2018 Flash Flood and Intense Rainfall Experiment

June 18 - July 20, 2018

Weather Prediction Center

College Park, MD

Findings and Results

Benjamin Albright *- Systems Research Group, NOAA/NWS/WPC/HMT*
Sarah Perfater *- I.M. Systems Group, NOAA/NWS/WPC/HMT**
**Current affiliation: Cherokee Nation Business, NOAA/OAR/OWAQ*

Table of Contents

1. Introduction

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

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

The 2018 experiment provided a real-time pseudo-operational environment in which participants from across the weather enterprise combined expertise to explore the utility of emerging model and ensemble guidance for improving flash flood forecasts. This year's experiment furthered efforts to rapidly incorporate the latest observational and model guidance into the decision making process, while also challenging participants to simulate the collaboration that occurs between the national centers and local forecast offices during flash flood events.

2. Science and Operations Goals

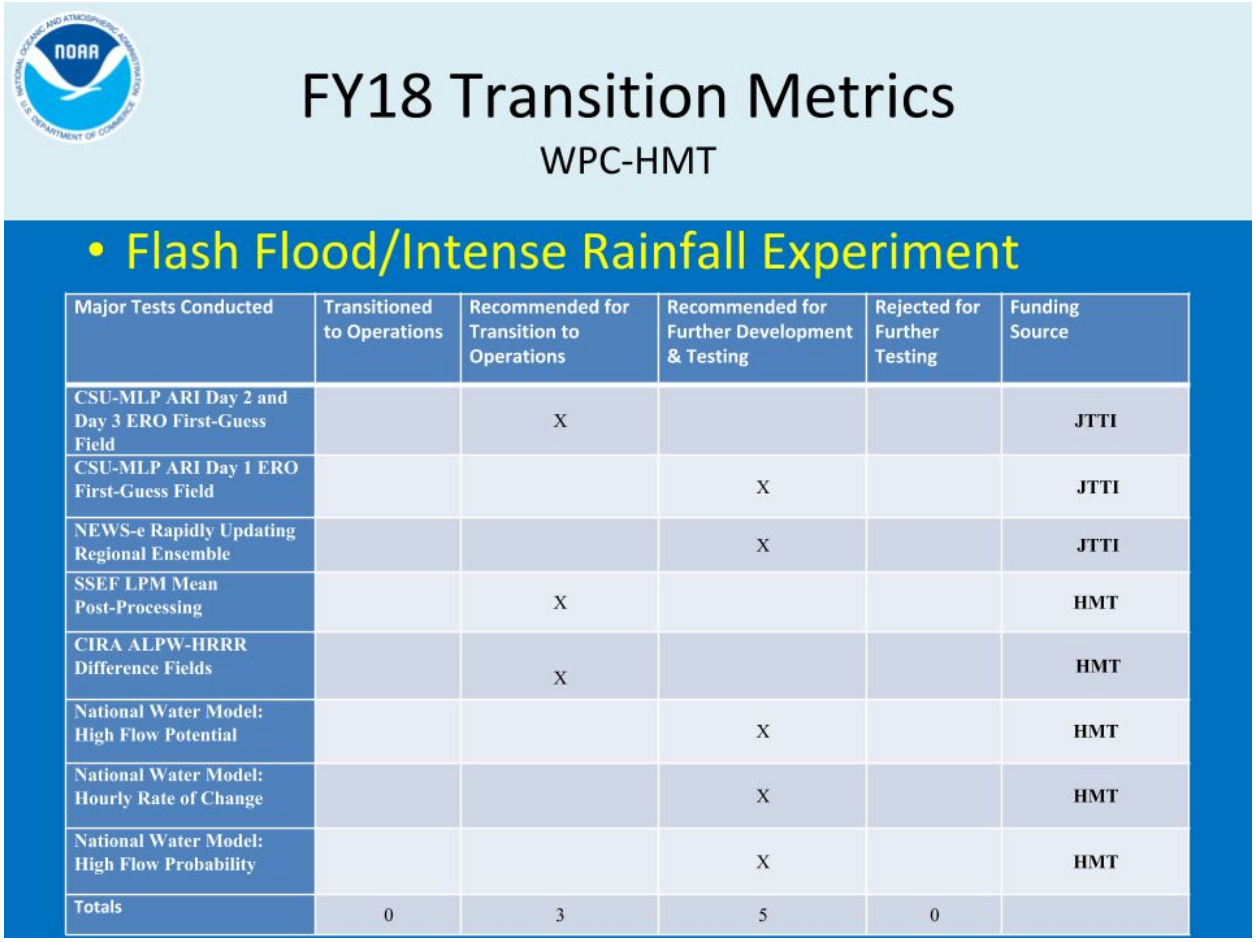
The 2018 experiment focused on the use of high resolution guidance to synthesize atmospheric and hydrologic guidance in an end-to-end forecast process to produce probabilistic flash flood forecasts in the short range (6-24 hours). To simulate the flow of information that occurs from a national center (e.g. WPC) to the local forecast offices, this year's experiment engaged the HMT-Hydro participants and the Science and Operations Officer-Development and Operations Hydrologist (SOO-DOH) community through screen sharing, video, and teleconference to discuss the experimental guidance with the goal of producing a collaborative 6-hour probabilistic flash flood forecast.

The goals of the 2018 Flash Flood and Intense Rainfall Experiment were to:

- Evaluate ways to maximize the utility of high resolution convection-allowing deterministic models and ensembles for short-term flash flood forecasts.
- Evaluate ways to maximize hydrologic guidance for the assessment of flood risk.
- Identify effective methods for synthesizing atmospheric and hydrologic guidance for rapid risk assessment and prediction of flash flooding.
- Identify utility of advanced remotely-sensed products and difference fields.

- Objectively and subjectively evaluate the utility of the Colorado State University Machine Learning Probabilities (CSU-MLP) "First Guess Field" for the Excessive Rainfall Outlook in the Marginal, Slight, Moderate, High Risk categories at Days 2 and 3.
- 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.

Table 1. *Research to Operations Transition Metrics for the 2018 FFaIR Experiment.*



FY18 Transition Metrics

WPC-HMT

• Flash Flood/Intense Rainfall Experiment

Major Tests Conducted	Transitioned to Operations	Recommended for Transition to Operations	Recommended for Further Development & Testing	Rejected for Further Testing	Funding Source
CSU-MLP ARI Day 2 and Day 3 ERO First-Guess Field		X			JTTI
CSU-MLP ARI Day 1 ERO First-Guess Field			X		JTTI
NEWS-e Rapidly Updating Regional Ensemble			X		JTTI
SSEF LPM Mean Post-Processing		X			HMT
CIRA ALPW-HRRR Difference Fields		X			HMT
National Water Model: High Flow Potential			X		HMT
National Water Model: Hourly Rate of Change			X		HMT
National Water Model: High Flow Probability			X		HMT
Totals	0	3	5	0	

3. Experiment Operations

Forecast Activities

The experiment was conducted for four weeks beginning June 18, 2018 in the WPC-OPC Collaboration Room at the NOAA Center for Weather and Climate Prediction (NCWCP) in College Park, MD:

Week 1: June 18 – 22, 2018 (Monday – Friday)

Week 2: June 25 – 29, 2018 (Monday – Friday)
Week 3: July 9 – 13, 2018 (Monday – Friday)
Week 4: July 16 – 20, 2018 (Monday – Friday)

Each morning, participants were paired with a WPC forecaster as part of a collaborative forecast team and were asked to use available experimental guidance to create a Day 1 Excessive Rainfall Outlook (ERO) defined experimentally as the probability of flooding rains occurring within 40 km of a point. The Day 1 ERO was valid over the contiguous United States (CONUS) and valid for a 21-hour period from 15 UTC to 12 UTC, using probability contours of 5% (marginal), 10% (slight), 20% (moderate) and 50% (high) to convey the risk.

After lunch, the participants also created a short-term, Probability of Flash Flooding (PFF1) forecast valid for six hours from 18-00 UTC using probability contours of 10% (slight), 20% (moderate), and 50% (high) that conveyed the likelihood of flash flooding occurring within 40 km of a point. The forecast was created over a limited domain within the CONUS.

Later in the afternoon, the participants used available experimental guidance to create a second short-term, Probability of Flash Flooding (PFF2) forecast valid for the six hours from 00-06 UTC again using probability contours of 10% (slight), 20% (moderate), and 50% (high). And lastly, a PFF3 was created as an update to the PFF1, valid for the three hours from 21-00 UTC and using the NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e) rapidly-updating ensemble information exclusively. The PFF exercises were designed to explore potential improvements to the WPC MetWatch Desk operations by using experimental guidance in a shorter, 6 hour timeframe.

SOO-DOH/HMT-Hydro PFF1 Collaboration and Daily Briefing

Each day an email message was distributed to the SOO/DOH, communities and other associated partners inviting all to join an afternoon teleconference call. Beginning at 12:30 pm EDT, interested partners joined remotely through a Mikogo screen sharing session to view N-AWIPS data and video broadcast of AWIPS data to collaboratively develop the 6-hour PFF1 forecast (valid 18-00 UTC). At 1:30 pm EDT, a conference call was used to showcase a PowerPoint (PPT) presentation that was built throughout the day by the participants. This PPT highlighted the two experimental forecasts (the Day 1 ERO valid 15 - 12 UTC and the 6-hour PFF1 valid 18-00 UTC) and a sampling of the experimental guidance that supported those forecasts. The product creation and briefings were designed to simulate collaboration between a national center and the WFO field offices when developing and communicating flash flood forecasts.

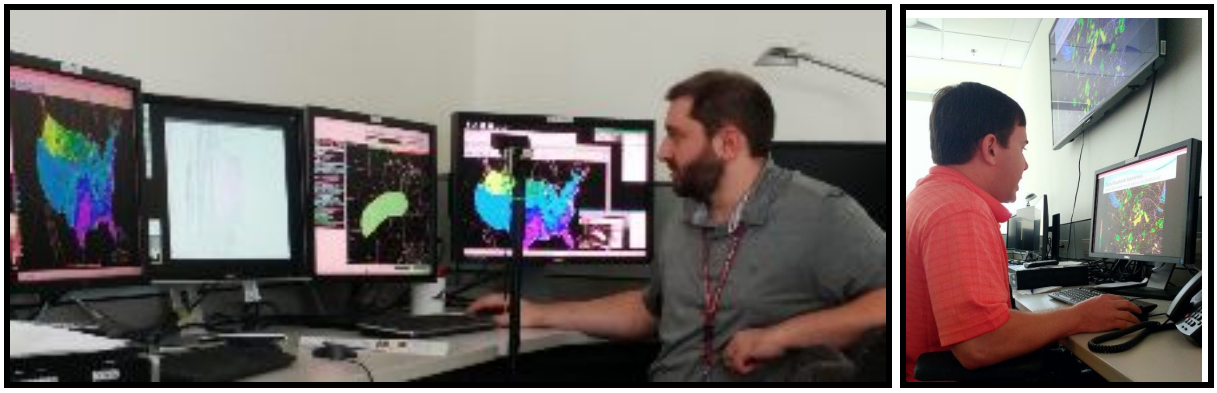


Figure 1. *Left: WPC forecaster, Alex Lamers, utilizes both AWIPS and N-AWIPS to collaboratively create the PFF1 in AWIPS with the participants in the room and those on the teleconference. Right: A FFaIR Experiment participant volunteer delivers the forecast PPT briefing via teleconference call and Mikogo.*

Verification

Each day participants subjectively evaluated eleven science questions presented by the testbed staff. These questions included evaluation of the experimental FFaIR forecasts as well as other experimental models and tools used during the forecasting process. Participants used white boards to rank each experimental guidance, tool, or forecast on a scale from 1 (very poor) to 10 (very good). Individual scores were then recorded each day and used for all statistics. Individual model and ensemble names were removed from the titles of the evaluation graphics so the participants did not know which model they were assigning scores to each day. This was done to attempt to remove preconceived biases that a participant might hold for any particular model. Table 2 shows the FFaIR forecasts, models, or tools evaluated by the science questions and the number of subjective scores each received throughout the experiment. The total number of scores were dependent on both model availability and the number of participants providing scores each day.

Table 2 *An overview of the science questions and associated total subjective scores in parenthesis provided by experiment participants.*

FFaIR Products Evaluated and the Number of Scores for Each Model/Cycle/Parameter Evaluated					
PRODUCT	**MODEL, CYCLE, or PARAMETER EVALUATED** *(Number of Scores)*				
NEWS-E 3HR PMM QPF	**18 UTC** *(192)*	**19 UTC** *(202)*	**20 UTC** *(203)*	**21 UTC** *(204)*	
HREFv2.1 Neighborhood/EAS Probabilities	**Neighborhood 0.5"** *(66*)*	**EAS 0.5"** *(66*)*	**Neighborhood 1.0"** *(67*)*	**EAS 1.0"** *(67*)*	
Day 1 QPF	**FV3-GFS (Model A)** *(214)*	**HRRRv3 (Model B)** *(173)*	**FV3-CAPS Thomp. (Model C)** *(192)*	**FV3-CAPS NSSL (Model D)** *(204)*	**NBMv3.1 (Model E)** *(194)*
Ensemble 6HR PMM	**SSEFX LPM**	**HRRRE**	**HREFv2.1**	**NCAR**	**SSEFX PMM**

QPF	(Ensemble A) *(212)*	(Ensemble B) *(191)*	(Ensemble C) *(135)*	(Ensemble D) *(146)*	(Ensemble E) *(212)*
National Water Model Parameters	**Peak Flow Arrival Time** *(80*)*	**High Flow Probability** *(69*)*			
CSU-MLP First Guess Field	**Day 1** *(201)*	**Day 2** *(192)*	**Day 3** *(192)*		
FFaIR Forecasts	**Day 1 ERO** *(202)*	**PFF1** *(181)*	**PFF2** *(202)*		
**Counts significantly lower due to data outages and transition to collecting comments/feedback only for these evaluations.*					

Quantitative precipitation forecasts (QPF) from the various models and ensembles were verified using Multi-Radar, Multi-Sensor Gauge Corrected (MRMS-GC) quantitative precipitation estimates (QPE) over the forecast period. Figure 2 shows an example image of MRMS-GC QPE on the left compared to "Model B" (HRRRv3) QPF on the right.

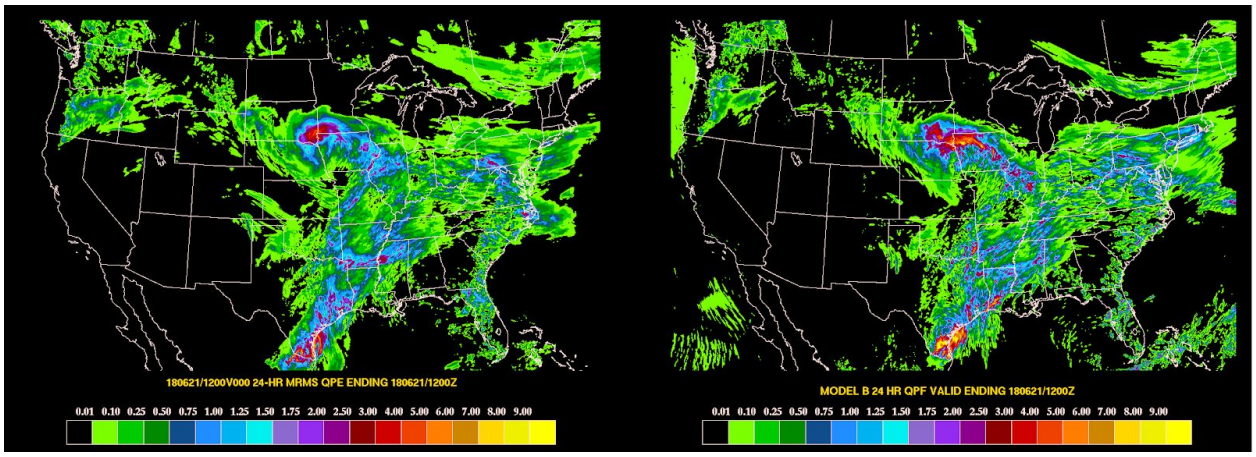


Figure 2. *(Right) MRMS-GC 24 hour QPE valid from 12 UTC June 20 - 12 UTC June 21, 2018 and (left) 24 hour QPF from "Model B" (HRRRv3) valid over the same time.*

Verification of WPC's operational EROs is based solely on 1-, 3-, or 6-hour Stage IV QPE exceeding corresponding 1-, 3, or 6-hour FFG over the CONUS. For the ERO, PFF, and CSU-MLP First Guess Field forecasts issued in FFaIR, the Unified Flood Verification (UFV) system was used. The UFV was used for the first time in the 2017 FFaIR Experiment in an effort to expand beyond just using FFG to verify the experimental forecasts. UFV uses a combination of 1/3/6 hour QPE > FFG, 1/6/24 hour QPE > 5 year average recurrence interval (ARI), and flash flood/flood LSRs and USGS gauge reports. A 40 km radius is then applied to each point considered a hit and all hits are combined and plotted on one map. An example of how a FFaIR ERO forecast was verified in the experiment is shown in Figure 3. The Day 1 ERO, valid 15 UTC June 20 to 12 UTC June 21, 2018 is contoured in panel A and overlaid with the UFV reports (green circles). Panel B in Figure 3 shows the 21 hour QPE valid over the forecast period. Panel C displays the practically perfect analysis which creates a neighborhood probabilistic forecast based on only the flash flood/flood LSRs received during the valid time. Practically perfect uses a 90 km

Gaussian smoother on the LSRs and serves as a representation of what the forecast should have been if the forecaster had prior knowledge of where the reports would be located.

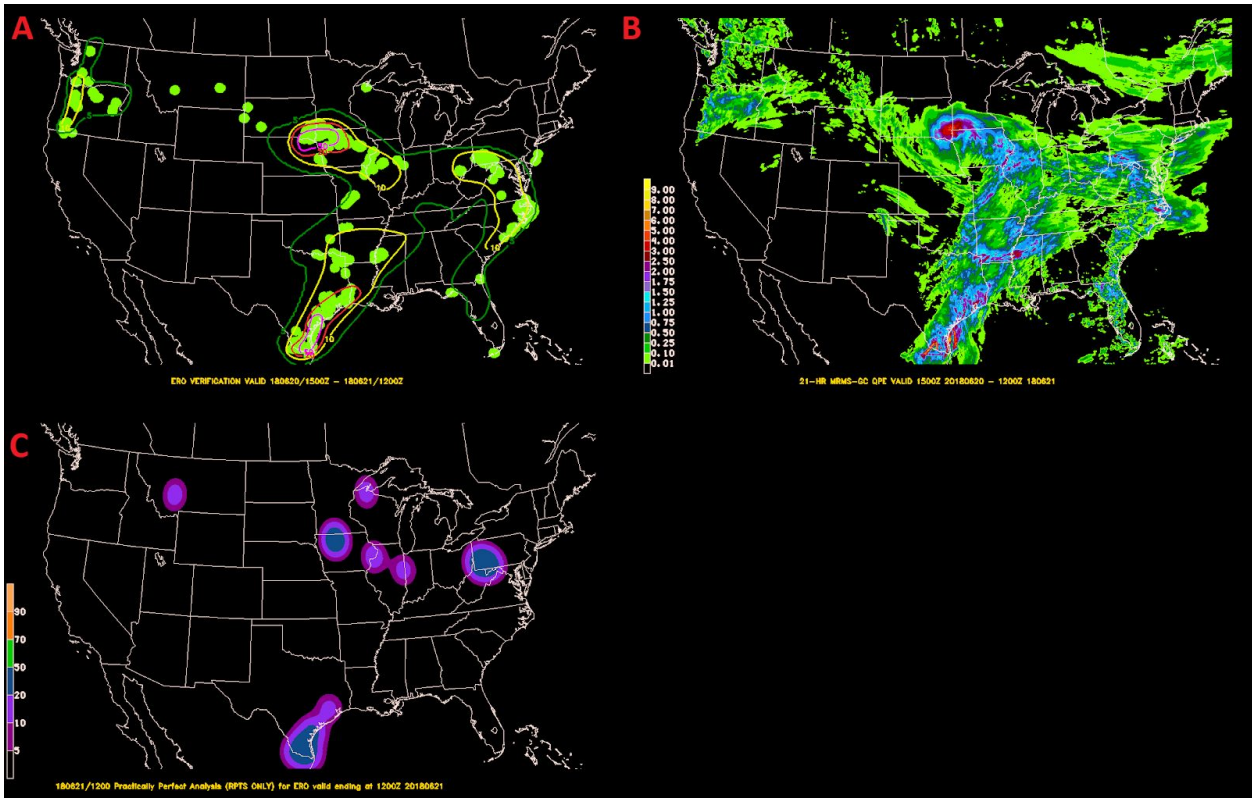


Figure 3. *(A) The Unified Flood Verification system (green dots) which includes 1-, 3-, 6-hour QPE exceeding FFG, 5 year 1-, 6-, 24-ARI exceedance, and flash flood LSRs, flood LSRs, and USGS gauge reports, where all hits have a 40 km neighborhood radius filter applied. The Day 1 ERO valid 15 UTC June 20 to 12 UTC June 21, 2018 is overlaid, (B) 21 hour MRMS-GC QPE, (C) practically perfect analysis based only on flash flood and flood LSRs.*

After the subjective verification, the Method for Object-Based Diagnostic Evaluation (MODE) was used to compare various forecasted QPF thresholds from several models to MRMS-GC QPE (see Appendix C for WPC MODE settings). MODE outputs various statistics comparing the forecasted objects (model QPF) to the observed objects (MRMS-GC QPE) including centroid distance, angle, and intersection area. The Gilbert Skill Score (GSS) and critical success index (CSI), commonly referred to as Equitable Threat Score and Threat score respectively, were also computed over the whole domain for several models. All model QPF and QPE were re-gridded to a common 5 km grid with a CONUS mask applied. An example of the MODE verification for the 36 hour forecast from "Model B" (HRRRv3) of 24 hour QPF at the 1 inch threshold valid at 12 UTC June 21, 2018 is shown in Figure 4. The overall performance of select models were tracked on a daily basis as well as cumulatively throughout the entire experiment using Roebber Performance Diagrams (Roebber, 2009), pictured in Figure 5. A Roebber Performance Diagram provides a way to visualize a number of measures of forecast quality including probability of detection, false alarm ratio, contingency bias, and CSI in a single diagram.

