observations, although the potential exists for broad data coverage as usage of the application expands. USGS stream gauges provide both an objective and automated source of information, but reports are limited to locations with gauged streams. Prior to being included in the database, the USGS gauge information is filtered in an attempt to restrict the archived data to observations that are most likely to be the result of flash flooding (basin < 2,000 km^2, rate of rise at least 1 ft hr^{-1} and 3 ft hr^{-1} at consecutive observations, total rise > 2 ft).

In addition to using this database to plot point observations for use during the verification process, the database was also used to explore the utility of the "practically perfect" analysis technique. This technique converts point observations into contoured areas using a Gaussian weighted function (Hitchens et al. 2013; Israel Jirak, personal communication) with the goal of producing the probabilistic forecast a forecaster would have issued had the location of all reported flash flooding been known in advance. At WPC this is done by placing flash flood LSRs, mPING reports, and USGS stream gauge exceedance locations onto a 5 km grid. Once on the grid, any grid cell within 40 km of an observation (consistent with the definition of our experimental forecasts) is assigned a 100% probability of flash flooding. These values are then smoothed over 80 km to yield an approximation of the probability of flash flooding within 40 km of a point. A similar technique has previously been used to evaluate severe weather forecasts at the Storm Prediction Center (SPC).

While the practically perfect output was not the basis of the subjective evaluations completed during the experiment, it was provided to participants as an alternative source of information during the verification process. Figure 13 shows an example of the practically perfect output and the corresponding QPE associated with the experimental excessive rainfall forecast valid 12 UTC 15 July 2014.

Overall, participants found this method of verification promising, but disliked its tendency to produce circular probability areas that often did not match the character of the observed precipitation. This phenomenon can be seen in Figure 14a across Texas, Oklahoma, and Arkansas, where widely spaced individual flash flood observations correspond to individual

probability “bull’s eyes”. In events with abundant observations, such as the area highlighted by the practically perfect analysis in the northern Mid Atlantic, the practically perfect technique produces a much more physically realistic picture of the event.

To help address the issue of probability “bull’s eyes”, participants suggested including additional sources of information in the practically perfect calculation such as flash flood warnings, QPE > FFG, flood LSRs and heavy rain LSRs. Similarly, there were several suggestions to weight the observations differently based on the perceived quality of the source and relevance of the observation type to flash flooding. For example, if flash flood warnings (predicted observation) were added to the practically perfect calculation, they might receive less weight in the probability calculation than a flash flood LSR (actual observation). Another option may be to leverage the QPE information (Fig. 14b) to influence the orientation of the contours drawn by the practically perfect verification.

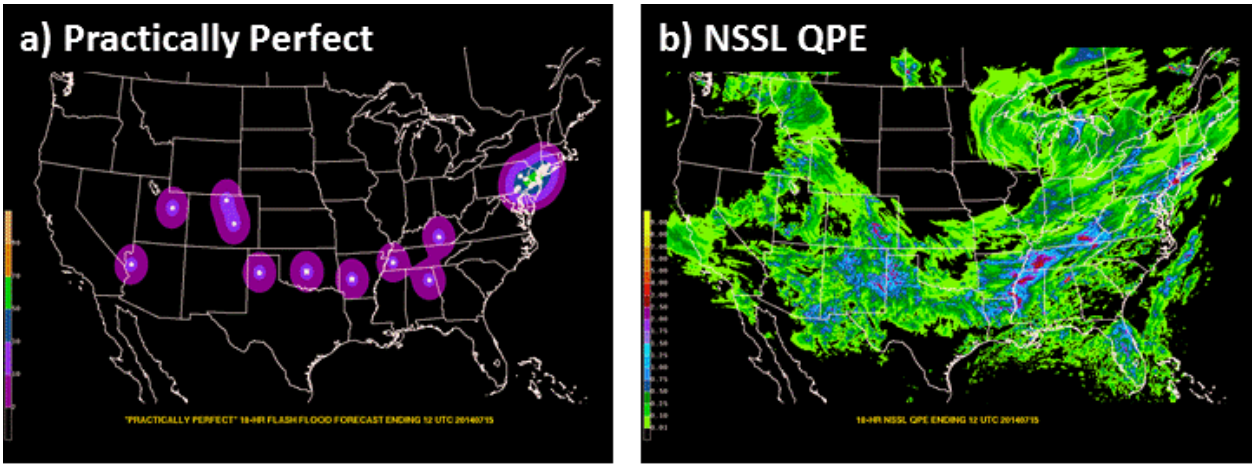


Figure 13. (a) Practically perfect verification output and (b) corresponding QPE for the 18 hour period ending 12 UTC 15 July 2014. The white circles in the practically perfect verification represent flash flood LSRs while the white squares represent mPING reports of flash flooding.

7. EXPERIMENTAL FORECAST PERFORMANCE

Figure 14 shows the results of the subjective verification for the three forecasts made each day: the 18-12 UTC flash flood outlook (FFO), the 18-00 UTC probabilistic flash flood forecast (PFF1), and the 00-06 UTC probabilistic flash flood forecast (PFF2). Each of the forecasts were compared to NSSL’s MRMS QPE, various flash flood observations (FFWs, areas of QPE-to-FFG exceedance, LSRs, USGS stream flow data, mPING reports, etc.), as well as plots of the ‘practically perfect’ forecast, and then subjectively rated as *good*, *fair*, or *poor*.

Overall, participants thought the 18 hour FFO performed well (Fig. 14), with 80% of forecasts receiving either a *good* or *fair* rating. In cases where the forecast was rated as *poor*, there were often multiple LSRs that were not covered by the 2% and/or 10% lines, or the 10% and 30% did not receive heavy precipitation or any observed flash flooding. Evaluators felt that the 2% should capture most, if not all, of the flash flood reports (LSRs), as well as capture any areas of

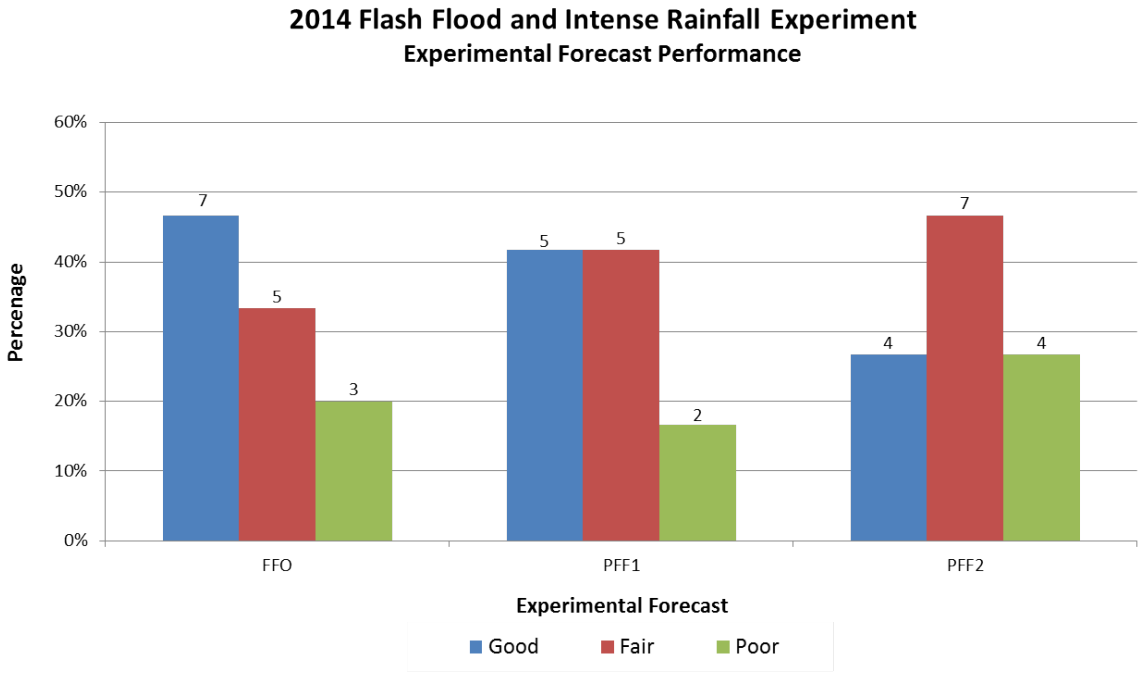


Figure 14. Showing the rating for each of the three experimental forecasts completed in FFaIR: the 18-12 UTC probabilistic flash flood outlook (FFO), the 18-00 UTC probabilistic flash flood forecast (PFF1) and the 00-06 UTC probabilistic flash flood outlook (PFF2). Participants were asked to rate each probabilistic forecast as good, fair, or poor.

heavy rain (usually defined as >1"); any 10% or 30% areas should feature heavy rainfall and have at least one, if not multiple, LSRs and FFWs for those areas to be a considered a successful forecast. There was also recurring discussion about the overall size and usefulness of the 2% contour, as some participants felt that a 2% contour was too lax a threshold which resulted in the contour(s) being drawn over too large an area to be effective. 7 of the 15 FFOs (~47%) in FFaIR featured at least 1/3 of the CONUS area in a 2% risk, often resulting in large areas that received little or no precipitation. However, of the six 30% contours issued in FFOs, four (66%) received at least one FF report in the defined area, with one other having areas that were within 40 km of a report just outside the contour; this suggests that areas that have an elevated risk of flash flooding can be identified accurately at lead times of up to 18 hours.