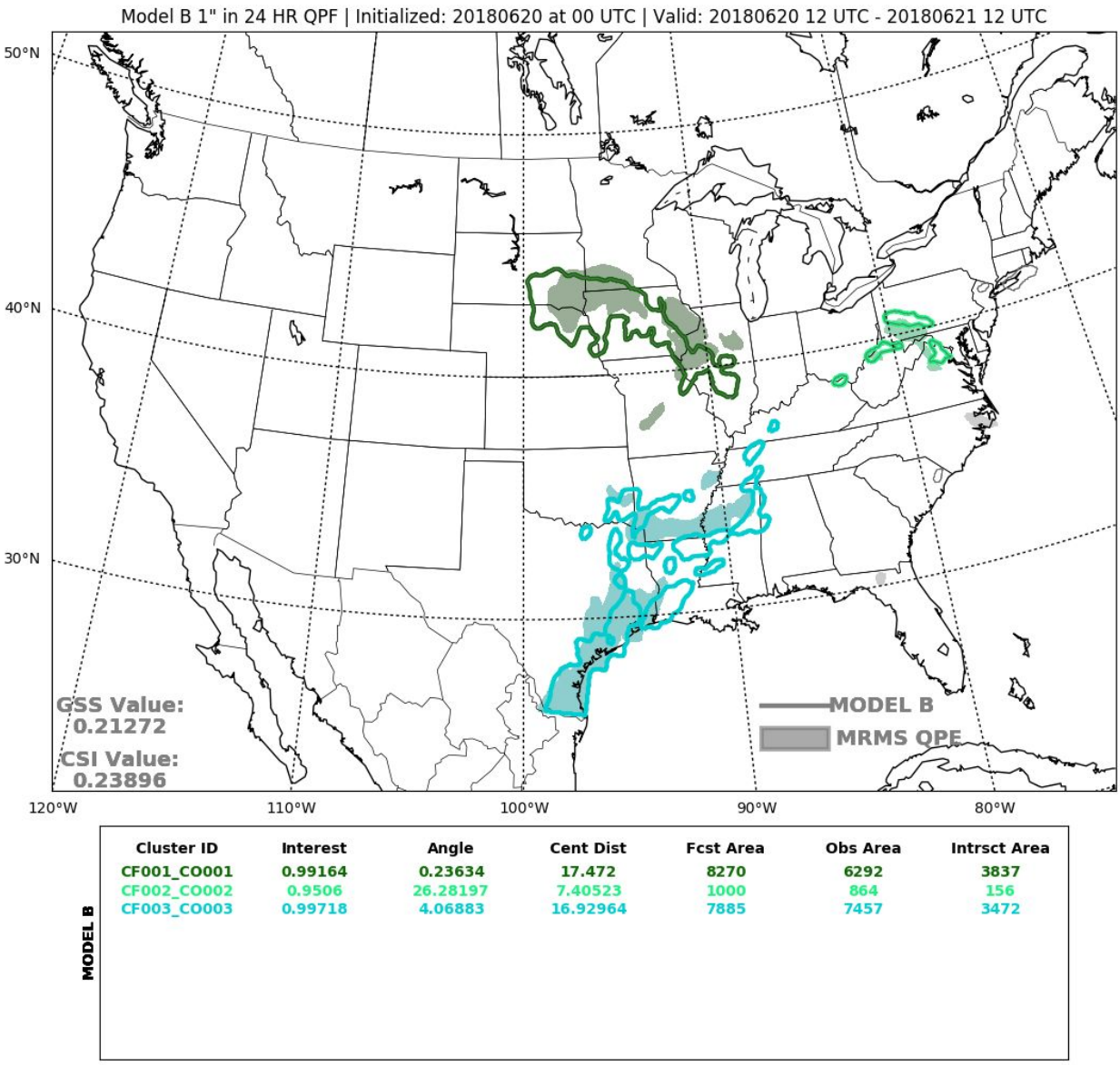


Figure 4. *MODE analysis for the 36 hour "Model B" (HRRRv3) forecast for 24 hour QPF at the 1 inch threshold valid from 12 UTC June 20 to 12 UTC June 21, 2018.*

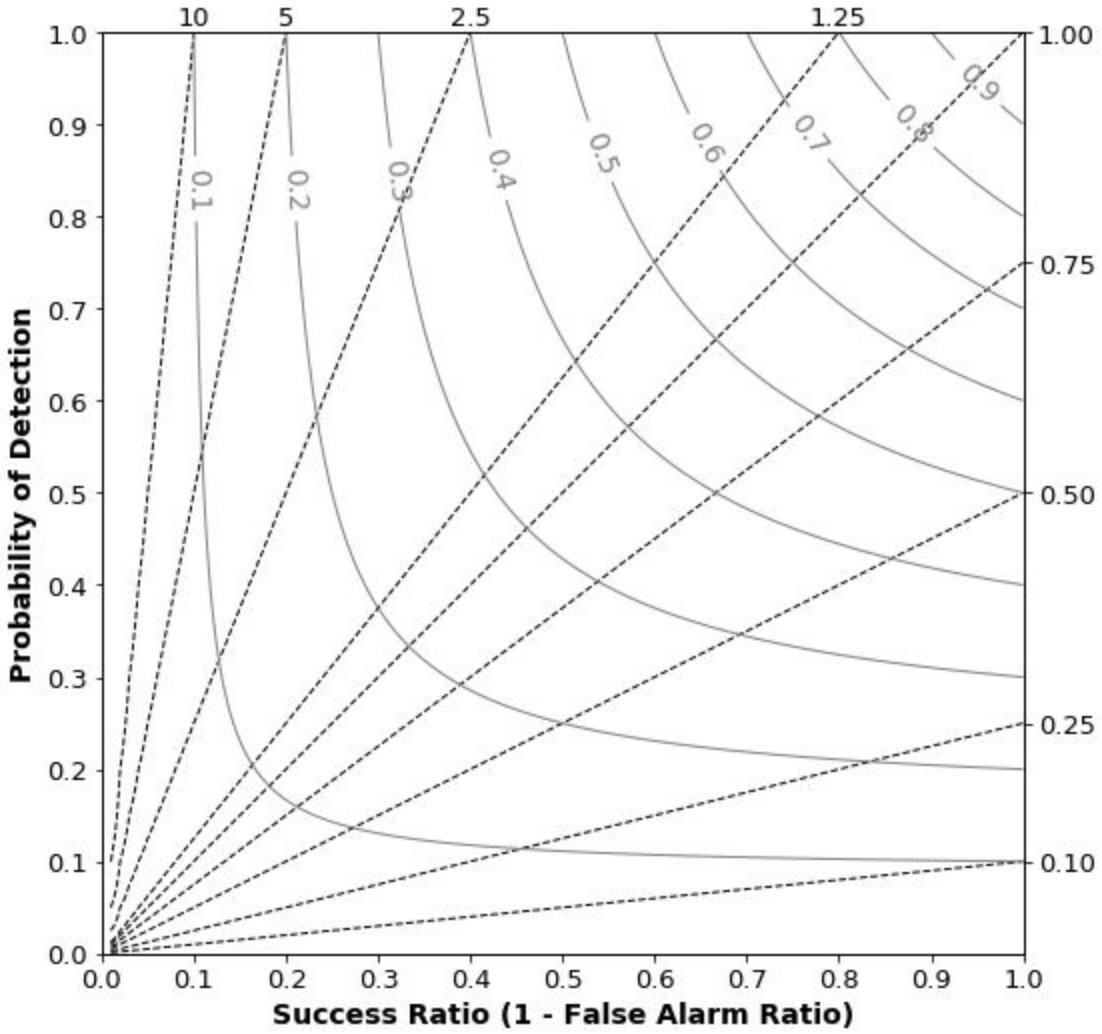


Figure 5. *Example of a Roebber Performance Diagram. Y-axis is the probability of detection, x-axis shows the success ratio (1 - false alarm ratio), dashed diagonal lines represent the bias, and curved solid lines represent critical success index.*

Featured Guidance and Tools for Experimental Forecasts

In addition to the full multi-center suite of operational deterministic and ensemble guidance, the 2018 FFaIR Experiment featured several experimental ensemble systems including the experimental Storm-Scale Ensemble Forecast (SSEFX) from the University of Oklahoma (OU) and Center for Analysis and Prediction of Storms (CAPS), the High-Resolution Rapid Refresh Ensemble (HRRRE) from Earth Systems Research Laboratory (ESRL), and the Experimental High Resolution Ensemble Forecast (HREFv2.1) provided by the Environmental Modeling Center (EMC). The experiment also featured several deterministic high-resolution models, including the High Resolution Rapid Refresh (HRRRv3) provided by ESRL and two 3-km FV3 CAM variants provided by OU/CAPS. Additionally, the experiment featured forecasts and probabilities derived from the National Water Model provided by the The Office of Water Prediction (OWP), the GFS run of the FV3 dynamical core provided by EMC, the National Blend of Models version

3.1, and the NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e). Table 3 summarizes the model data that was the focus of the experiment. More detailed information about each model is provided in Appendix C.

Table 3. *Featured 2018 FFaIR deterministic and ensemble model guidance (Experimental guidance is in the darker shade)*

Provider	Model	Resolution	Forecast Length	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
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 1 km and 250 m grids.
NSSL/HDSC/NERFC /CSU	Precipitation Recurrence Data (Atlas 14)	5 km	6 and 24 hour (2, 5, 10, 25 and 100 year intervals)	Precipitation frequency estimates based on historical observations.
OWP	National Water Model (NWM) Post-Processed Visualizations	250 m 1 km	From 1 hour to 10 days (product dependant)	Utilize a dataset of recurrence flows for each stream reach derived from a 23-year retrospective analysis of the NWM (v1.0)
ESRL/GSD	HRRRv3	3 km	Hourly out to 36 hours every 3 hours Hourly out to 18 hours every hour	Experimental version of the HRRR, hourly updating, convection allowing
ESRL/GSD	HRRR Ensemble (HRRRE)	3 km	36 hours at 00 UTC 18 hours at 12 UTC	9 forecast members, 36 DA members , Full-CONUS domain at 00 UTC, Sub-CONUS domain at 12 UTC, stochastic. 00/12 UTC cycles
NCAR	NCAR Ensemble	3 km	36 hours at 00 UTC 24 hours at 12 UTC	9 forecast members, 80 DA members, Full-CONUS domain at 00 UTC, Sub-CONUS domain at 12 UTC, continuously cycling initial conditions. 00/12 UTC cycles
EMC	HREFv2.1	3 km	36 hours	Experimental version of HREF with 10 members which produces ensemble mean precip in three different forms, and precipitation probability of exceedance of QPF, FFG, and RIs. 00/12 UTC cycles

OU/CAPS	SSEFX	3 km	60 hours	15-member (13 ARW+2 FV3) ensemble forecast). 00 UTC cycle
EMC	FV3-GFS	13 km	10 days	3D hydrostatic dynamical core; vertically Lagrangian; GFS analyses initialization/physics. 00 UTC cycle.
MDL	NBMv3.1	2.5 km	Hourly out 36 hrs 3-hrly to Day 8 6-hrly Days 8-10	Runs every hour with 15 different deterministic and ensemble systems
NSSL	NEWS-e	3 km	6 hour forecasts initialized every hour starting from 1800 UTC out o 0400 UTC	HRRRE analysis/boundary conditions; regional domain, every 15 min, EnKF assimilation; forecast files every 5 minutes

Synoptic Overview and Highlights of Daily Impacts Throughout the Experiment

Throughout the four weeks of the 2018 FFaIR Experiment, the participants dealt with a wide variety of forecasting challenges including mesoscale convective systems (MCS), deep tropical moisture, and monsoon activity in the Southwest United States. Figure 6A shows the 500 hPa heights over the United States during the first half (June 18-29, 2018) of the experiment and ()B shows the 500 hPa height anomalies. The highest geopotential heights at 500 hPa were suppressed to the southern tier of the United States. A slight trough is evident in the Northwest as well as off the coast of New England. The first two weeks were dominated by unusual summertime synoptic scale low pressure systems that moved through the Central Plains and anomalous tropical moisture that brought extremely heavy rainfall to the Texas Gulf Coast during week 1. The 500 hPa geopotential height anomalies in Figure 6B highlight the slightly lower than normal heights associated with the low pressure systems in the central U.S., particularly when considering that the majority of CONUS experienced positive height anomalies. A WPC surface analysis in Figure 7 shows an occluded 1000 hPa low in the central U.S. analyzed at 12Z on June 21, 2018. The low associated with the tropical moisture in Texas is also evident; this system meandered across the region for several days. During the second week, another large scale occluded low pressure system tracked across the central U.S., although it was more progressive bringing heavy rainfall both out ahead of the associated fronts and around the low center. Figure 8A shows 1000-500 hPa total column precipitable water anomalies over the first half of FFaIR. Most of the eastern half of the United States had anomalously high precipitable water values, particularly the Texas Gulf Coast region, the upper Mississippi Valley, and the Midwest.

Figure 6C shows the 500 hPa heights over the United States during the second half (July 9-20, 2018) of the experiment and 6D shows the 500 hPa height anomalies. Compared to the first half (Figure 6A), much higher 500 hPa heights extended north over most of the country with a large ridge in the central U.S. and minor troughs off both the Northeast and Northwest coasts. Height anomalies were above normal over the entire U.S. with the highest positive anomalies in the Northwest. During the second half of the experiment, the Southwest monsoon was extremely active with heavy rainfall and flash flooding reports occurring each day. The

precipitable water anomalies in Figure 8B for weeks three and four of FFaIR illustrate the extremely moist monsoonal environment over the Southwest. Other areas that had anomalously high precipitable water values included the central Gulf Coast north into the Central Plains.

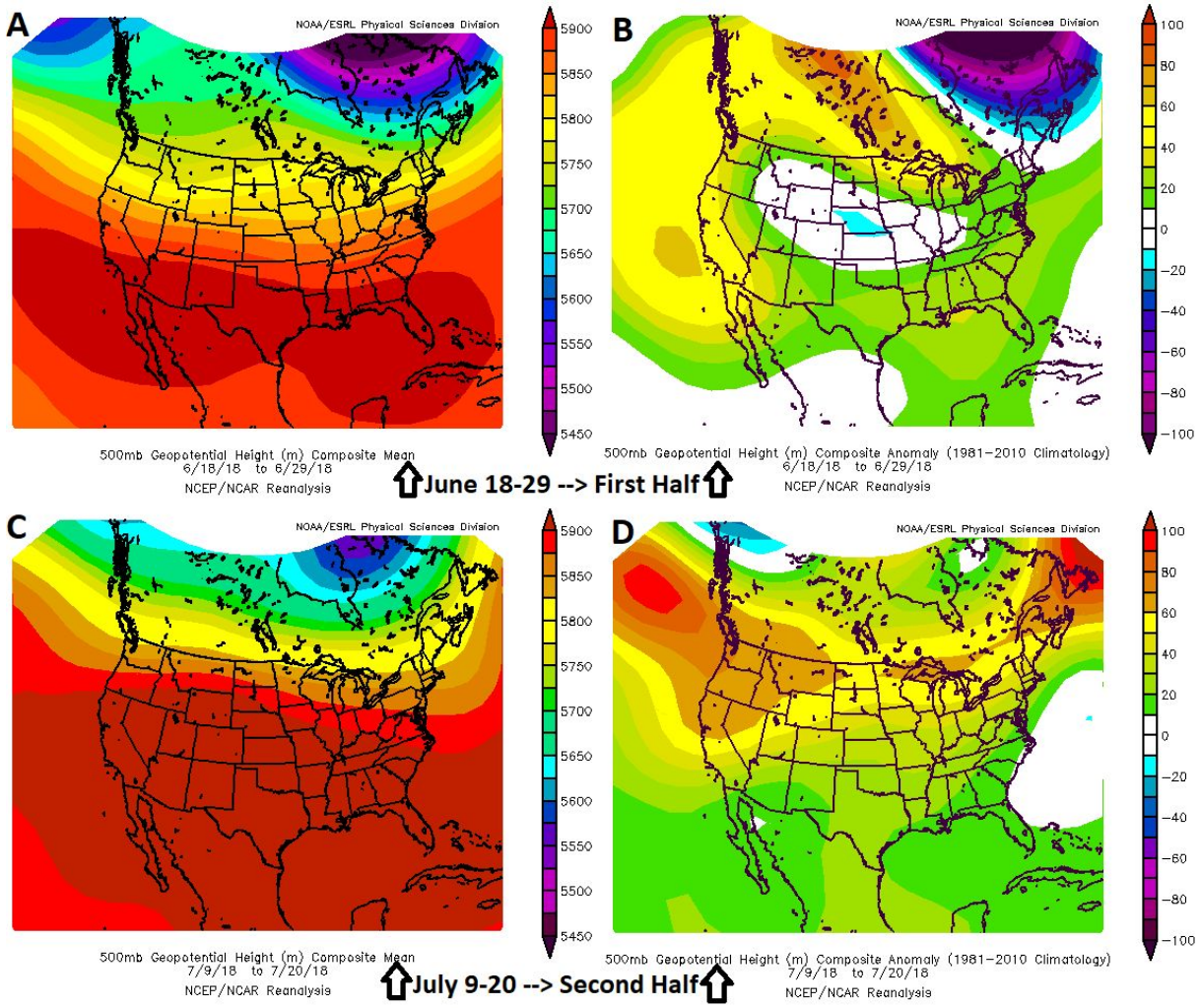


Figure 6. *(A) 500 hPa mean geopotential height and (B) 500 hPa geopotential height composite anomalies for the first half of FFaIR covering June 18 - June 29, 2018. (C) 500 hPa mean geopotential height and (D) 500 hPa geopotential height composite anomalies for the second half of FFaIR covering July 9 - 20, 2018. Images generated from the NCEP/NCAR Reanalysis provided by NOAA/ESRL/Physical Sciences Division (http://www.esrl.noaa.gov/psd/data/composites/day/).*

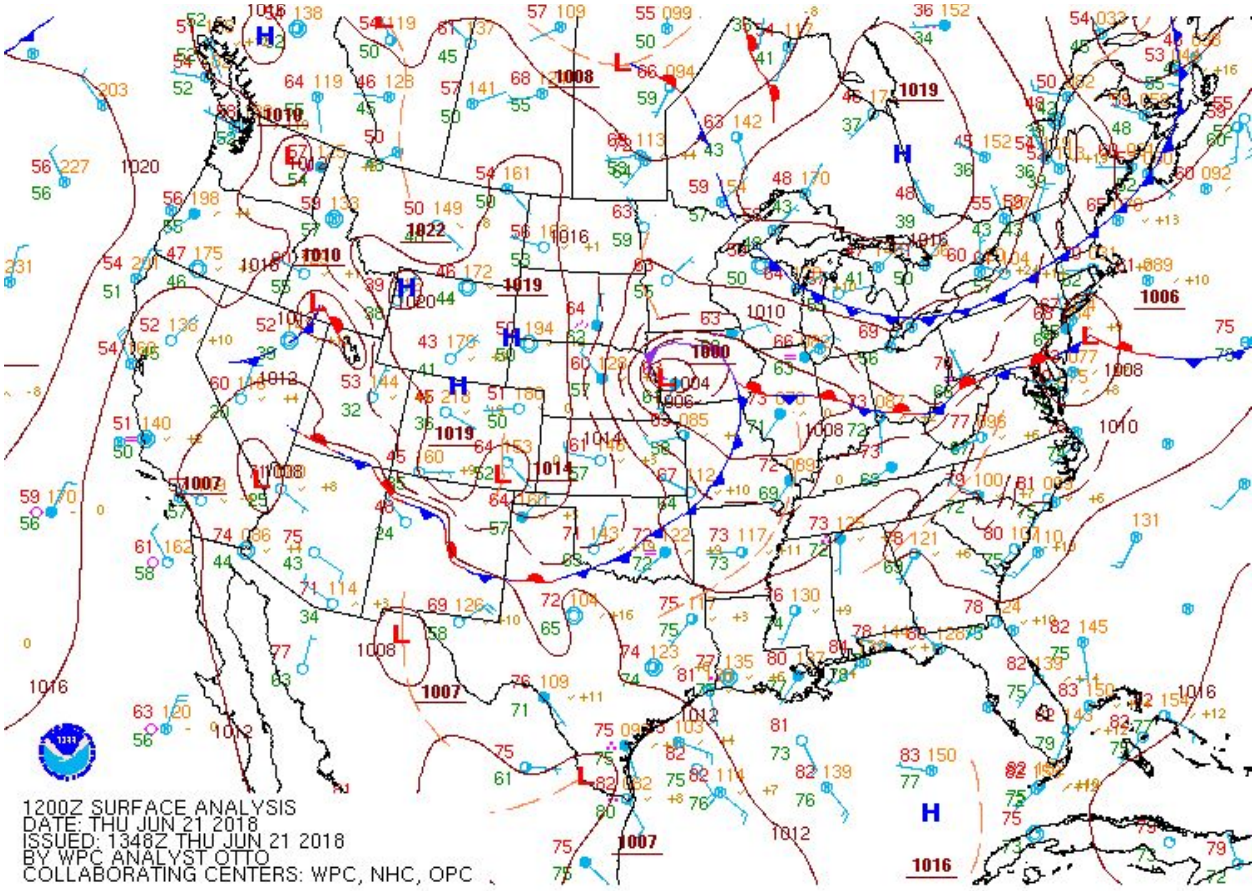


Figure 7. *WPC surface analysis valid at 12Z June 21, 2018.*

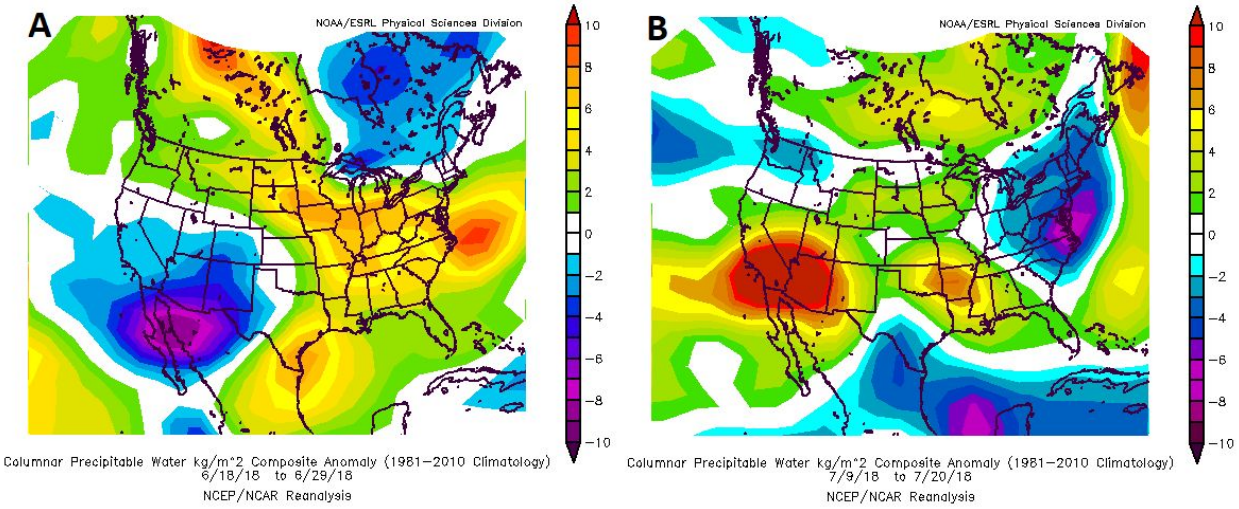


Figure 8. *1000 - 500 hPa precipitable water composite anomalies for (A) the first half of FFaIR (June 18 - June 29, 2018) and (B) the second half of FFaIR (July 9 - 20, 2018). Images generated from the NCEP/NCAR Reanalysis provided by NOAA/ESRL/Physical Sciences Division (http://www.esrl.noaa.gov/psd/data/composites/day/).*

There were a number of high impact flash flooding events that occurred during the four weeks of the 2018 FFaIR Experiment. Tables 4-7 in Appendix A give a weekly overview of the forecasts issued each day, the geographic areas highlighted, and any noteworthy impacts. During week one, multiple days of heavy rain from persistent tropical moisture led to over a foot of rain in many Texas Gulf Coast locations. Figure 9 shows the accumulated rainfall in that region from 12Z June 18 - 12Z June 22, 2018. Flash flood emergencies were issued in the far southern part of the state as well as to the north in Rockport and Port Aransas, TX.