When considering the 6 hour probabilistic flash flood forecasts (Fig. 14), the 18-00 UTC forecasts[3] (PFF1, 83% rated as *good* or *fair*) performed slightly better than the 00-06 UTC forecasts (PFF2, 73% rated as *good* or *fair*); this difference is most noticeable in the percentage of forecasts rated as *good* (42% to 27%). Table 5a shows that the overall effectiveness of the 10% and 30% contours decreases in PFF2 compared to PFF1, as the percentage of forecasts that were validated with at least one FF observation decreases.

[3] *Due to an altered schedule of activities, the PFF1 forecast was not created on Friday, meaning there were only 12 PFF1 forecasts created in comparison to the 15 FFO and PFF2 forecasts.*

Table 5. (a) Showing the number of 10%, 30% and 50% areas drawn in the probabilistic flash flood forecasts (PFF1 and PFF2) in FFaIR and the corresponding amount of those areas that received at least one flash flood observation, and (b) showing the subjective ratings (Figure 15) of PFF1 and PFF2 forecasts based on whether the forecast area was in the east or west portion of the U.S.

a)

	# of 10% areas	10% areas with reports	Percent	# of 30% areas	30% areas with reports	Percent	# of 50% areas	50% areas with reports	Percent
PFF1	13	6	46%	7	4	57%	2	2	100%
PFF2	16	4	25%	7	2	29%	1	1	100%

b)

East	# of Forecasts	Poor (1)	Fair (2)	Good (3)	Avg Score (out of 3)
PFF1	8	0%	50%	50%	2.5
PFF2	10	10%	50%	40%	2.3

West	# of Forecasts	Poor (1)	Fair (2)	Good (3)	Avg Score (out of 3)
PFF1	4	50%	25%	25%	1.75
PFF2	5	40%	60%	0%	1.6

Table 5b shows the split in forecast performance for forecasts were made in the eastern United States, defined as the eastern 2/3 of the CONUS (east) beginning east of the Rocky Mountains, and the western United States. While the sample size of forecasts in the west is smaller, there is a clear degradation in forecast quality in the western U.S., with a drop in average forecast score and sharp increase in the percentage of 'poor' forecasts for both PFF1 and PFF2 forecasts. This same decrease in forecast quality was also seen in the numerical model guidance (Section 4). Some of this drop in the forecast rating can likely be attributed to verification problems in the west; dispersed population and inconsistent radar coverage make accurate QPE information and flash flood observations difficult to achieve. This can lead to forecasts being downgraded during the subjective evaluation process, as evaluators were often presented with scattered precipitation data and few, if any, observations indicating flash flooding that would verify their forecast. The combination of decreased model skill and the general lack of forecaster familiarity with small scale features and geographical details that lead to flash flooding in the western U.S. results in a decline in forecast skill that needs to become a focused point of improvement.

8. SUMMARY AND OPERATIONAL IMPACTS

The second annual Flash Flood and Intense Rainfall Experiment was conducted from July 7 – 25, 2014 at the NOAA Center for Weather and Climate Prediction in College Park, MD. Over the course of the three week experiment, 23 forecasters, researchers, and model developers used a variety of innovative high resolution model and ensemble output to issue a series of experimental probabilistic flash flood forecasts. While there are still a number of areas that need improvement, this year's experiment once again highlighted the utility of convection-allowing guidance in the flash flood forecast process. In particular, a number of the experiment findings are relevant to operational forecasters focused on the flash flood threat:

- In order to improve flash flood forecasts, **the NWS needs to have a more consistent approach to the flash flood forecast problem, ranging from definition, to warning practices, to reporting.** Lack of clarity regarding what is and is not a flash flood, combined with inconsistent reporting practices from WFO to WFO, leads to a moving target when it comes to identifying where flash flooding has occurred.
- **Convection-allowing ensembles such as the SSEO are invaluable to the flash flood forecast process, especially when paired with hydrologic data.** The neighborhood probability of QPF > FFG again proved to be a useful forecast tool, and the 06 and 18 UTC runs and additional FFG ratios were found to provide additional value to forecasters. In addition, high resolution models were found to provide valuable QPF guidance out to at least 36 hours.
- While numerous challenges remain**, the development of a multi-source flash flood verification database represents a significant step forward in the effort to improve flash flood forecasts.** During the experiment this database was leveraged to explore the utility of the practically perfect analysis technique. While initial impressions of this technique were favorable, there are a number of potential avenues for improvement that would result in a more robust verification tool.
- While high resolution models have proven to be a useful forecast tool east of the Rockies, **flash flood forecasting in the western U.S. represents a more significant challenge.** Improving forecast skill in this region will require both improved model guidance and a focused effort to build forecaster understanding of the factors that govern flash flooding in the complex topography in the western U.S.
- **WPC should continue to pursue changes to the current Excessive Rainfall Outlook product** to highlight broader areas at risk of flash flooding. While further testing will be required to determine whether numerical probabilities or descriptive categories are a more effective approach to conveying the flash flood risk, during the experiment the forecast teams were consistently able to successfully distinguish between areas with a risk for flash flooding and those without.
- The daily forecasts briefings provided useful information to the EWP's experimental flash flood watch and warning activities, however **continued effort needs to be made to maximize the benefit of cross-testbed interactions.**
- RFC and WFO participants suggested several hydrologic datasets that can be used as an alternative to flash flood guidance. **WPC will work to make these datasets available to forecasters and consider integrating them into the neighborhood probability technique.**