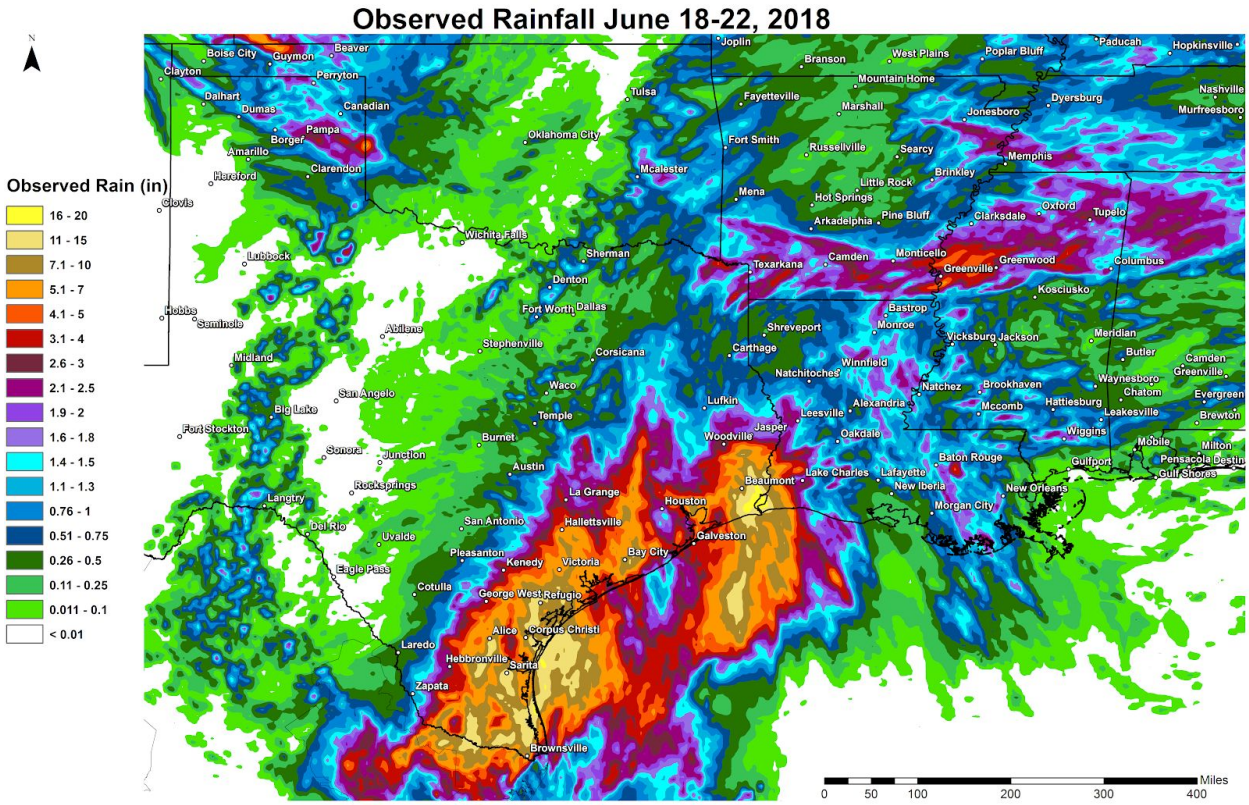


Figure 9. *MRMS-GC QPE valid 12Z June 18 - 12Z June 22, 2018.*

Also during week one, high impact events occurred in both Rockford, IL and Pittsburgh, PA where heavy rainfall in a short amount of time led to multiple water rescues and road closures in both cities. Week three saw multiple flash flood reports throughout the Southwest United States associated with monsoonal moisture. On July 12, multiple hikers had to evacuate the Havasu Falls portion of the Grand Canyon due to flash flooding. Finally during the fourth week on July 17, a cold frontal passage produced storms with very heavy rainfall out ahead and along the front in New England and the Mid-Atlantic. Particularly hard hit were areas in Massachusetts and also in Washington D.C. area where 25 vehicles were stranded due to flood waters on a highway in northern Virginia.

4. Atmospheric Guidance Results

Deterministic Models and National Blend of Models v3.1 Performance

Several deterministic models were featured during the 2018 FFaIR experiment both for forecast guidance and evaluation. A primary science goal was to determine the utility of Day 1 high-resolution deterministic QPF from the 13-km FV3-GFS, 3-km HRRRv3, and two 3-km FV3s from OU/CAPS: One utilizing Thompson microphysics and the other with NSSL microphysics.

The participants were presented with a display of Day 1 24-hour QPFs (Figure 10B) alongside the MRMS-GC QPE (Figure 10A) for that same time period and asked to evaluate the QPF on a scale from 1 (very poor) to 10 (very good) based on factors such as areal extent, magnitude, placement, and timing. Subjective scores and comments were collected for available guidance each day during the experiment.

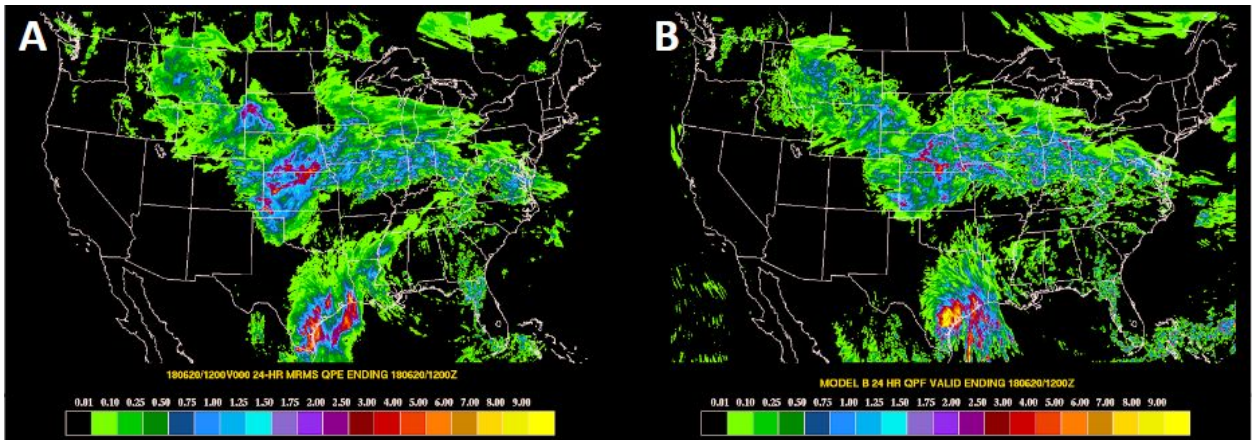


Figure 10. *An example of a how the experimental model QPF was evaluated with (A) 24-hr MRMS-GC QPE and (B) 24-hr deterministic model QPF from HRRRv3 ("Model B") in this comparison.*

FV3-GFS Findings

The participants scored the FV3-GFS an average of 6.28, with the lowest score being a 2 and the highest a 10 (the only deterministic model to receive the highest possible score of 10). A box plot of all subjective scores for all the models tested is shown in Figure 11 with the FV3-GFS on the far left.

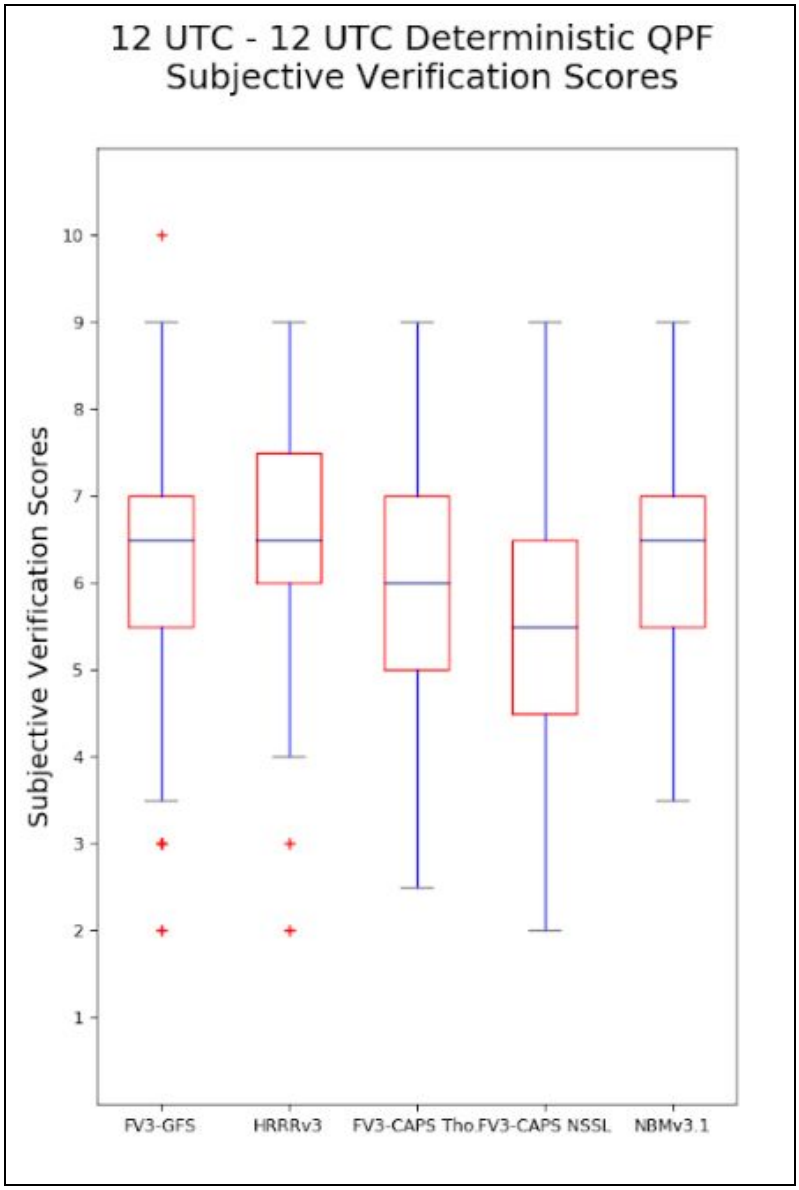


Figure 11. *Box plot of the subjective scores for the FV3-GFS, HRRRv3, FV3-Thompson, FV3-NSSL, and NBMv3.1 24-hour QPF over the course of the experiment. Red plus symbols denote outliers.*

The model did exceptionally well capturing the areal extent and location of precipitation events. The FV3-GFS had the coarsest resolution among the experimental models and is not convection-allowing, therefore, as expected, there were frequent comments regarding magnitudes of precipitation that were too low, or underdone, especially in regions of convection. These characteristics and a dry bias are shown in the performance diagram in Figure 12A for a 0.5 in. and more noticeably in (B) at the 1.0 in. threshold, where the FV3-GFS is the dark green circle in the diagrams.

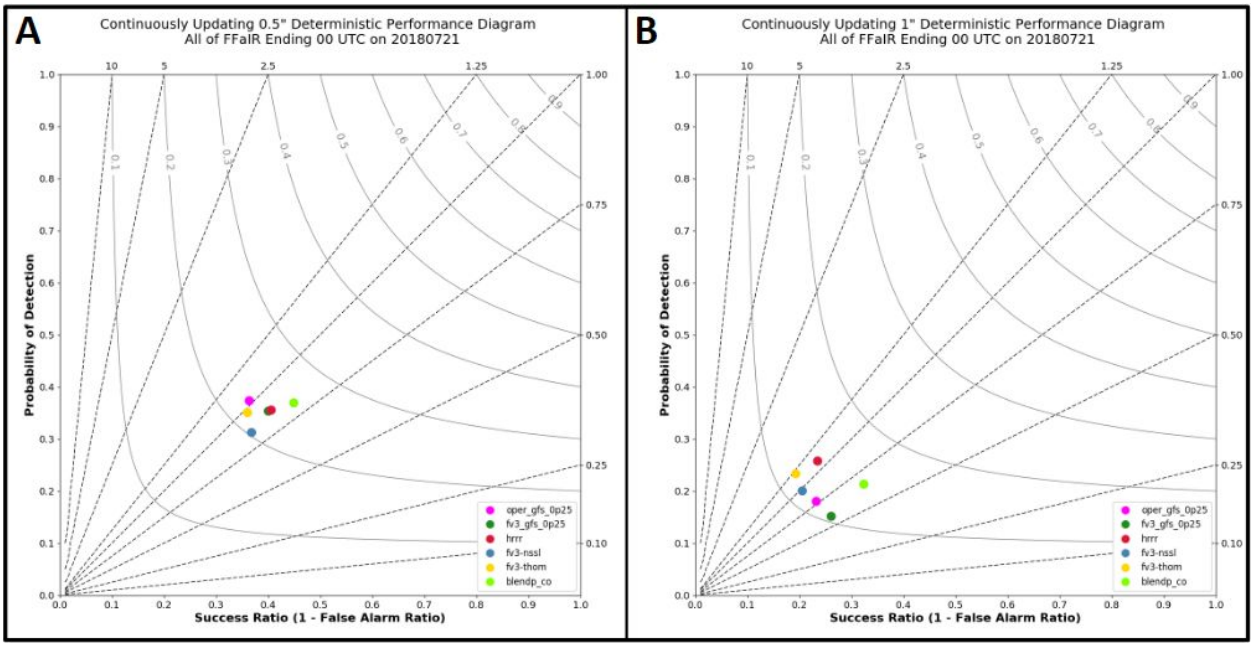


Figure 12. *Performance diagram displaying the FV3-GFS (dark green), HRRRv3 (red), FV3-NSSL (blue), FV3-Thompson (yellow), NBMv3.1 (light green), and the operational GFS (magenta) over the four weeks of the 2018 FFaIR Experiment.*

HRRRv3 Findings

The experimental HRRRv3 became operational midway through the FFaIR Experiment. It received the highest subjective average score, 6.54 out of 10, among the deterministic high resolution guidance, but had the fewest scoring opportunities due to sporadic outages. The highest score it received was 9 and the lowest a 2.

The HRRRv3 was considered favorably the majority of test days, however, participants often noted that the QPF was underdone, especially for more marginal or weakly-forced events. Conversely, in areas of strong synoptic forcing or enhanced convection, the magnitude of the precipitation would often be too high. Figure 13 is an example from June 28-29 where both of these issues were prevalent. From the performance diagrams in Figure 12, the HRRRv3 (red) had a slight dry bias at the 0.5 in. threshold and a slight wet bias at the 1.0 in. threshold. The model had the second highest CSI value among the experimental models tested at both thresholds.

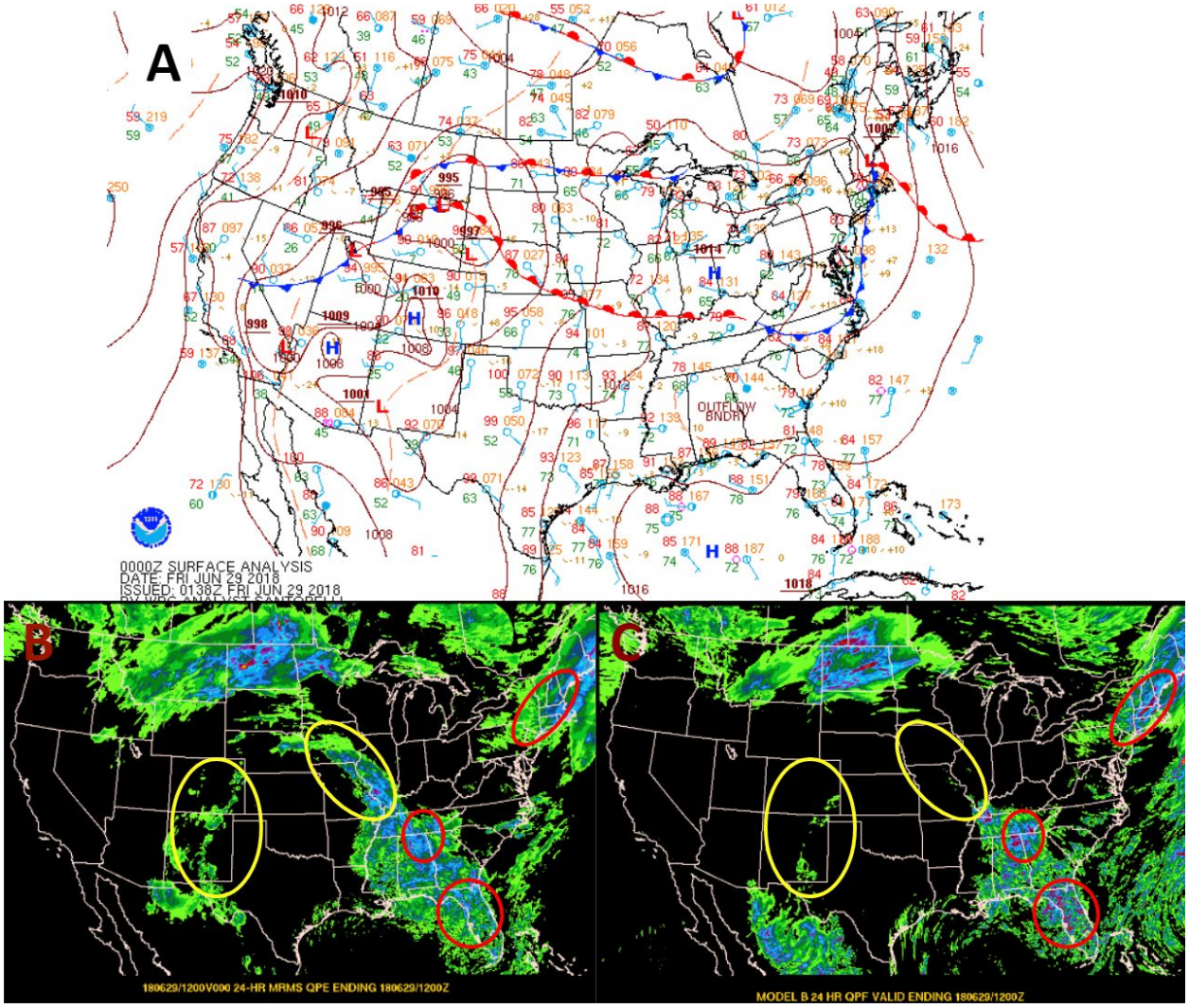


Figure 13. *(A) The WPC surface analysis valid 00z June 29, 2018 (mid-way through the 24 hour valid period), (B) 24 hour MRMS-GC QPE, and (C) 24 hour QPF from the HRRRv3, both valid from 12 UTC June 28 - 12 UTC June 29, 2018. Yellow circles highlight areas where the HRRRv3 QPF (B) was lighter than the QPE (A) and red circles indicate areas where the QPF was heavier than the QPE in this example.*

FV3-NSSL and FV3-Thompson Findings

Two 3-km deterministic FV3 models were provided by the OU CAPS team, one with NSSL microphysics and the other with Thompson microphysics. The FV3-NSSL achieved an average score of 5.89 out of 10, and the FV3-Thompson slightly lower at 5.57 out of 10.

The FV3-NSSL regularly struggled to produce organized precipitation over the CONUS. Participants frequently commented on the low magnitude and scattered nature of the precipitation. Figure 14 shows a MODE analysis of 24 hour QPF from both the FV3-Thompson and FV3-NSSL at the 0.5 in. threshold with three separate areas highlighted in the red circles where MRMS-GC QPE was not matched by MODE to QPF from either model. At times, erroneously high QPF maxima would occur related to certain features which the participants

found distracting and could be misleading in the forecast process. The FV3-Thompson, likewise, often struggled to produce precipitation leading to low magnitudes and scattered representations but did slightly better than the FV3-NSSL in organization, timing and location accuracy. Figure 15 shows that the the FV3-Thompson had slightly higher CSI values at the 0.5/1.0/2.0 in. thresholds when compared to the FV3-NSSL. There were days when the FV3 models performed well, as both models were given scores as high as a 9 out of 10 for some events, but the erratic errors in the rainfall prediction tended to create forecaster distrust in the both FV3 models over the experiment.

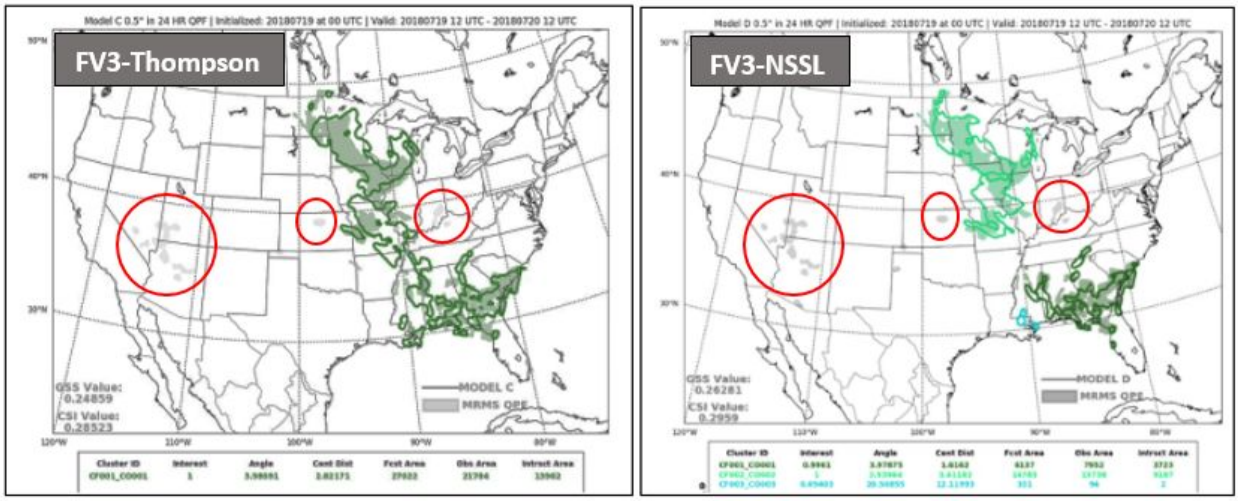


Figure 14*. Objective MODE performance of the 24-hour 0.5" QPF from the FV3-Thompson (left) and FV3-NSSL (right) both valid at 12 UTC July 20, 2018. Noted in the red circles are areas of MRMS QPE that were not matched by MODE to the model QPF.*

National Blend of Models v3.1 Findings

The National Blend of Models version 3.1 (NBMv3.1) was examined with the deterministic guidance in order to evaluate its 24 hour mean QPF over the entire CONUS. The NBMv3.1 had an average subjective score of 6.47 out of 10, with a highest score of 9 and lowest of 3.5. The NBMv3.1 QPF achieved the best CSI score among the deterministic models at both the 0.5 in. (0.263) and 1.0 in. (0.156) thresholds as shown in Figure 15.

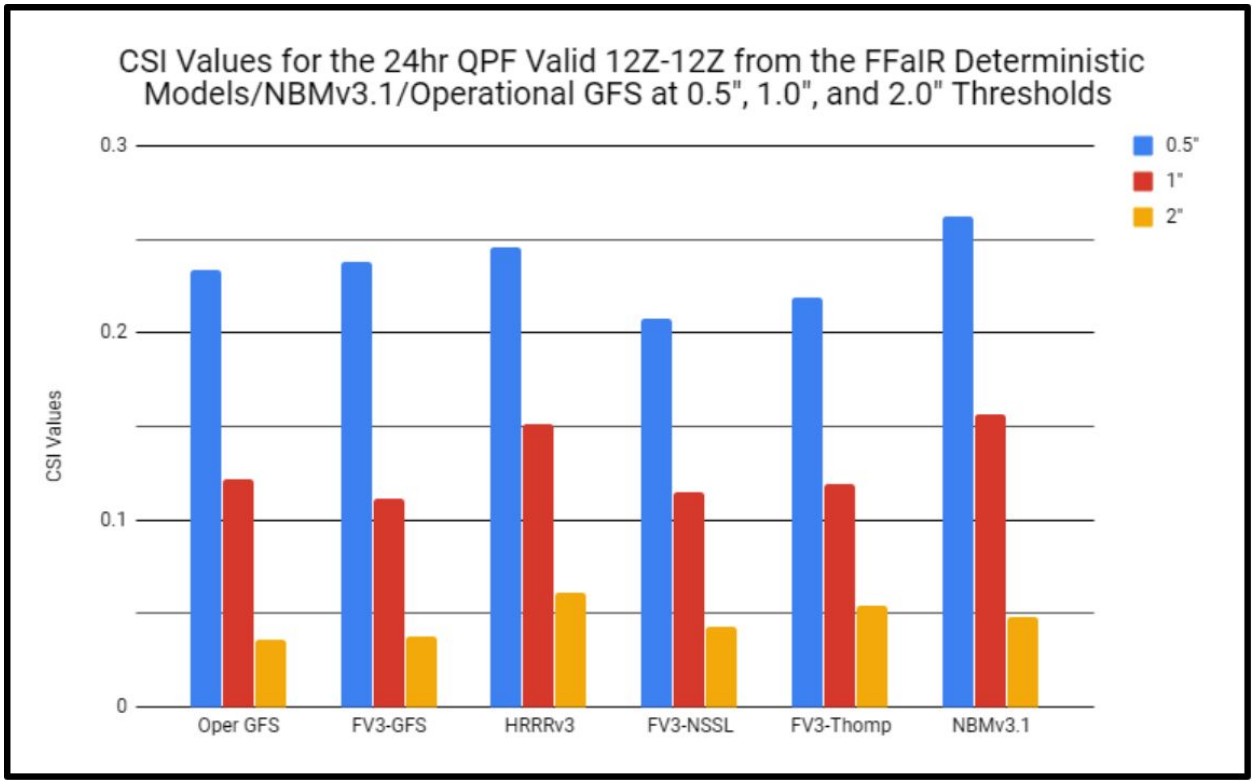
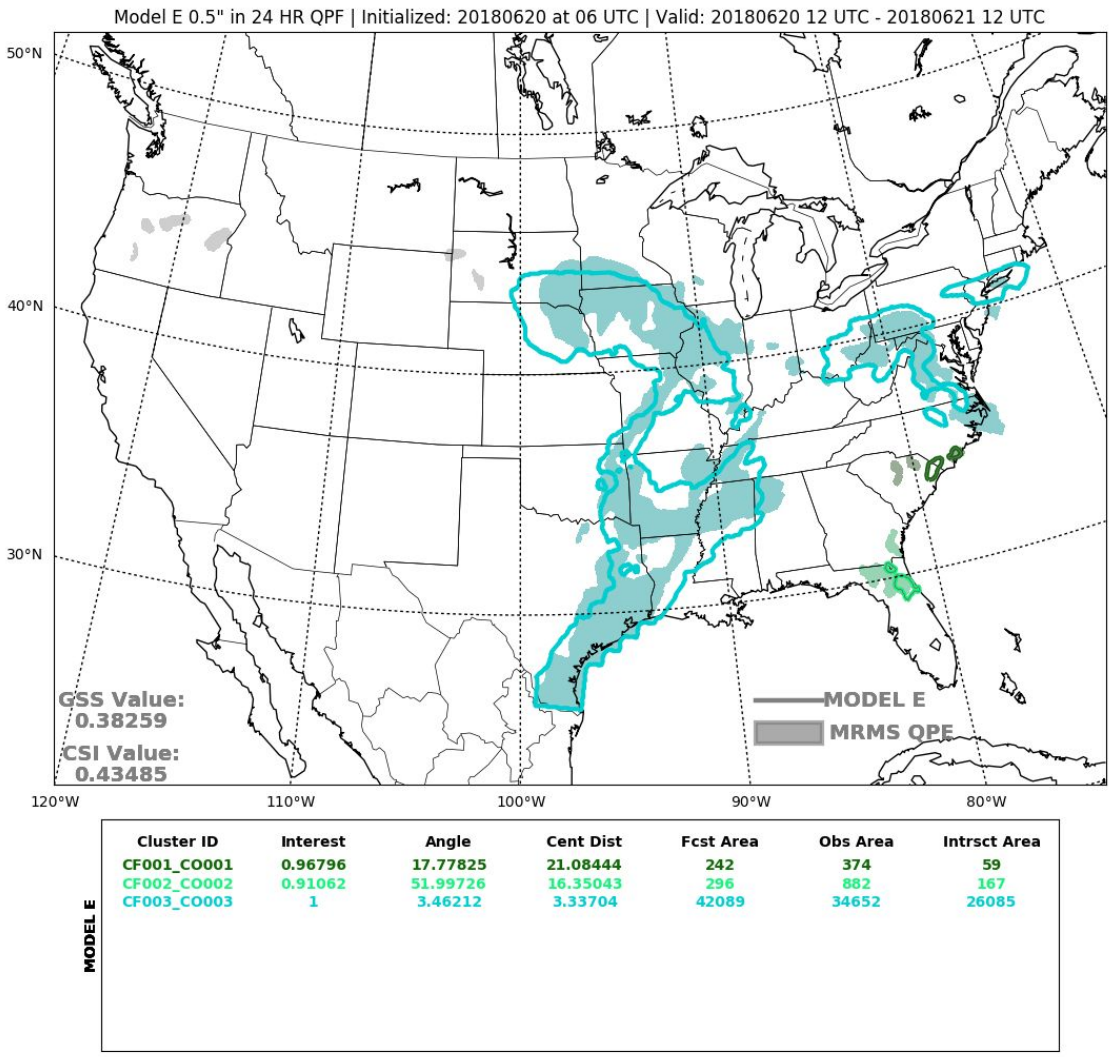


Figure 15. *CSI values of the 0.5", 1" and 2" QPF thresholds from the experimental deterministic models and NBMv3.1 evaluated during the 2018 FFaIR Experiment. The operational GFS model is included as well for comparison.*

The NBMv3.1 received frequent praise throughout the experiment for its accurate prediction of areal extent, location and overall representation of the main areas of precipitation over the CONUS. Figure 16 shows a MODE analysis of the 24 hour QPF at the 0.5 in. threshold for the NBMv3.1 valid at 12 UTC June 21, 2018. Object 3 (light blue) in particular shows how well the NBMv3.1 (outline) matched with the MRMS-GC QPE (shaded) in terms of areal coverage. The forecast object's centroid distance was offset by just 3.33 grid squares, the major axis angle difference was only 3.46 degrees, and 62% of the forecast area intersected the observation area. The NBMv3.1's CSI value for this example was 0.435. Participants noted that for several events the extent of the light precipitation was too large and diffuse. Despite the inclusion of the high-resolution convective-allowing models through the first 36 forecast hours, the NBMv3.1 QPF was often underdone with the highest areas of precipitation, often in convective environments. The subjective results aligned with the objective results as the NBMv3.1 had the highest CSI values at both 0.5 in. and 1.0 in. but the most highest dry bias at 0.5 in. and second highest at 1.0 in.



MODEL E

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.96796	17.77825	21.08444	242	374	59
CF002_CO002	0.91062	51.99726	16.35043	296	882	167
CF003_CO003	1	3.46212	3.33704	42089	34652	26085

Figure 16. *MODE results of 24-hr NBMv3.1 0.5" QPF (contoured) compared to the MRMS-GC QPE (shaded) valid 12 UTC June 20 - 12 UTC June 21, 2018.*

Recommendations

The HRRRv3 went operational during the experiment in July, however it is recommended that future versions examine the low bias cases in the Southwest monsoon environment and several of the cases where the HRRRv3 had a high bias in convective or strongly forced regions. The NBMv3.1 is scheduled to be operational in early fall 2018. WPC-HMT recommends improving the low bias of the higher QPF thresholds (participants noted this factor most often) for future versions of the NBM. The FV3-GFS is also scheduled to be operational in the winter of 2018/19. Despite the small sample size of 26 cases, WPC-HMT recommends monitoring the dry bias noted in the FV3-GFS compared to the operational GFS at the 0.5/1.0/2.0 in. thresholds moving forward. Continued development work is needed for high resolution FV3 members from CAPS using the NSSL and Thompson microphysics. A much longer testing period is recommended to thoroughly investigate the issues highlighted in the limited (20) cases evaluated during FFaIR.

Ensemble Guidance Performance

Ensemble Local Probability Matched Mean and Probability Matched Mean Performance
The probability matched mean (PMM) QPF was evaluated from four different ensemble systems: the SSEFX, the HRRRE, the NCAR ensemble, and the HREFv2.1. In addition to the four PMMs, a local probability matched (LPM) mean method from the SSEFX was also assessed. The LPM mean method calculates the PMM over smaller sub-domains rather than in the entire model domain. More information on both the LPM method from the SSEFX and the other ensemble systems can be found in Appendix C. For each ensemble, the 18-24 hour forecast was subjectively evaluated with a 1 (very poor) to 10 (very good) score and the forecasts were valid from 18-00 UTC over the same region the participants made their PFF1. Each ensemble QPF was shown separately alongside the MRMS-GC QPE, an example of which is in Figure 17. The names of each ensemble system were replaced by generic names on the images. Due to data availability issues at times throughout the experiment, Figure 18 shows the number of scores collected for each ensemble system throughout the four weeks of the experiment.

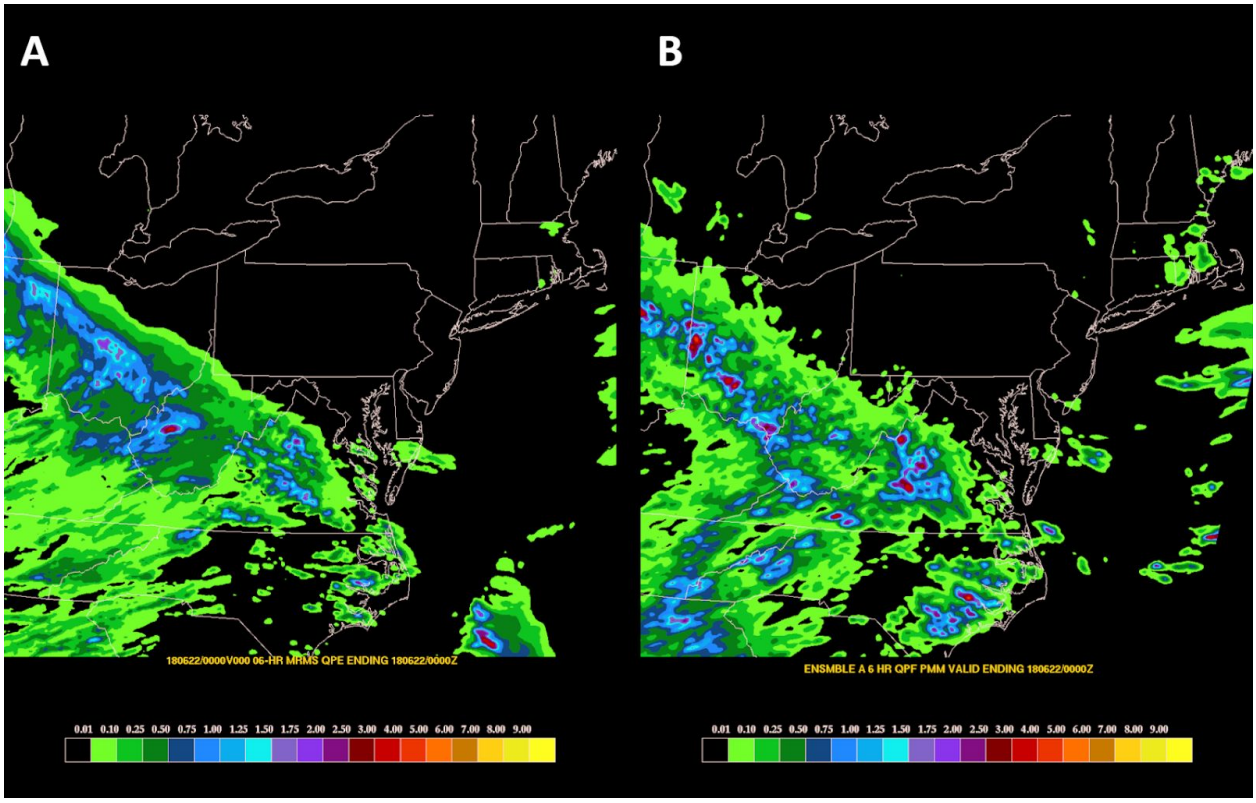


Figure 17. *(A) Six hour MRMS-GC QPE valid 18 UTC June 21 - 00 UTC June 22, 2018 and (B) Six hour LPM QPF from the SSEFX (Ensemble A) valid over the same time period.*

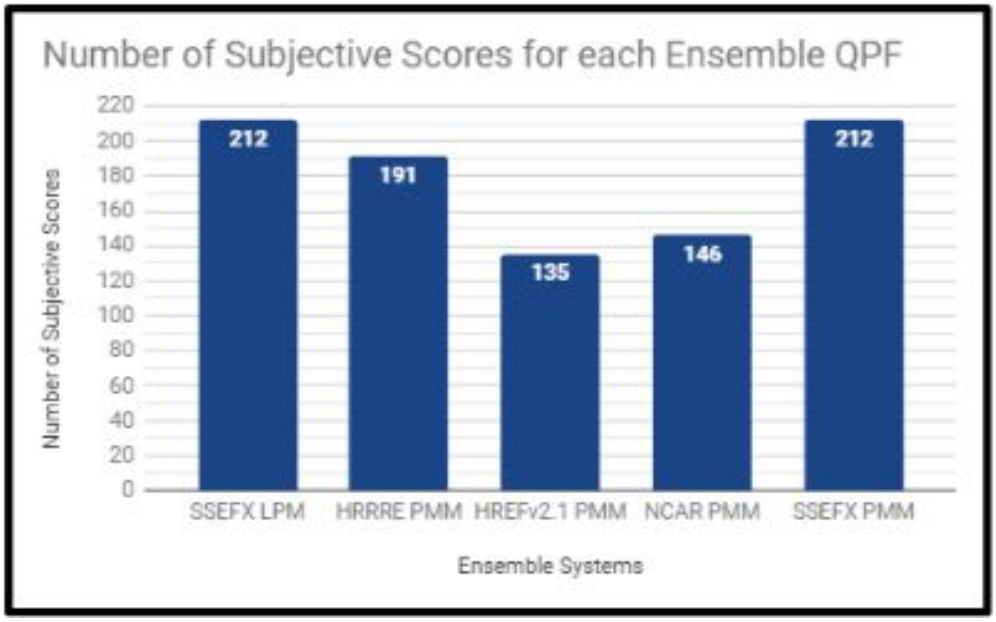


Figure 18. *Number of subjective scores recorded for each ensemble system throughout the four weeks of the experiment.*

Findings

Figure 19 is the box plot for all the subjective scores for the five PMM/LPM QPFs from the four ensembles. The SSEFX LPM method had the highest average subjective score at 6.03. The HREFv2.1 PMM had the next highest average score of 5.80, followed by 5.16 from the SSEFX PMM, 4.83 from the HRRRE PMM, and 4.50 from the NCAR PMM. Participants found that the SSEFX LPM method worked well in many cases in reducing the magnitude of the QPF when comparing it directly to the SSEFX PMM and the other ensemble PMMs. Figure 20 shows one example where the PMM from the SSEFX was significantly heavier than the LPM in the highlighted region of southern Wisconsin/northern Illinois. A common observation between the other three ensemble systems (HREFv2.1, HRRRE, NCAR) was that the PMM QPF often had high magnitudes when compared to the observations. Participants found the spatial extent of the overall areas of precipitation was handled well for most of the ensemble systems. The NCAR PMM struggled the most with structure and spatial coverage with the most common feedback from participants being the areal extent was too narrow. Figure 21 shows an example where the NCAR PMM did not have enough coverage, especially in southeastern Texas.

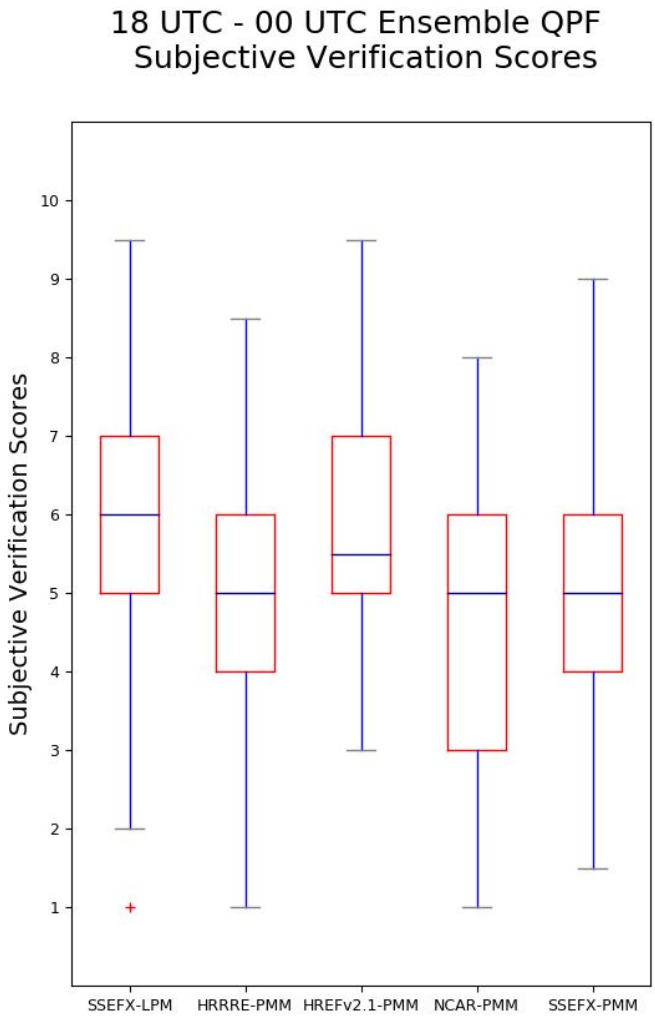


Figure 19. *Box plot for all the subjective scores over the four weeks of the experiment for the SSEFX LPM, HRRRE PMM, HREFv2.1 PMM, NCAR PMM, and SSEFX PMM 6 hr QPF valid 18-00 UTC. Red plus symbols denote outliers.*

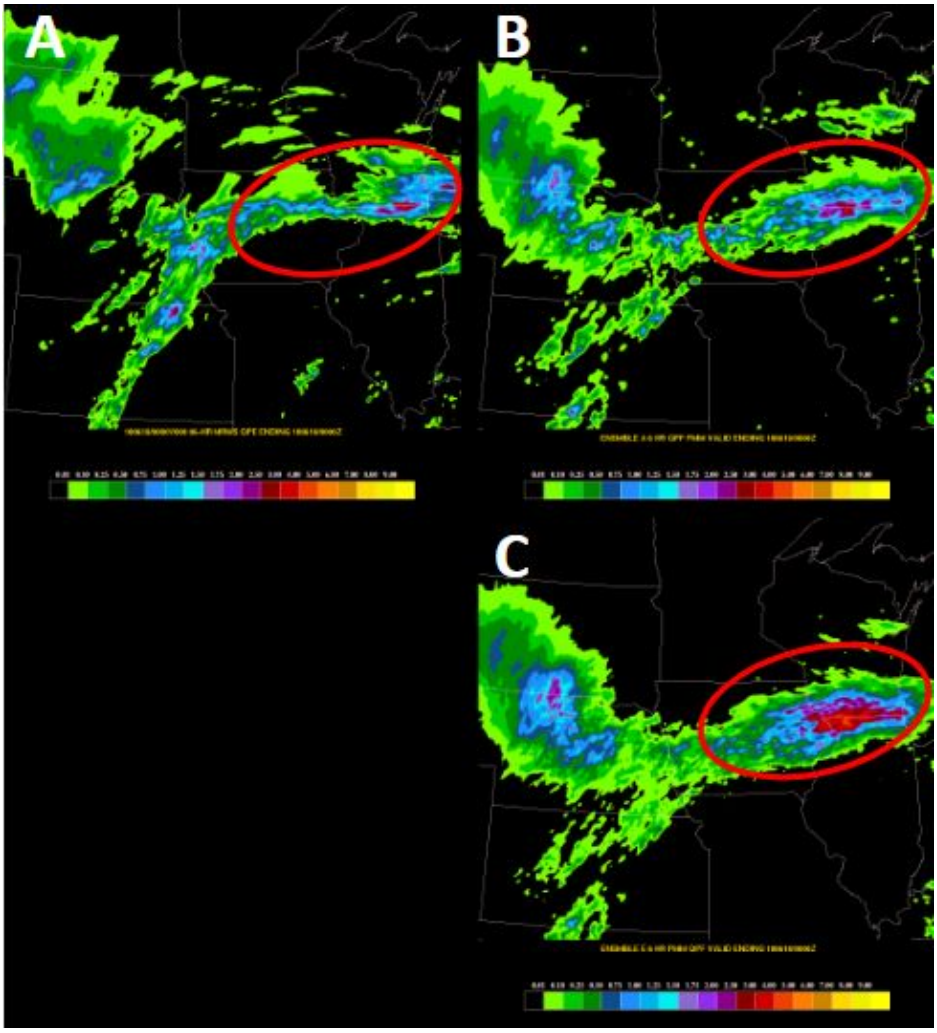


Figure 20. *(A) 6 hour MRMS-GC QPE valid from 18 UTC June 18 - 00 UTC June 19, 2018, (B) SSEFX LPM 6 hour QPF, and (C) SSEFX PMM 6 hour QPF valid over the same time period.*