The Flash Flood and Intense Rainfall Experiment provided a unique opportunity to gather participants with varying backgrounds together to discuss the challenges associated with flash flood forecasting. The results of this experiment will help guide the continued development of WPC's MetWatch Desk as well as the proposed changes to the operational Excessive Rainfall Outlook. In the coming months, HMT-WPC will work to implement the lessons learned, including investigating the proposed alternatives to flash flood guidance and refining the practically perfect verification technique.

ACKNOWLEDGEMENTS

The Flash Flood and Intense Rainfall Experiment would not have been possible without the dedication of a host of individuals including Faye Barthold (HMT-WPC), Tom Workoff (HMT-WPC), Brian Cosgrove (OHD), Wallace Hogsett (WPC/Weather Analytics), JJ Gourley (NSSL), and Dave Novak (WPC). In particular, while on an NRAP assignment at WPC Brian Cosgrove was instrumental in the development of the flash flood verification database and implementation of the practically perfect verification technique. Ligia Bernardet (ESRL) and Craig Hartsough (ESRL) provided access to the ExREF and VLAPS HIRES during the experiment, while Binbin Zhou (EMC) provided the NCASE. WPC forecasters Rich Otto, Patrick Burke, and David Roth helped lead participants through the experimental forecast process.

REFERENCES

Ebert, E.E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 53-66.

Hitchens, N.M., H.E. Brooks, and M.P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting,* **28**, 525-534.

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. Sever Local Storms, Nashville, TN. Amer. Meteor. Soc. P9.137.

Schwartz, C.S., and Coauthors, 2009: Next-day convection-allowing WRF model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev*., **137**, 3351-3372.

Schwartz, C.S., and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263-280.

Zhuo, B., J. Du and G. DiMego, 2014: Aviation, Convection, Energy and Fire-Weather Ensemble Products in NCEP's Regional Ensemble Forecast Systems**.** 6th NCEP Ensemble Users Workshop, College Park, MD.

APPENDIX A
Participants

Week	WPC Forecaster	WFO/RFC	Research/Academia	EMC
July 7 – 11	Rich Otto	Patrick Ayd (BIS) Ron Horwood (NERFC) Scott Lincoln (LMRFC)	Russ Schumacher (CSU) Greg Herman (CSU)	Matt Pyle
July 14 – 18	Patrick Burke	Jason Elliot (LWX) Ryan Husted (GLD) Randy Graham (SLC) Greg Waller (WGRFC)	Kelly Mahoney (ESRL)	Eric Aligo
July 21 – 25	Dave Roth	Lara Pagano (MHX) Alex Tardy (SGX) Andrew Pritchett (MRX) Andy Dean (SPC)	Ellen Sukovich (ESRL) Craig Hartsough (ESRL)	Jacob Carley Geoff Manikin

APPENDIX B
Daily Schedule

8:00am – 9:30am Use 00 UTC guidance to issue an experimental Excessive Rainfall Outlook valid 18 – 12 UTC. Contours will indicate a 2%, 10%, and 30% probability of flash flooding within 40 km of a point. Prepare forecast discussion.

9:30am – 11:00am Subjective evaluation of yesterday's experimental forecasts and model guidance.

11:00am – 11:30am WPC-CPC map discussion

11:30am – 12:30pm Lunch

12:30pm – 2:00pm Use 00 UTC and any available 12 UTC guidance to issue an experimental probabilistic flash flood forecast valid 18 – 00 UTC. Contours will indicate a 10%, 30%, and 50% probability of flash flooding within 40 km of a point. Update forecast discussion.

2:00pm – 2:45pm EWP forecast briefing

2:45pm – 4:00pm Use 12 UTC guidance to issue an experimental probabilistic flash flood forecast valid 00 – 06 UTC. Update forecast discussion.

4:00 – 4:30pm Group discussion