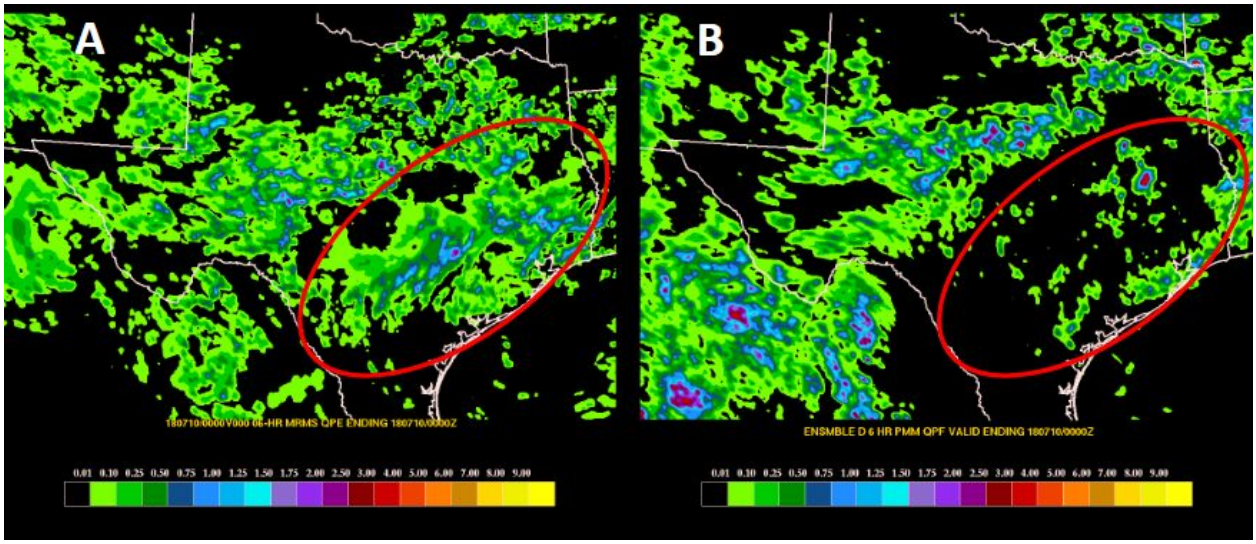


Figure 21. *(A) 6 hour MRMS-GC QPE valid from 18 UTC July 9 - 00 UTC July 10, 2018, (B) NCAR PMM 6 hour QPF valid over the same time period.*

The 6 hour PMM and LPM QPF from the four ensemble systems were objectively verified over the four weeks of the experiment. It is important to note that the objective verification was done over the entire CONUS domain and not just in the limited domains assessed during the subjective evaluation. Figure 22 shows the performance diagram for each ensemble at the 0.5 in. (A) and 1 in. (B) thresholds and Figure 23 highlights just the CSI values of each ensemble at the same thresholds. At the 0.5 in. threshold, all except for the NCAR PMM are above the 0.1 CSI threshold. The SSEFX PMM and HREFv2.1 PMM had the highest CSI but also a wet bias. The HRRRE PMM and SSEFX LPM both had lower CSI values but had very little bias. At the 1.0 in. threshold, all ensemble systems have a CSI below 0.1, but still have very similar results to those at the 0.5 inch threshold. The SSEFX PMM and HREFv2.1 PMM have the highest CSI but the bias is wetter than at the 0.5 in. threshold. Similarly, the SSEFX LPM and HRRRE PMM have lower CSI values and exhibit a slight wet bias. The NCAR PMM has a dry bias and lowest CSI value at each threshold.

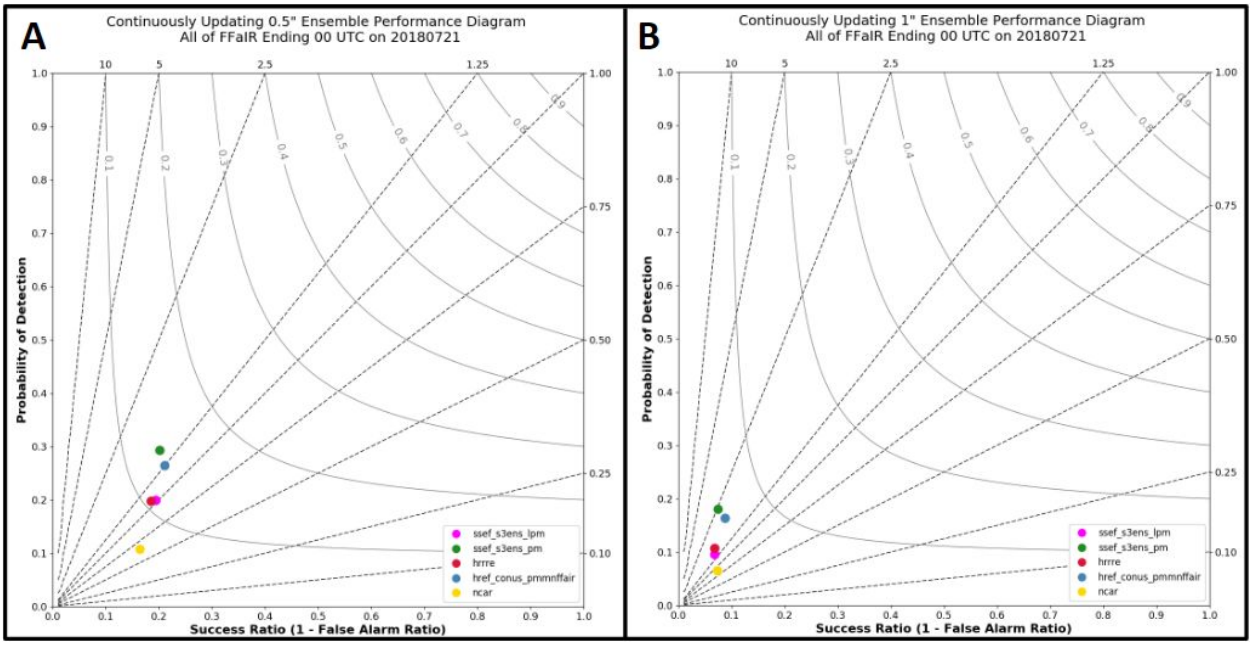


Figure 22. *Performance diagrams for the SSEFX LPM (magenta), SSEFX PMM (green), HRRRE PMM (red), HREFv2.1 PMM (blue), and NCAR PMM (yellow) 6 hour QPF over the four week experiment at (A) 0.50 inch and (B) 1.0 inch threshold.*

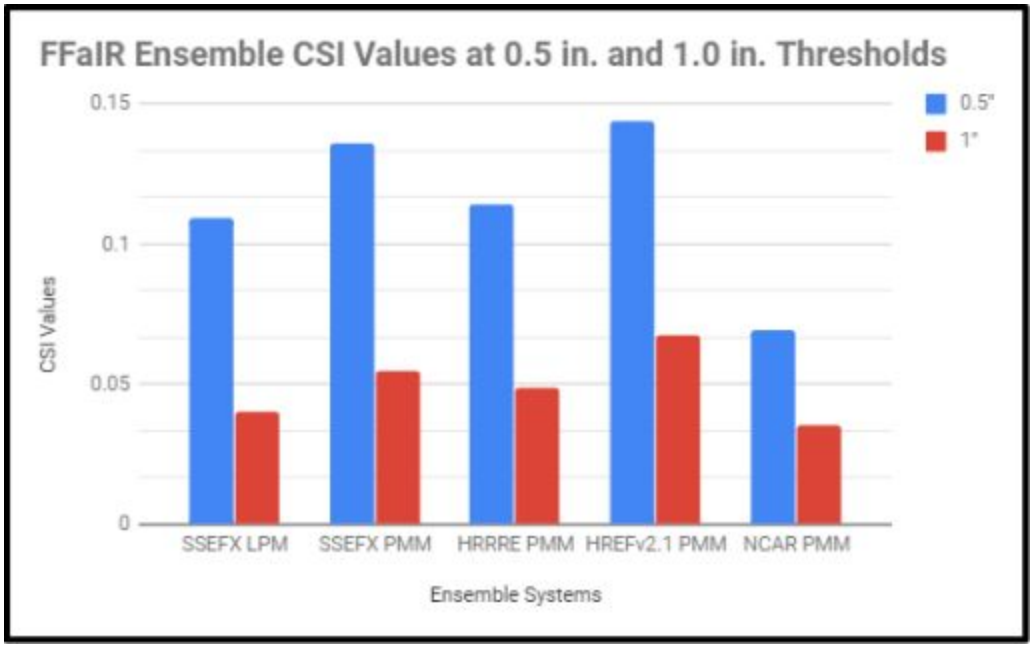


Figure 23. *CSI values for the SSEFX LPM, SSEFX PMM, HRRRE PMM, HREFv2.1 PMM, and NCAR PMM over the four weeks of FFaIR at 0.5 in. (blue) and 1.0 in. (red).*

Recommendations

The SSEFX LPM method had the highest subjective average score and best bias in the objective results. Its CSI value was similar to the HRRRE and slightly lower than the HREFv2.1 PMM and SSEFX PMM. WPC-HMT recommends that other ensemble systems begin testing and using the

LPM method in the calculation of the PMM as it has shown promise in bringing down the higher values seen at times in the traditional PMM method. WPC-HMT also recommends the HRRRE/NCAR to run on a full CONUS domain for all cycles. Newer 12 UTC guidance was often used for the PFF1 and PFF2 products in the afternoon, but for areas in the Southwest, these ensembles could not be used because a full CONUS domain was only available at 00 UTC. With a full CONUS domain for all cycles, future evaluations of the 12 UTC cycles in comparison with the 00 UTC cycles would be useful.

HREFv2.1 Neighborhood Probability and Ensemble Agreement Scale (EAS) Probability Evaluation

Two different methods for displaying probabilities of QPF from the HREFv2.1 were tested during the 2018 FFaIR Experiment. The first is the more traditional 40 km neighborhood probability method. A 40 km neighborhood probability represents the fraction of ensemble members that exceed a specified threshold anywhere within 40 km of a point. The second method tested was the Ensemble Agreement Scale (EAS) (Blake et al. 2018; Roberts and Lean 2008). EAS probabilities are effectively point probabilities that acknowledge the existence of spatial uncertainty in a forecast. They represent the fraction of points from all members within a radius of influence around each grid point that exceed a threshold. The size of the radius/filter is variable, and is determined by EAS similarity criteria - the smallest possible radius exists where the ensemble members have the highest agreement. More details on this and the HREFv2.1 can be found in Appendix C. Each probability method was evaluated by participants at two different thresholds: 0.5 in./6 hours and 1.0 in./6 hours. The probabilities were compared to MRMS-GC QPE and participants were asked to comment on each probabilistic method at each threshold and give opinions on how useful each might be in the forecast process. Figure 24 is an example of how the two probability methods from the HREFv2.1 were evaluated during the experiment.

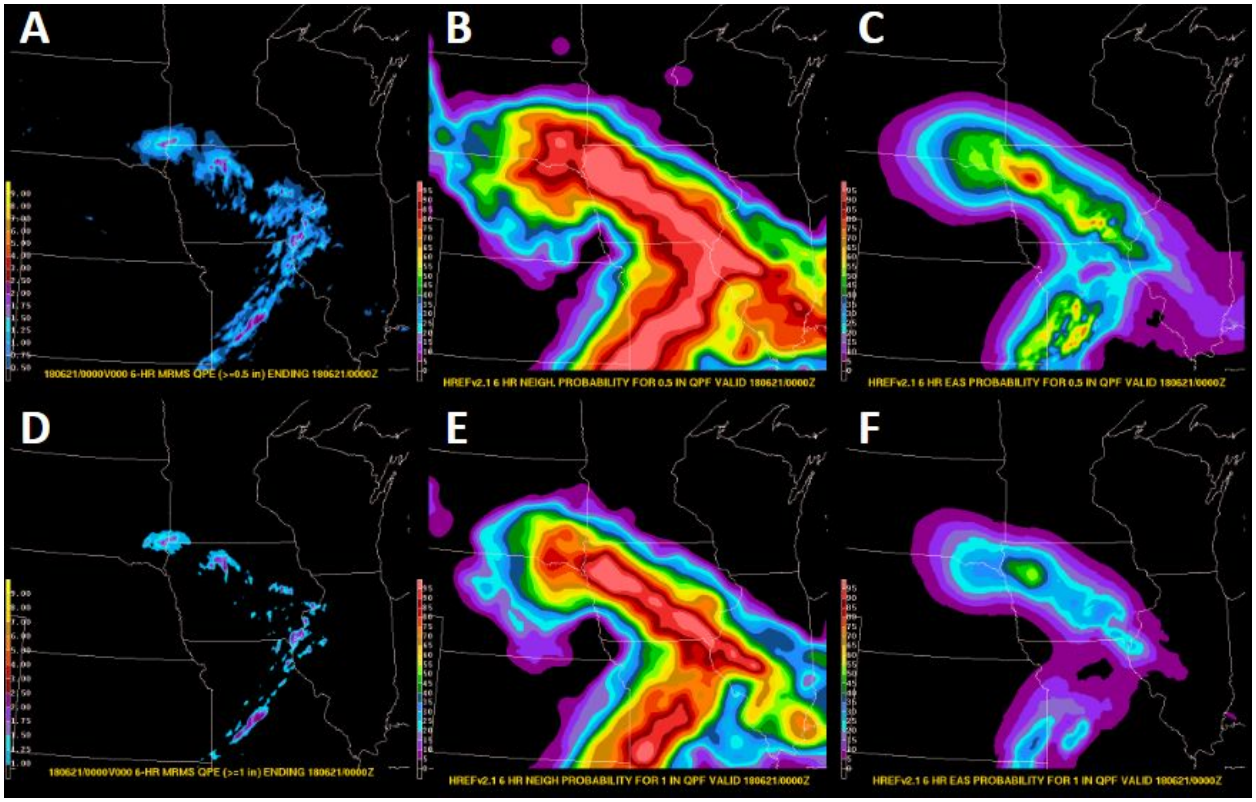


Figure 24. *(A) 6 hour MRMS-GC QPE >=0.5 in. values; (B) HREFv2.1 6 hour 40 km neighborhood probabilities for 0.5 in. QPF; (C) HREFv2.1 6 hour EAS probabilities for 0.5 in. QPF; (D) 6 hour MRMS-GC QPE >=1.0 in. values; (E) HREFv2.1 6 hour 40 km neighborhood probabilities for 1.0 in. QPF; (F) HREFv2.1 6 hour EAS probabilities for 1.0 in. QPF; all valid 18 UTC June 20 - 00 UTC June 21, 2018.*

Findings

Comparing the two different probability methods from the HREFv2.1 generated abundant discussion and opinions among the FFaIR participants. The QPF threshold and the synoptic setup were two of the biggest factors that forecasters weighed when giving their opinions on the method. At the lowest threshold of 0.5 in./6 hours, participants favored the EAS method because the overall coverage of the higher probabilities was typically more narrow and more focused, especially in strongly forced scenarios. Figure 25 is an example of a strong synoptic scale cold front that moved through the Mid-Atlantic and New England where at a 0.5 in./6 hours, participants felt that the EAS probabilities in panel C did a better job at focusing the higher probabilities rather than having a very broad >=95% area like the neighborhood probabilities in panel B.

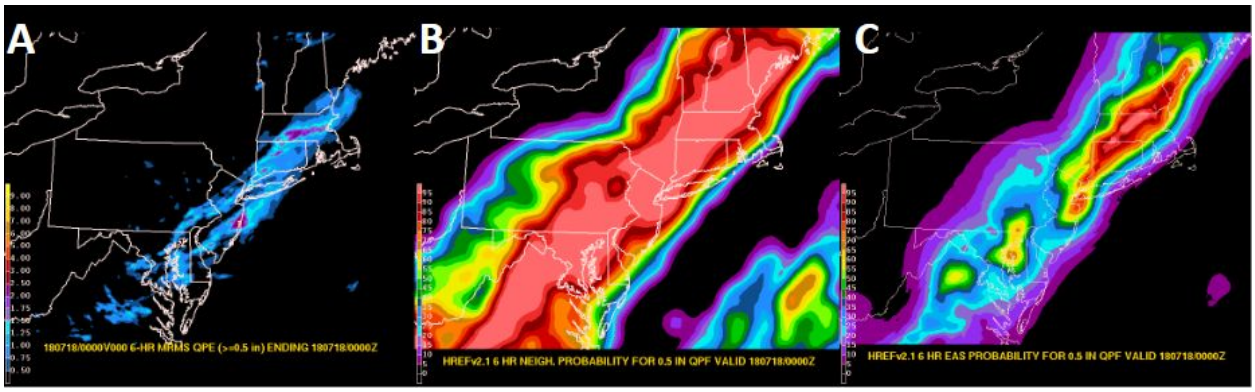


Figure 25. *(A) 6 hour MRMS-GC QPE >=0.5 in. values; (B) HREFv2.1 6 hour 40 km neighborhood probabilities for 0.5 in. QPF; (C) HREFv2.1 6 hour EAS probabilities for 0.5 in. QPF all valid 18 UTC July 17 - 00 UTC July 18, 2018.*

In more weakly forced regimes, such as the Southwest monsoon shown in Figure 26, the EAS (panel C) was preferred again at 0.5 in./6 hours. Despite the probabilities being lower for the amount of activity, participants liked how there was less noise than the neighborhood method in panel B where there are many individual high probability areas throughout the domain with no clear signal.

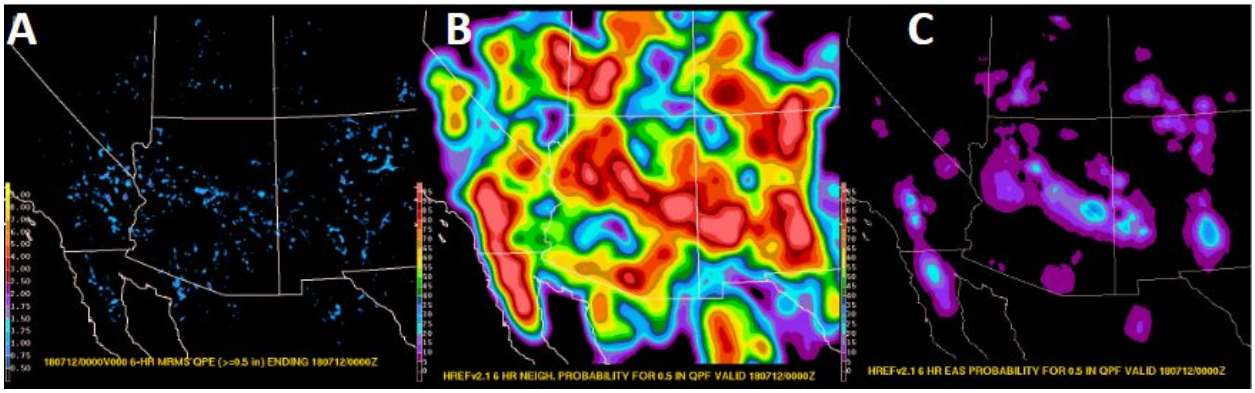


Figure 26. *(A) 6 hour MRMS-GC QPE >=0.5 in. values; (B) HREFv2.1 6 hour 40 km neighborhood probabilities for 0.5 in. QPF; (C) HREFv2.1 6 hour EAS probabilities for 0.5 in. QPF all valid 18 UTC July 11 - 00 UTC July 12, 2018.*

At the higher threshold of 1.0 in./6 hours, participants preferred the neighborhood probabilities more often as the EAS probabilities often struggled to produce high enough probabilities. This is likely due to the EAS probabilities being based on point probabilities, resulting in a lower chance of getting a hit among the ensemble members at higher QPF thresholds Often the EAS probabilities did not go higher than 30%. Figure 27 is an example where the neighborhood probabilities in panel B were much preferred over the EAS probabilities in panel C. Figure 28 is another example in a more weakly forced regime where the EAS probabilities (panel C) at 1.0 in./6 hours give less than 5% probabilities in Georgia and Alabama where widespread areas of rainfall of at least an inch was observed in the QPE. The neighborhood probabilities in this

example (panel B) were much preferred by the participants due to the coverage and values of the probabilities in Georgia and Alabama.

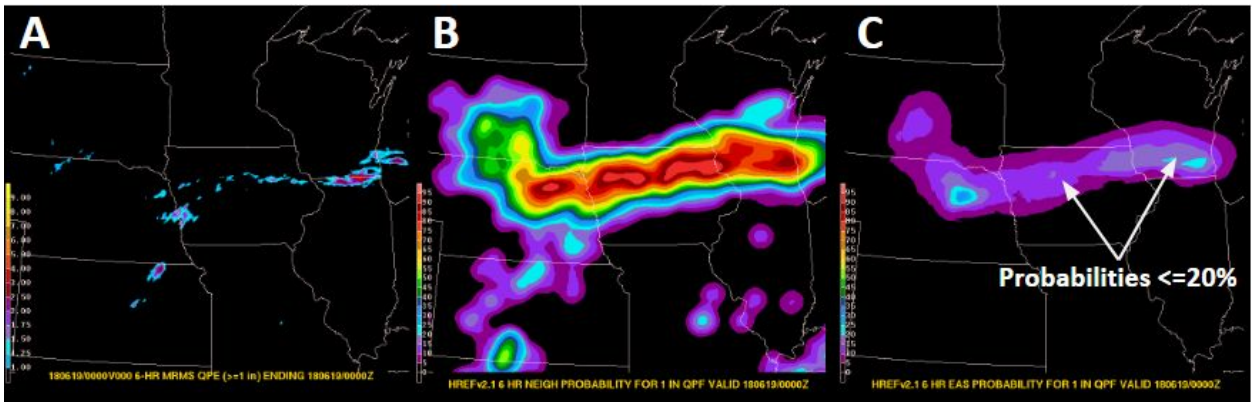


Figure 27. *(A) 6 hour MRMS-GC QPE >=1.0 in. values; (B) HREFv2.1 6 hour 40 km neighborhood probabilities for 1.0 in. QPF; (C) HREFv2.1 6 hour EAS probabilities for 1.0 in. QPF all valid 18 UTC June 18 - 00 UTC June 19, 2018.*

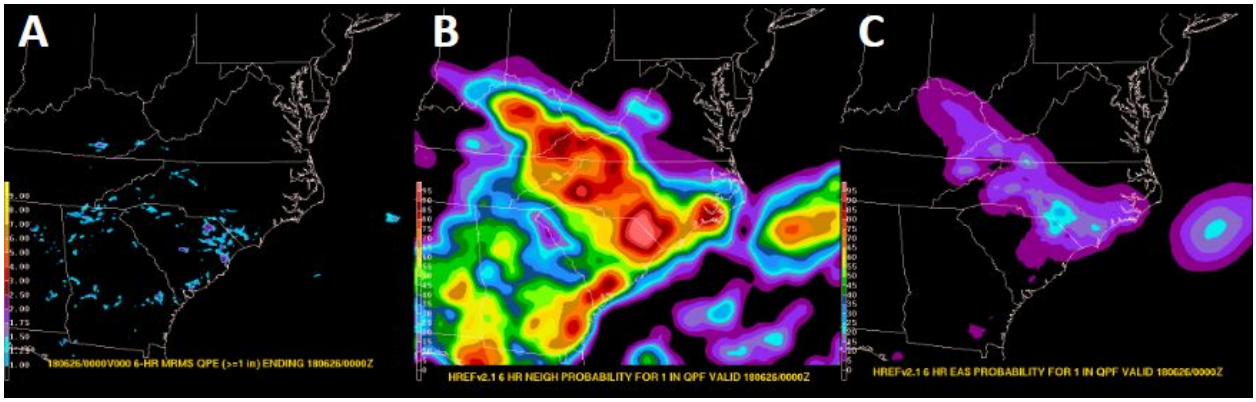


Figure 28. *(A) 6 hour MRMS-GC QPE >=1.0 in. values; (B) HREFv2.1 6 hour 40 km neighborhood probabilities for 1.0 in. QPF; (C) HREFv2.1 6 hour EAS probabilities for 1.0 in. QPF all valid 18 UTC June 25 - 00 UTC June 26, 2018.*

Recommendations

WPC-HMT recommends that both the neighborhood and EAS probability methods be made available as options for forecasters. Both proved to have positive and negative feedback for different situations and overall participants felt each could contribute positively to the forecast process. The EAS method was generally favored at lower thresholds and in both strongly and weakly forced regimes. In weakly forced patterns, the EAS method often had probabilities that were too low but still focused on the correct areas whereas it was often difficult to discern focus areas using the neighborhood probability method at lower thresholds. At a higher threshold of at least 1.0 in./6 hours, the neighborhood probability method was more generally preferred as the probabilities in the EAS method were often below 30% in all situations. Ideas for further research and development include exploring changing the similarity criteria in the EAS method to allow for higher probabilities for higher QPF thresholds and possibly differing the similarity criteria by regions in the United States.

NEWS-e Performance and Results

Multiple cycles of the NEWS-e were subjectively and objectively evaluated during the 2018 FFaIR. The NEWS-e was run every hour beginning at 18 UTC and provided hourly output out to six hours. The domain was limited to a 900 km square area chosen daily by the FFaIR participants. For the evaluation, the 3 hour PMM QPF valid 21-00 UTC was subjectively evaluated on a scale from 1 (very poor) to 10 (very good) from the 18, 19, 20, and 21 UTC cycles to determine whether the newer cycles benefited from the real-time observational data that feeds into the ensemble system. More detailed information on the NEWS-e can be found in Appendix C. Figure 29 is an example image of how the NEWS-e data was presented to participants during subjective evaluation. The 3 hour MRMS-GC QPE was in the top left, 18 UTC 3 hour PMM QPF top center, 19 UTC 3 hour PMM QPF top right, 20 UTC 3 hour PMM QPF bottom center, and 21 UTC 3 hour PMM QPF bottom right.

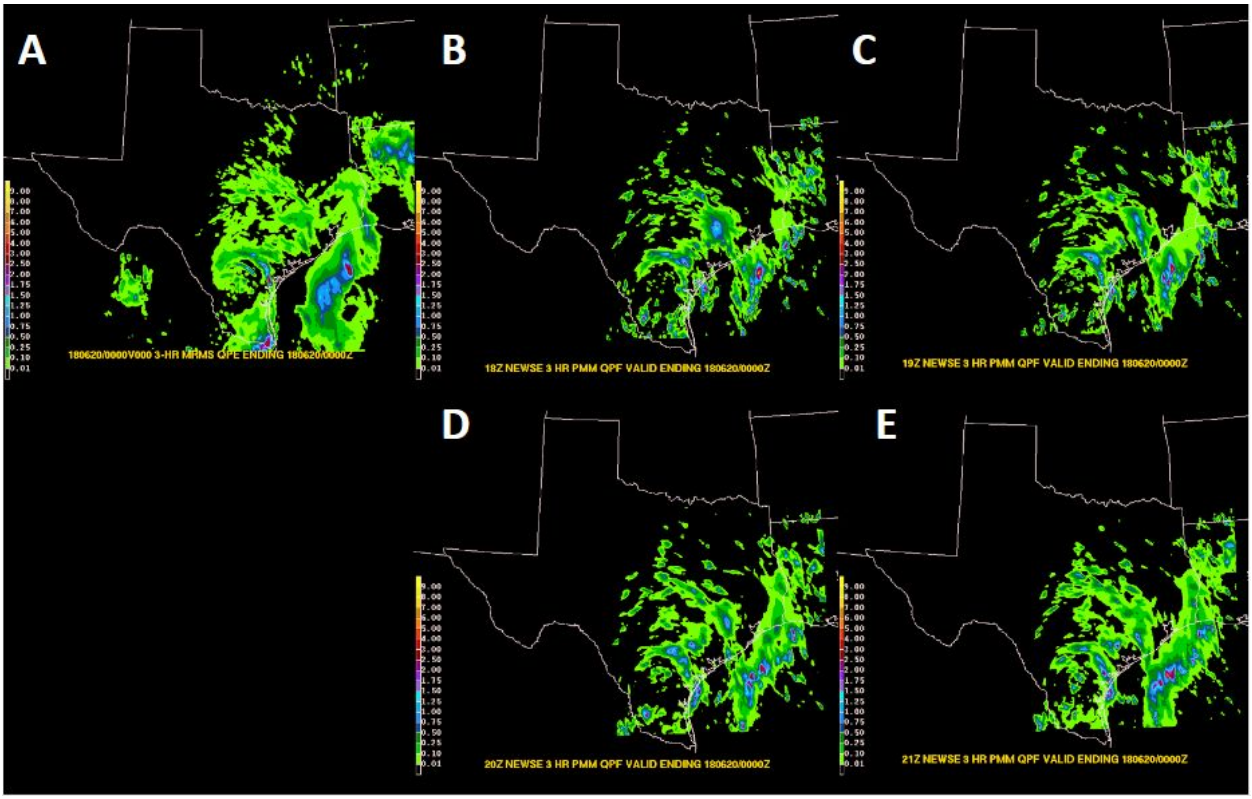


Figure 29. *(A) 3 hour MRMS-GC QPE valid 21 UTC June 19 - 00 UTC June 20, 2018; (B) 18 UTC 3 hour PMM QPF, (C) 19 UTC 3 hour PMM QPF, (D) 20 UTC 3 hour PMM QPF, and (E) 21 UTC 3 hour PMM QPF all valid over the same time period.*

Findings

All four NEWS-e cycles were rated similarly in the subjective evaluation. Figure 30 shows a box plot of the 18 to 21 UTC NEWS-e subjective evaluation scores that were evaluated over the four week experiment. The latest cycle, 21 UTC, had the highest average score of 6.99 out of 10. The 20 UTC cycle had the next highest average score at 6.69 followed by the 18 UTC cycle at 6.62 and finally the 19 UTC cycle at 6.53. The 21 UTC cycle also had the smallest standard

deviation of all the cycles at 1.37 compared to the 18 UTC cycle that had the largest at 1.48. Participants often found very subtle differences between the four cycles which led to scores being very similar. The two later cycles from both the scores and comments were found to be consistently better, however there were a few cases in which the two later cycles scored worse. An example of one such instance is Figure 31 where the NEWS-e produces too much QPF in the main band of rainfall circled in red in each newer cycle.

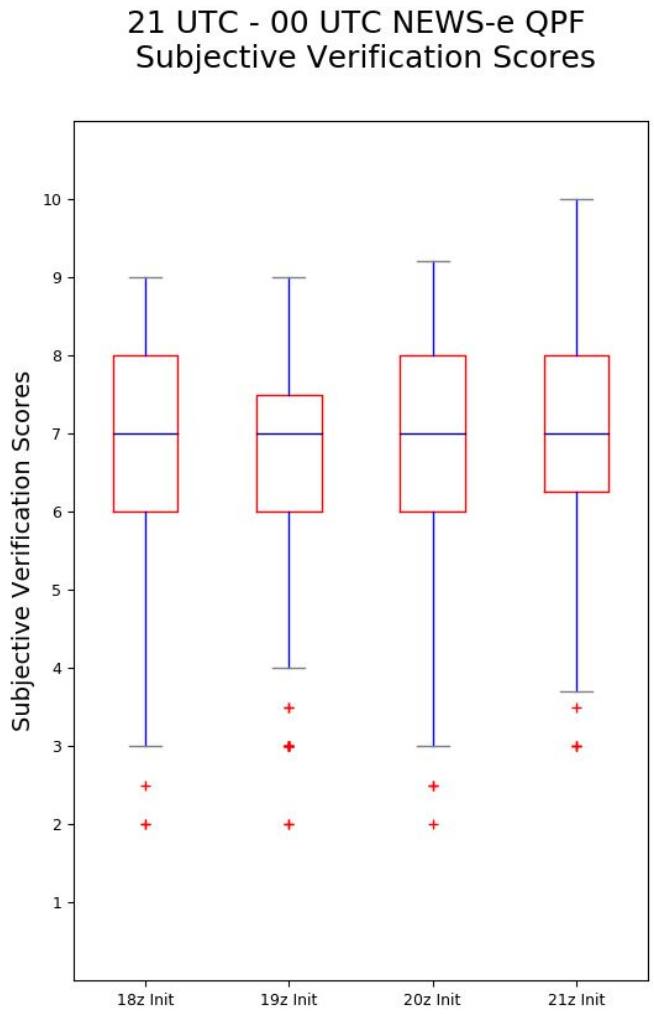


Figure 30. *Box plot of the subjective scores for the 18/19/20/21 UTC cycles of the NEWS-e throughout the four weeks of the experiment. Red plus symbols denote outliers.*

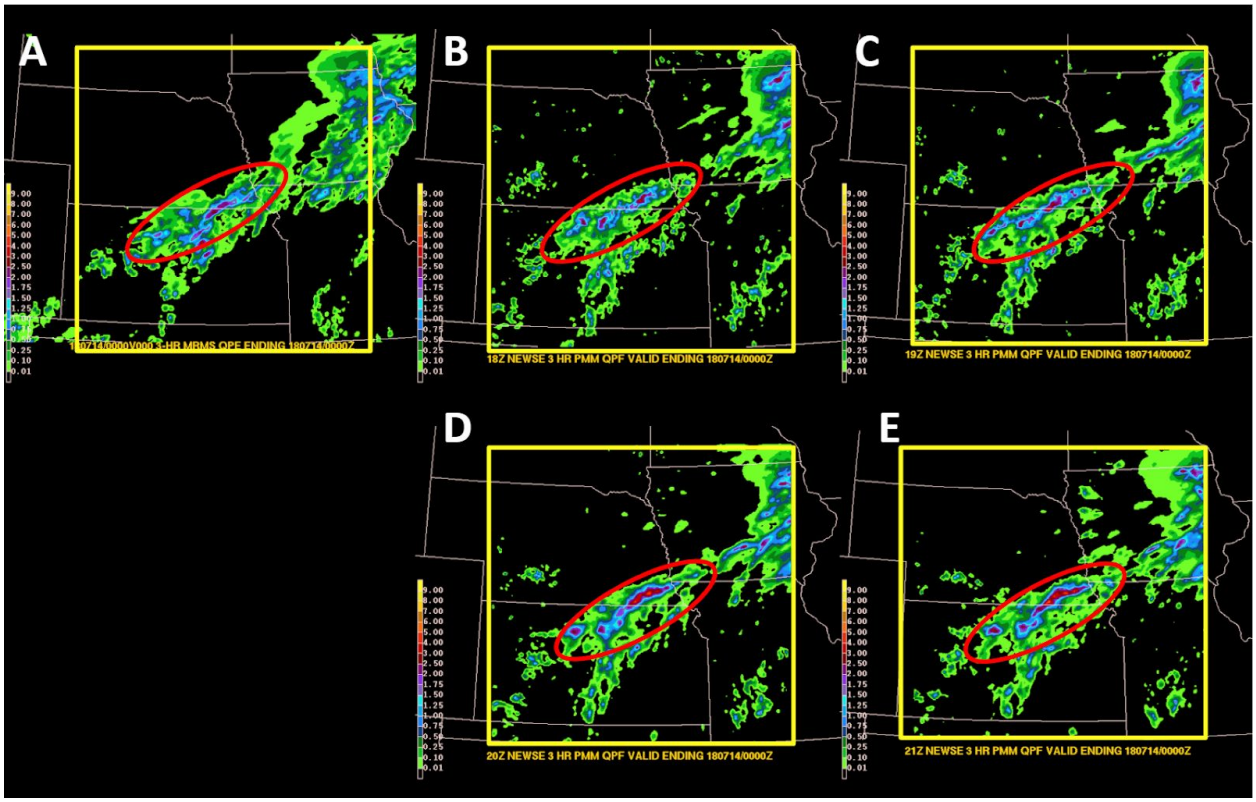


Figure 31. *(A) 3 hour MRMS-GC QPE valid 21 UTC July 13 - 00 UTC July 14, 2018; (B) 18 UTC 3 hour PMM QPF, (C) 19 UTC 3 hour PMM QPF, (D) 20 UTC 3 hour PMM QPF, and (E) 21 UTC 3 hour PMM QPF all valid over the same time period. The yellow box indicates the extent of the NEWS-e domain for this day.*

The case in Figure 31 was unusual and participants often noted that if the NEWS-e overpredicted in earlier cycles it would adjust to lighter values by the 20 UTC and 21 UTC model runs. It was also noted that in lighter precipitation events the model struggled to produce areas of precipitation early on, but tended to adjust correctly with time. Despite later runs adjusting QPF correctly towards heavier or lighter precipitation forecasts, the NEWS-e 3 hour PMM tended to remain overall too heavy in convective areas and too light in more weakly forced or stratiform regimes. Figure 32 shows performance diagrams valid for all four weeks of the experiment for all four NEWS-e cycles at four different thresholds: (A) 0.1 in., (B) 0.25 in., (C) 0.50 in., and (D) 1.0 in.

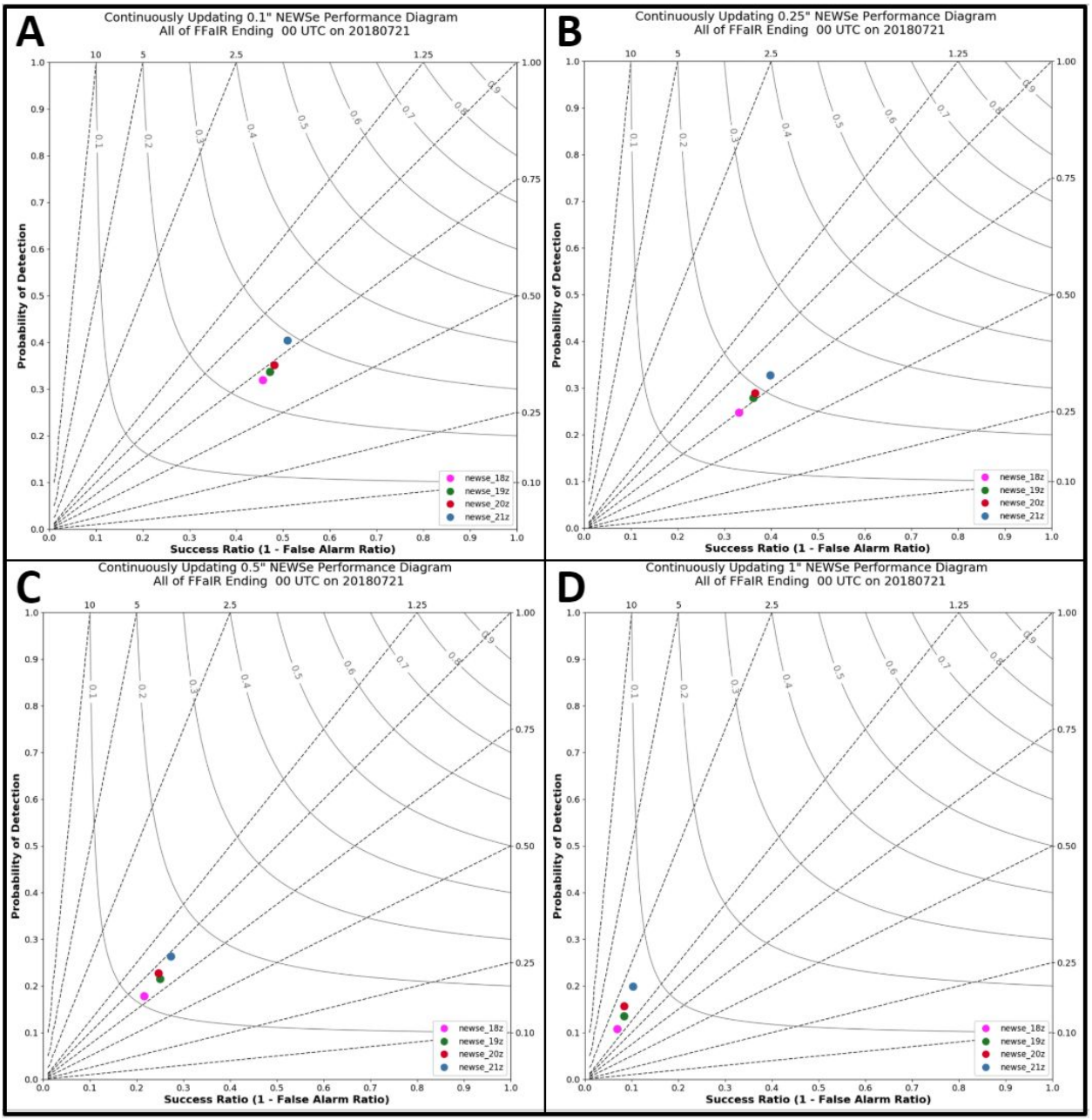


Figure 32. *Performance diagrams for the 18 UTC (magenta), 19 UTC (green), 20 UTC (red), and 21 UTC (blue) valid for all four weeks of FFaIR at (A) 0.1 in., (B) 0.25 in., (C) 0.5 in., and (D) 1.0 in.*

The objective verification supports the participant feedback showing that at lighter thresholds of 0.1 in. and 0.25 in., all four cycles have a noticeable dry bias and at the higher threshold of 1.0 in., all four cycles have a wet bias. The objective results also show that each newer cycle had an overall better CSI score, where 18 UTC is magenta, 19 UTC is green, 20 UTC is red, and 21 UTC is blue in the figure. Figure 33 shows the CSI values of the 18/19/20/21 UTC NEWS-e cycles at the 0.1 in. (blue), 0.25 in. (red), 0.5 in. (yellow), and 1.0 in. (green) thresholds.

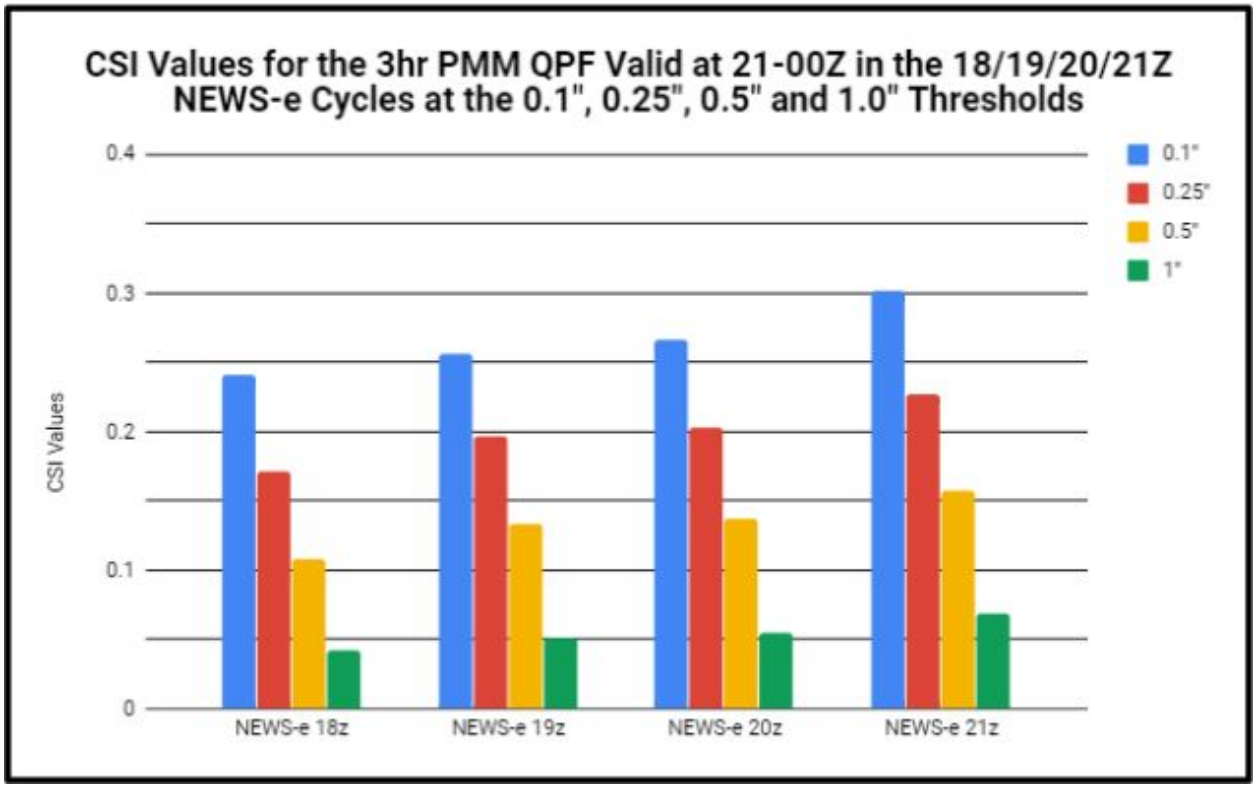


Figure 33. *CSI values for the 18/19/20/21 UTC cycles of the NEWS-e 3 hour QPF valid 21-00 UTC throughout all four weeks of the experiment at 0.1 in. (blue), 0.25 in. (red), 0.5 in. (yellow), and 1.0 in. (green).*

Recommendations

WPC-HMT recommends the NEWS-e for further testing. One major limitation this year was that the NEWS-e was initialized off of the HRRRE which did not cover the full CONUS in its domain. This prevented any NEWS-e domains west of the Rocky Mountains and therefore no Southwest monsoon cases were able to be evaluated. WPC-HMT would like to see Southwest monsoon cases from the NEWS-e as they are important in terms of impacts and difficult in terms of predictability. Cases like the one highlighted in Figure 31, where the newer cycles were worse than the earlier cycles, should be examined more closely to see if a cause can be determined. Because of the similarity between four consecutive cycles that was seen in this year's subjective evaluation, it is recommended that a future evaluation examines cycles with greater lead time in order to examine the effects of the assimilation of observations. It is also recommended that future examinations explore some of the probabilistic products.

5. Hydrologic Guidance Results

National Water Model Experimental Products Feedback

Due to the complexity of the National Water Model (NWM) access and products available during the 2018 FFaIR Experiment, the collection of data focused on comments regarding performance and utility to the flood forecast process rather than scoring and metrics. This

section will highlight the collection of comments from the participants over the course of the 2018 FFaIR Experiment. Comments focused on the High Flow Potential, High Flow Probability, and Peak Flow Arrival Time products.

Overall comments on the NWM included a need for the products to be aligned scientifically and temporally so they can be used in better concert with one another. The varying calculations, calibration, forcings, colors and units resulted in participant confusion and steep learning curves that took time away from the overall forecast process. Requests were made for a more flexible legend that can be customizable and take up less space, the capability to filter stream reaches that are already reacting to high flow and the highest order mainstem rivers, more information about water exceeding bankfull (inundation), more overlays such as hydrographs and USGS stream gauges, and more variable QPF forcings other than the HRRR and GFS. Forecasters desire stream information that has a direct relationship with impacts, which is generally not provided by flow anomalies. It was noted often that the relationship of stream response to flash flooding is not always one to one, so these products cannot aid in the prediction of flash flooding. Many felt the NWM products provided additional situational awareness but can rarely be used as key guidance for flash flood forecasting.

High Flow Potential

The High Flow Potential product depicts stream reaches expected to be above their 1.5-year recurrence flow as an estimate of bankfull discharge, an example of which is shown in Figure 32. The short range forecast is forced by the operational deterministic HRRR QPF.

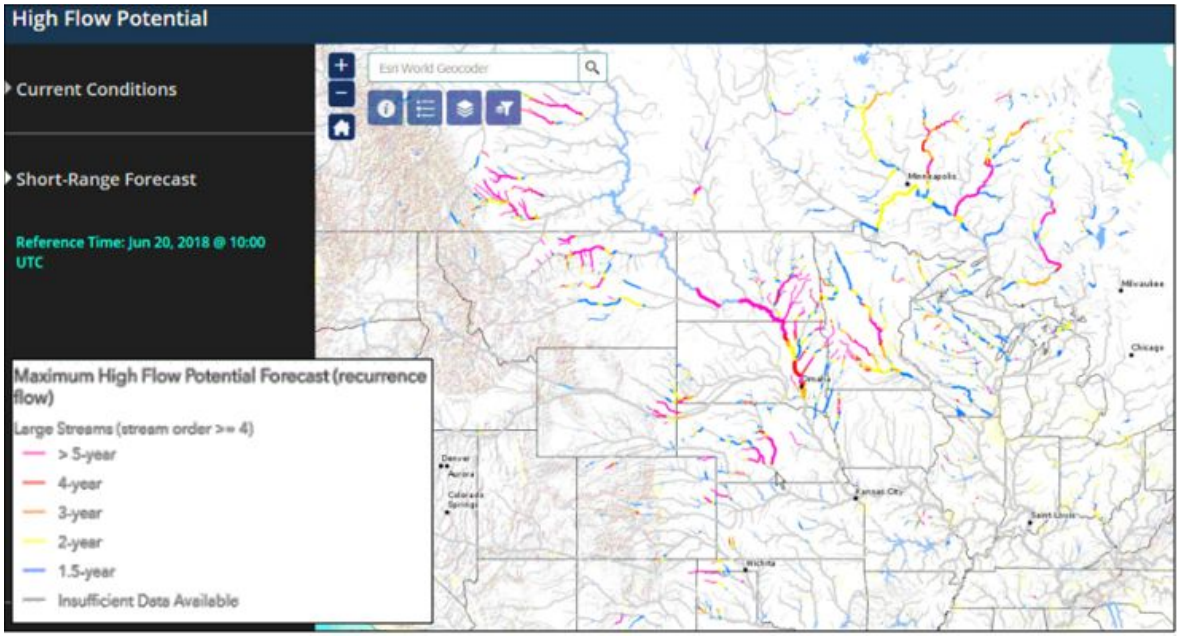


Figure 32. *An example of a short-range High Flow Potential forecast valid 10 UTC June 20, 2018. Note the magenta streams currently above the 5 year recurrence flow at the time of the forecast initialization.*

Findings

The High Flow Potential product was deemed to have the most forecast utility among the available NWM products due to its relationship with streamflow ARIs, improved calibration against the 24-year NWM climatology, and situational awareness. The potential for streamflow to exceed ARI thresholds draws the eye to areas vulnerable to flooding. To maximize its utility, it is best used in concert with other products such as soil moisture, QPF, and FFG. For quick decision making related to triggering flash flood warnings, forecasters found this less useful.

Recommendations

High Flow Potential was often confused with High Flow Probability. A more intuitive name suggestion is "High Flow Magnitude." The ability to click on and view associated hydrographs as a reference would be preferred to increase confidence. Longer temporal range with the short-term forecast would be preferred for those cycles where the HRRR goes out 36 hours (0000 UTC, 0600 UTC, 1200 UTC, and 1800 UTC). References to flood stage would add value. A rate of change in flow (perhaps a difference field) relative to flood benchmarks and/or previous forecast runs is desired to pull out more targeted information related to flood risk.

High Flow Probability

The High Flow Probability product depicts the probability that stream reaches will be at or above their 1.5-year recurrence flow between 6 and 8 hours beyond the forecast initialization time. An example of this product is shown in Figure 33. The short-range High Flow Probability forecast is forced by the QPF derived from a time-lagged HRRR ensemble comprised of the past 9 operational HRRR runs.

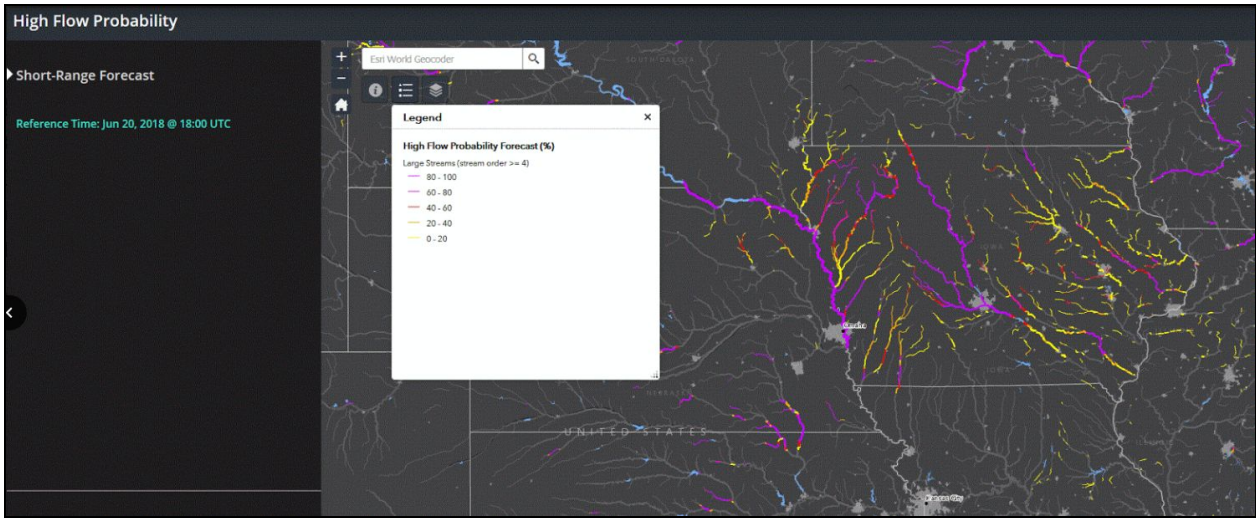


***Figure 33**. An example of a High Flow Probability short-range forecast initialized 18 UTC June 20, 2018 and valid between 00 UTC-02 UTC June 21, 2018.*

Findings

The NWM's version of the time-lagged ensemble HRRR as a QPF forcing increased forecaster confidence in the High Flow Probability as opposed to the deterministic NWM suite of products. Participants agreed that the probabilities highlighted the regions of hydrologic concern over QPF alone, especially when zooming in to smaller reaches. Some forecasters felt they could

trust the spatial distribution of the QPF to fall in the correct basin(s); this product provides useful detail to the county level. However, others felt that heavy rainfall and hydrologic response is too regional for this product to be useful.

Recommendations

As stated earlier, High Flow Probability was often confused with High Flow Potential, therefore a more intuitive name is recommended. Participants felt the arbitrary 6-8 hour valid time is limiting and would prefer flexibility with the time range selection. Although the time-lagged HRRR is an improvement on a single deterministic run and ensemble probabilities are preferred, participants desired a different ensemble forcing, namely the HREF.

Peak Flow Arrival Time

The Peak Flow Arrival Time product, an example of which is shown in Figure 34, depicts the time when stream reaches are expected to be at or above their peak flow based on the 1.5-year recurrence flow. The short range product is forced by the deterministic HRRR QPF.

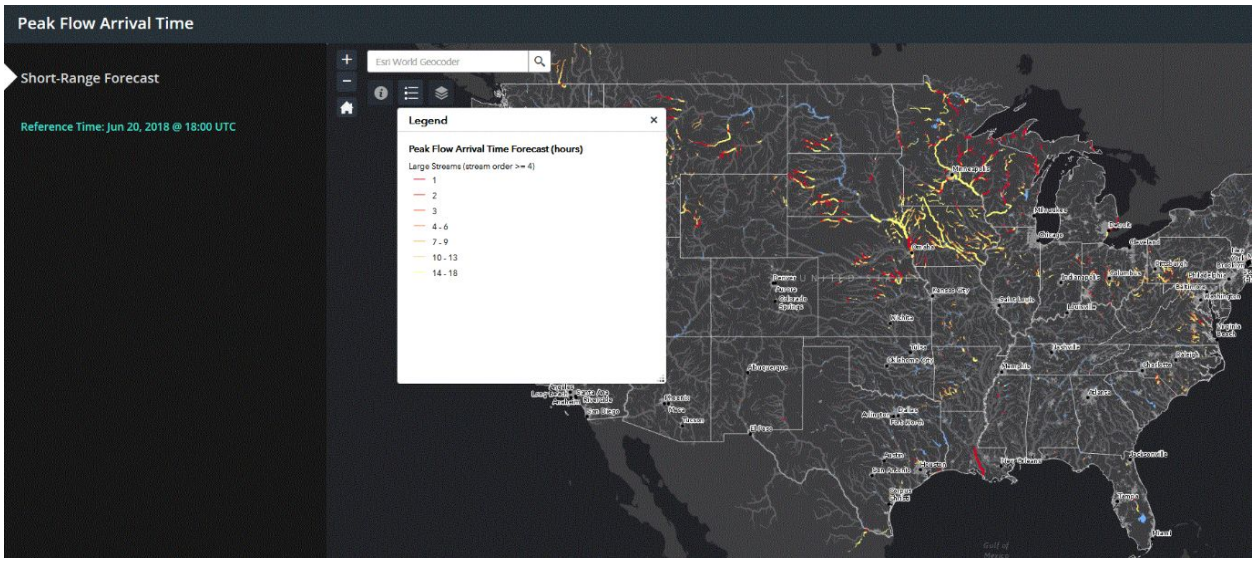


Figure 34. *An example of a Peak Flow Arrival Time short-range forecast initialized 18 UTC June 20, 2018 and valid for 18 hours.*

Findings

The Peak Flow Arrival Time had a small amount of forecast value for providing the ability to zoom down to small reaches and identify detailed stream response ("hot spots") at small scales, dependant upon the HRRR QPF. The forecasters noted that the metadata contained within the streams reaches, especially the "Time to Return to Normal Flow" estimation, was useful when determining the time scale of the high flow and possible flash flood vulnerability of the basin. This product was best used in tandem with the High Flow Arrival Time, ARIs, FFG, and QPF products rather than on its own. The information provided is not enough to be a useful product when discerning or communicating potential flooding impacts. Participants felt that given the time scales within which they have to assess flood risk and make decisions, there are too many

other products at their disposal that give them the information needed. Therefore, they are unlikely to spend time with the Peak Flow Arrival Time.

Recommendations

Participants were often distracted by the areas already at high or peak flow at the initialization time and desired a way to filter areas that were newly impacted or of highest risk. Embedded real-time hydrographs and stream gauges would add value. A density plot at the CONUS scale would aid forecasters in more quickly identifying areas of highest concern.

CSU-MLP First Guess Field for Days 1, 2 and 3 Performance and Results

The Colorado State University Machine Learning Probabilities (CSU-MLP) provides a first guess for the 24-hour Excessive Rainfall Outlook (ERO) for Days 1, 2 and 3. The product provides a probabilistic outlook for QPF exceeding the 1-year, 24-hour ARI using thresholds of 5, 10, 20, and 50%. An example of the Day 1 first guess field is shown in Figure 35. Day 1 utilizes the deterministic QPF from the deterministic NSSL-WRF and the Days 2 and 3 utilize the ensemble mean QPF from the GEFS Reforecast (GEFS-R). For more details on the CSU-MLP, please refer to the section within Appendix C.

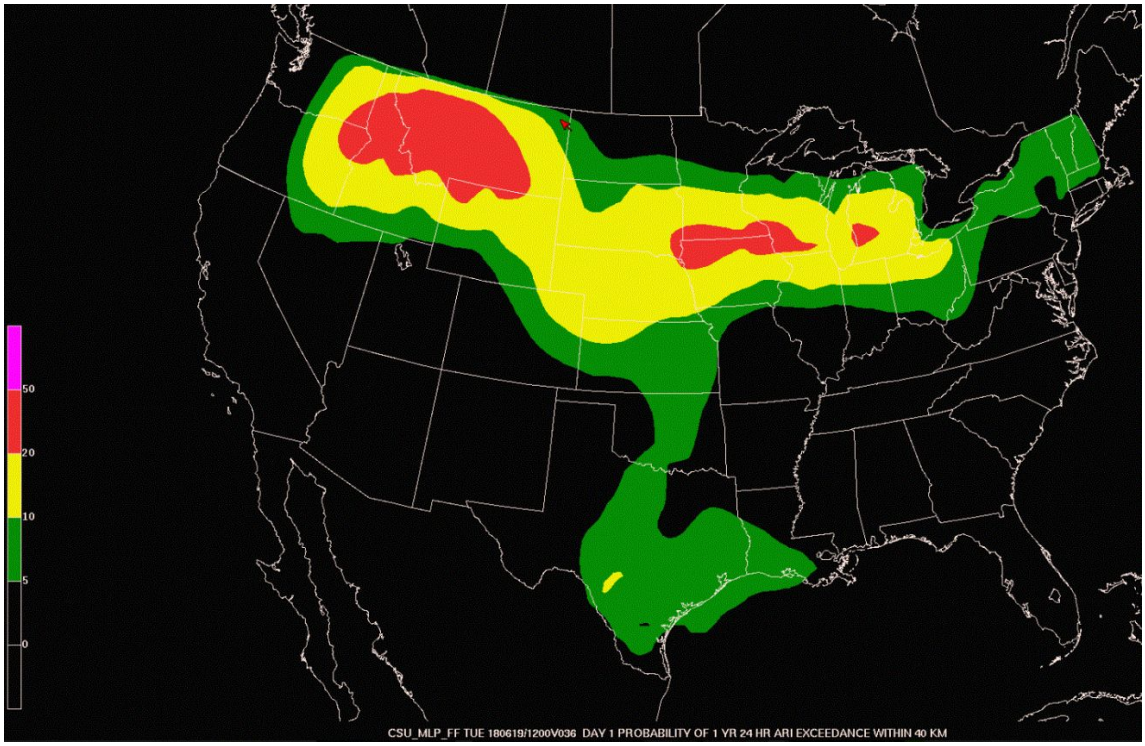


Figure 35. *An example of the CSU-MLP Day 1 ERO First Guess valid at 12 UTC on June 19, 2018.*

The participants used the ERO first guess as guidance when creating the experimental Day 1 ERO. It was also subjectively evaluated along with the Day 2 and Day 3 on a scale of 1 (very poor) to 10 (very good). The overall quality of the product, including areal coverage of the

probabilistic contours and their values, were evaluated against the UFV system described in the verification section.

Findings

Figure 36 is box plot showing the subjective scores over the course of the experiment for the Day 1, Day 2, and Day 3 CSU-MLP First Guess fields. The Day 1 CSU-MLP First Guess ERO average score was 6.27 out of 10. Probabilities in the Southwest, central plains and the Northeast were overall well-captured. However, probabilistic values tended to be often underdone in Texas, along the Gulf Coast and the southeastern coastal regions. Probabilities were over-confident throughout the upper plains, including Montana and the Dakotas. Areal coverage on many days tended to be too narrow. Participants noted that in some areas, multiple small contours would cluster together and create a distracting forecast when a single contour would have been easier to interpret. An example of this can be seen in Day 1 CSU-MLP First Guess field valid for 12 UTC July 10, 2018 in Figure 37 where there are six separate 10% contour lines in the Southwest. Also noted were risk areas having some displacement issues, most often to the east of the observed heavy rainfall and flooding.

Verification of the CSU-MLP was performed throughout the 4-week experiment. Figure 38A shows the probability of being in a slight probability contour from the Day 1 CSU-MLP forecast compared to (B) in the Day 1 operational WPC ERO. Three main areas stand out when compared to the operational ERO: Montana, the Central Plains of North and South Dakota and Nebraska, and the Southwest United States. Figure 39 shows the average fractional coverage of the Day 1 CSU-MLP First Guess ERO and the operational Day 1 ERO using FFG only and the UFV system as verification. The Day 1 CSU-MLP First Guess ERO falls within the defined probability definitions for each category, except for the marginal category using FFG only verification. The fractional coverage for all categories is on the low end of the categorical definition. The operational EROs are much higher than the Day 1 CSU-MLP ERO and often exceed the upper bound of the categorical definition.

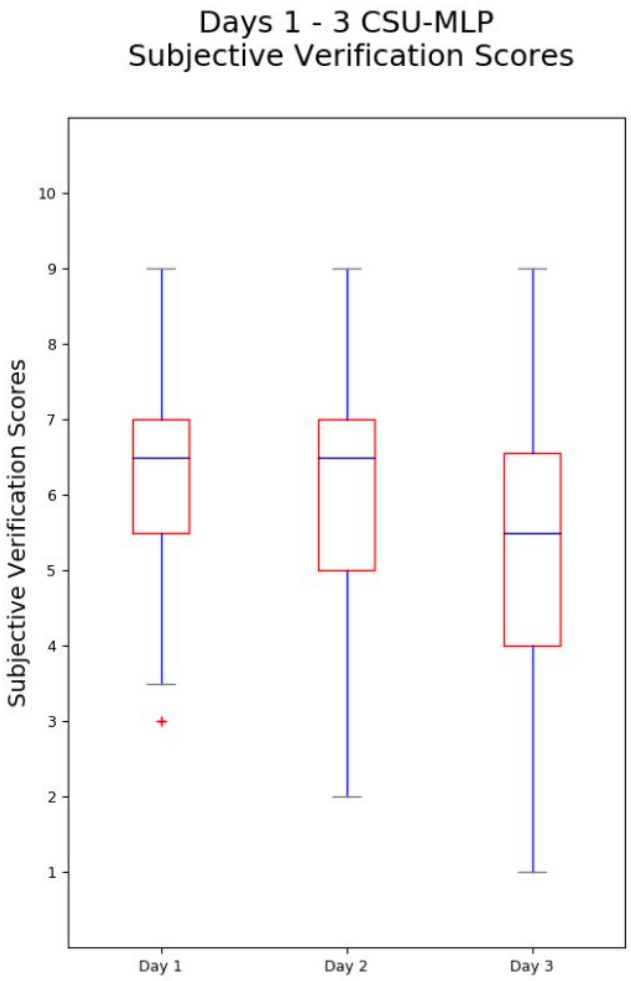


Figure 36. *Box plot for the total subjective scores over the four weeks of the experiment for the CSU-MLP 24-hour probabilistic ERO First Guess product for Days 1, 2, and 3. Red plus symbols denote outliers.*

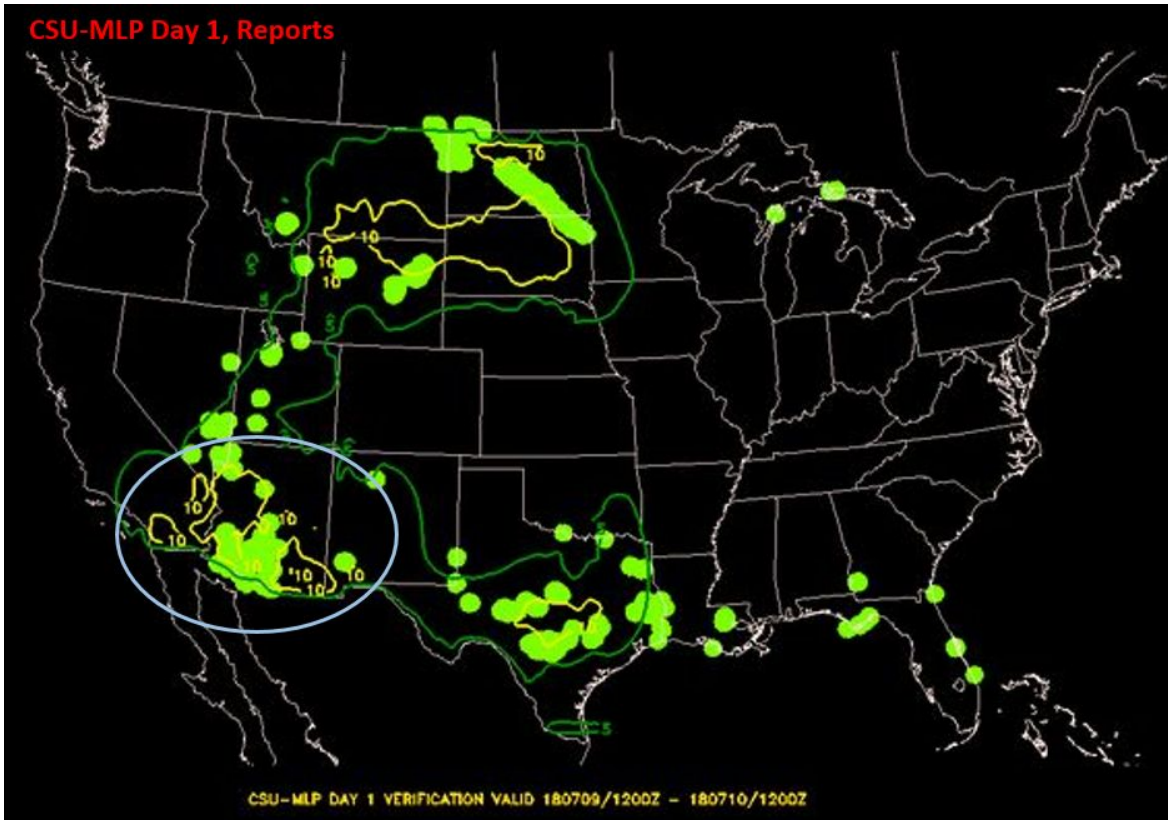


Figure 37. *An example of the 24-hour Day 1 CSU-MLP ERO First Guess (contoured) overlaid on reports from the UFV system (green circles) valid at 12 UTC July 10, 2018. Area highlighted shows multiple 10% contours clustered in the southwest rather than a broader single contour.*

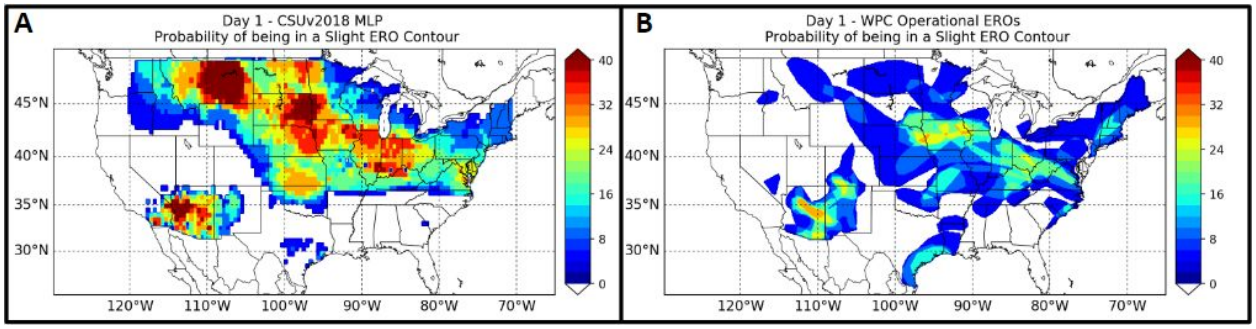


Figure 38. *(A) Probability of being in a "slight risk" Day 1 CSU-MLP First Guess ERO contour and (B) "slight risk" operational ERO contour over the four week experiment.*

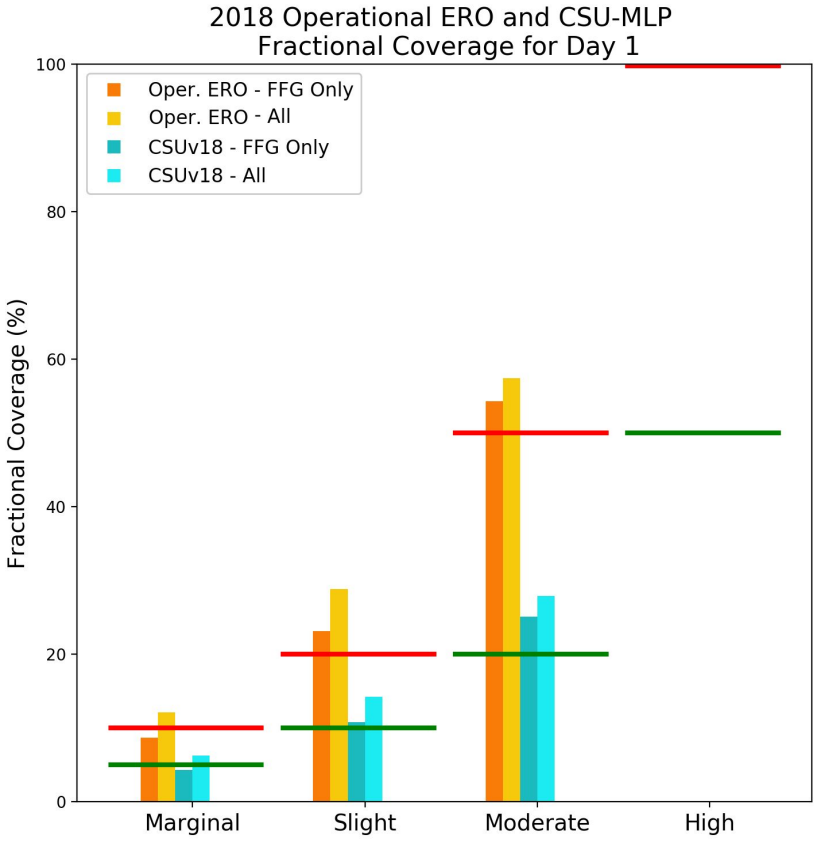


Figure 39. *Fractional coverage of the 2018 Day 1 CSU-MLP First Guess ERO (blue) and operational EROs (orange) issued over the same time period for each probabilistic category. Green horizontal lines represent the lower defined bound for each threshold and red horizontal lines represent the highest defined bound.*

The Day 2 CSU-MLP First Guess ERO, which uses the GEFS-R, had an average subjective score of 6.18 out of 10, just 0.09 lower than the Day 1 forecast. The participants often favored the Day 2 product over any other, commenting that it out-performed the Day 1 in areal coverage and confidence. Figure 40 is an example valid 12 UTC June 18 - 12 UTC June 19, 2018 where the Day 2 CSU-MLP First Guess (B) was rated higher than the Day 1 First Guess (A).

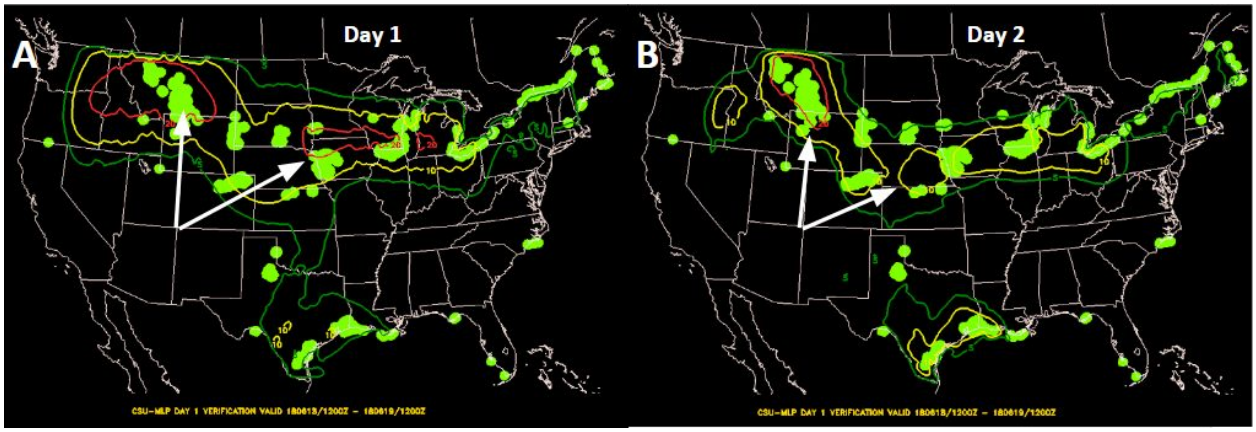


Figure 40. *(A) Day 1 CSU-MLP First Guess ERO (contour) and (B) Day 2 CSU-MLP First Guess ERO (contour) with UFV reports (green circles) valid 12 UTC June 18 - 12 UTC June 19, 2018.*

However, there were still similar issues with probabilistic values and displacement in the Day 2 First Guess. The Day 2 product was often over-confident through the central and upper Plains. Some days featured several small moderate contours that were not justified. The Southwest was well-captured most days. The Gulf coast and southeastern coastline struggled with areal extent being too narrow and slightly underdone, but was better captured than at Day 1.

The Day 3 CSU-MLP First Guess ERO, as expected with a longer lead time, had the lowest average score of 5.39 out of 10. Whereas not significantly degraded from the Day 2, the Day 3 struggled many days with areal extent, orientation, and probabilistic values. An example of some of these issues can be seen in Figure 41 for a Day 3 forecast valid 12 UTC June 21, 2018. Marginal events were more difficult to capture than strongly-forced synoptic events, for which it occasionally did well. Contours, particularly the moderate, tended to be confusing with many small contours across the CONUS rather than one, continuous forecast. As opposed to the narrow coverage from Days 1 and 2, the Day 3 marginal contours tended to be too broad. Participants noted that the Day 3 First Guess field correctly adjusted probabilities lower in some cases where the Day 2 First Guess probabilities were viewed as overconfident.

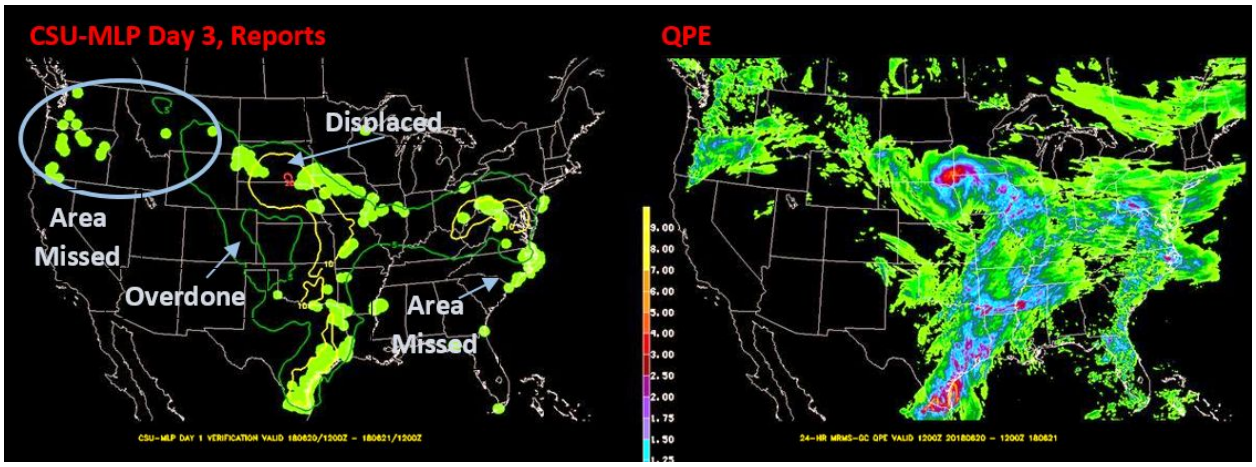


Figure 41. *An example of the 24-hour Day 3 CSU-MLP ERO First Guess valid 12 UTC 20180621. Several regions are highlighted where the CSU-MLP struggled with location, areal extent, and confidence.*

Figure 42 shows the probability of being within a "slight" probability contour throughout the entire experiment from the (A) Day 2 CSU-MLP First Guess ERO, (B) the Day 2 operational ERO, (C) the Day 3 CSU-MLP First Guess ERO, and (D) the Day 3 operational ERO. The probability of being within a slight contour from the first guess field was much higher than the operational ERO at both Day 2 and Day 3. Areas with especially high probabilities include Montana, Wyoming, South Dakota and Nebraska as well as further east into Illinois and Indiana and the Southwest U.S. in Arizona. Day 2 and Day 3 CSU-MLP First Guess EROs look similar overall. There are some subtle shifts of slight coverage further westward in the Southwest, shifts further eastward in Virginia, and generally higher coverage through Illinois, Indiana, and Ohio when moving from the Day 3 forecast to Day 2.

Average fractional coverage of the Day 2 CSU-MLP First Guess ERO and operational WPC ERO is shown in Figure 43. At Day 2, the CSU-MLP forecasts are all below the lowest probability threshold for each risk category with the exception of the slight and moderate categories when using the UFV system for verification. The operational ERO at Day 2 is the opposite and is at the high end or exceeds the marginal, slight, and moderate thresholds. These results are similar for the Day 3 forecasts shown in the same figure. Figure 44 shows the Day 2 and 3 CSU-MLP First Guess ERO BSS referenced against the Day 2/3 operational EROs. Areas where the BSS is positive, the CSU-MLP First Guess EROs performed better than the operational EROs and worse when the BSS is negative. The BSS shows that through the first half of the experiment, the CSU-MLP forecasts were neutral to slightly worse than the operational EROs. There was a large drop off in skill during third week with some improvement in the fourth week, but generally the CSU-MLP forecasts were worse in the second half of FFaIR. Figure 45A shows the area under the relative operating characteristic (AuROC) for the Day 2 CSU-MLP First Guess ERO and operational ERO. Figure 45B shows the Day 3 forecasts. AuROC relates the cumulative hit rate to the corresponding false alarm rate, with higher values being better. According to this measure, the CSU-MLP First Guess ERO performed better than the operational WPC ERO at both Day 2 and Day 3. Objectively, both the Day 2 and Day 3 CSU-MLP First Guess EROs had much lower fractional coverage for the marginal, slight, and moderate categories and were

either on the low-end or below calibration for each probabilistic category. Despite the low fractional coverage, the AuROC measure indicates the CSU-MLP First Guess EROs were better than the operational EROs. This may be caused by the corresponding false alarm rate being lower for the CSU-MLP forecasts compared to the operational EROs. It is important to note that Brier Skill Score and AuROC are different skill metrics and thus a single conclusion cannot be formed by either skill score alone.

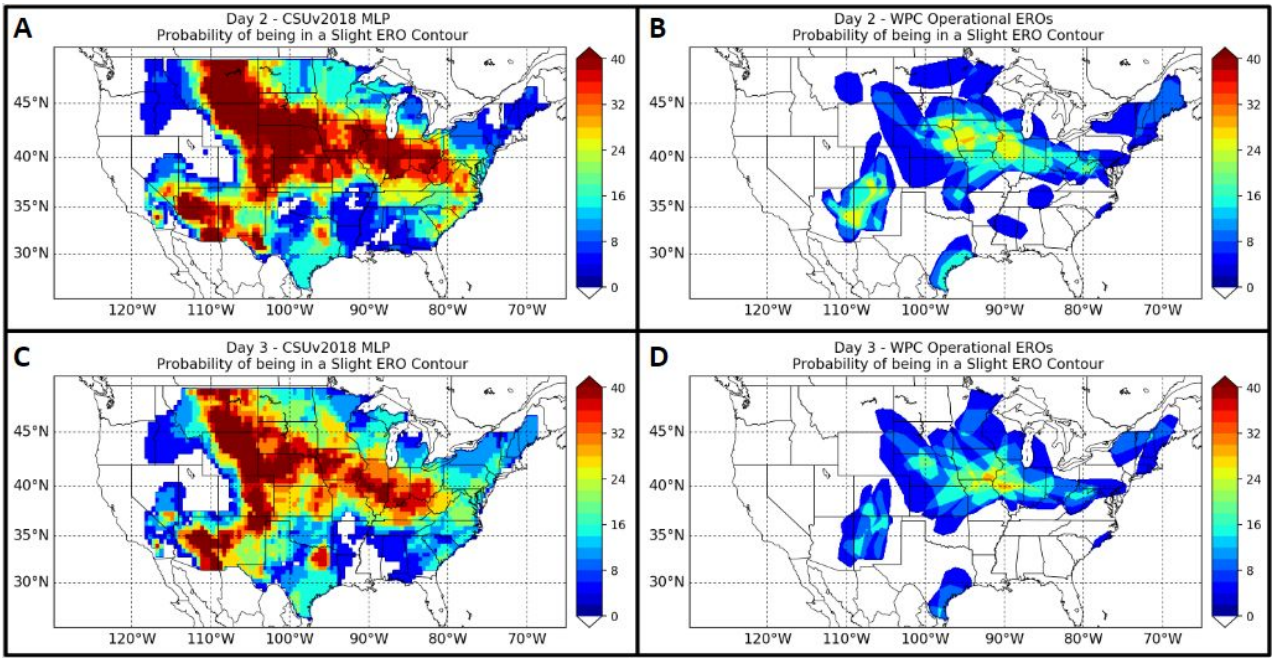


Figure 42. *(A) Probability of being in a "slight risk" Day 2 CSU-MLP First Guess ERO contour, (B) "slight risk" Day 2 operational ERO contour, (C) probability of being in a "slight risk" Day 3 CSU-MLP First Guess ERO contour, and (D) "slight risk" Day 3 operational ERO contour over the four week experiment.*

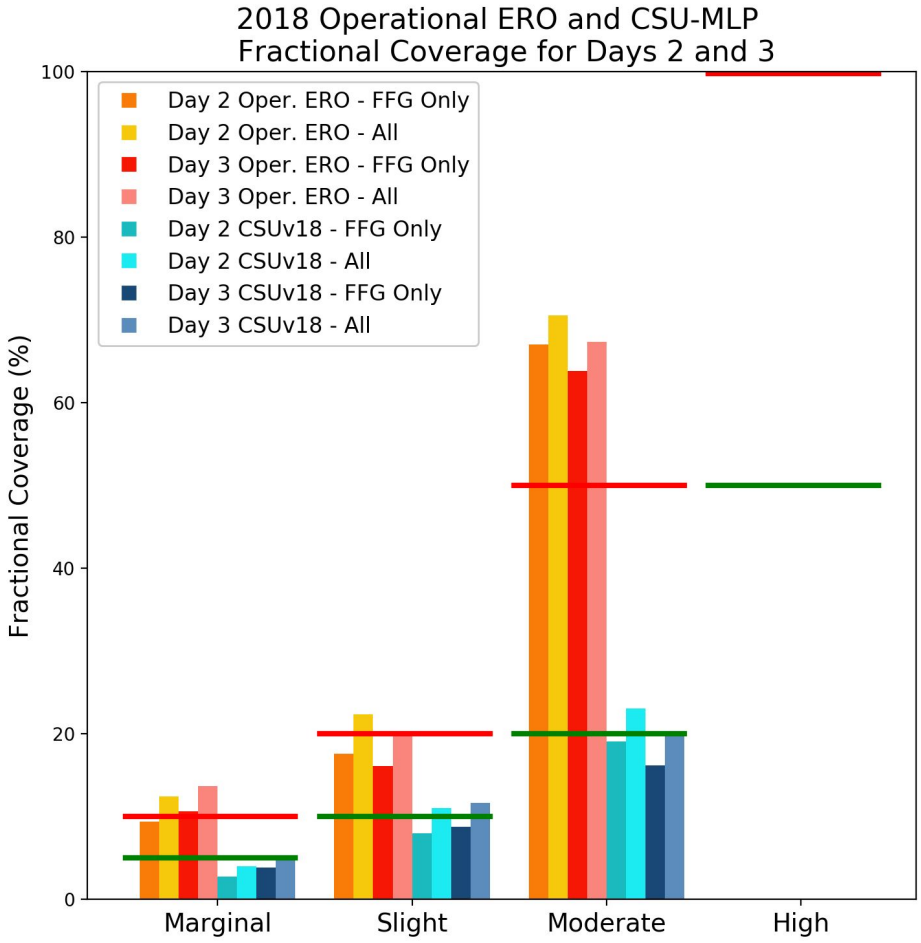


Figure 43. *Fractional coverage of the 2018 Day 2 (teal/cyan) and Day 3 (dark blue/light blue) CSU-MLP First Guess ERO and the Day 2 (orange/yellow) and Day 3 (red/pink) operational EROs issued over the same time period for each probabilistic category. Green horizontal lines represent the lower defined bound for each threshold and red horizontal lines represent the highest defined bound.*

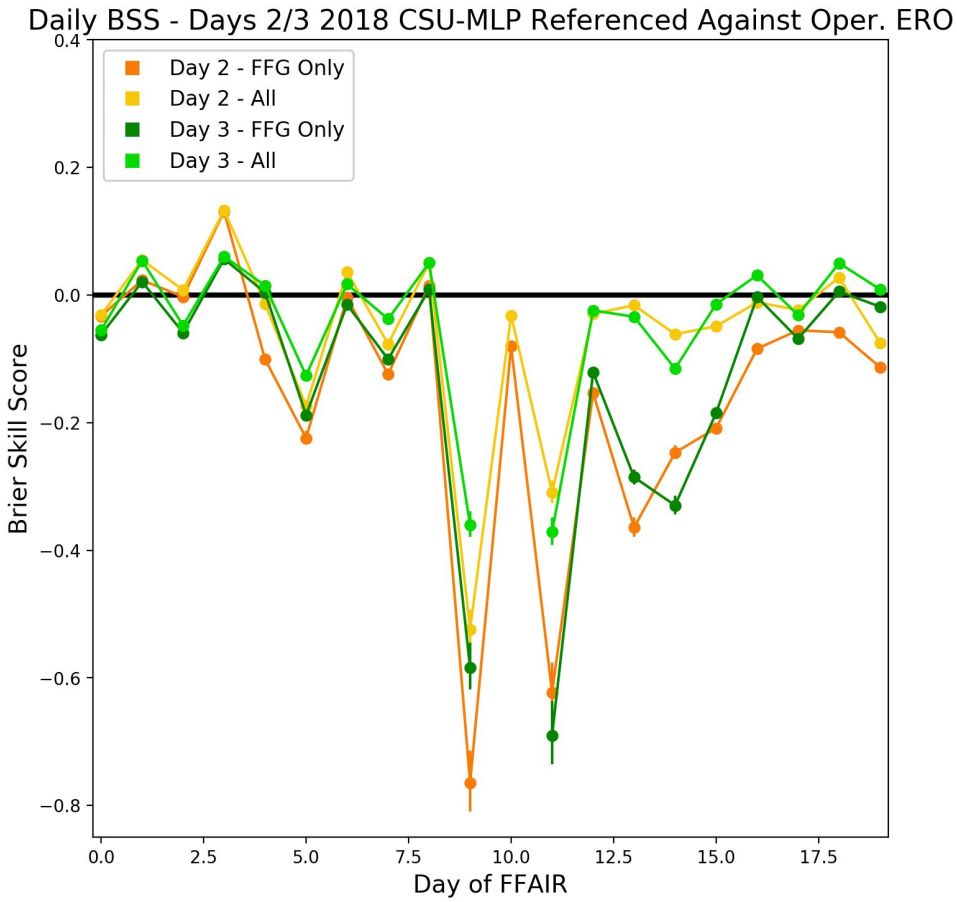


Figure 44. *Daily BSS for the Day 2 (orange/yellow) and Day 3 (green/light green) CSU-MLP First Guess EROs referenced against the Day 2 and 3 operational EROs throughout the entire experiment. Positive values represent days where the CSU-MLP First Guess ERO had better skill than the operational ERO, negative values represent worse skill.*

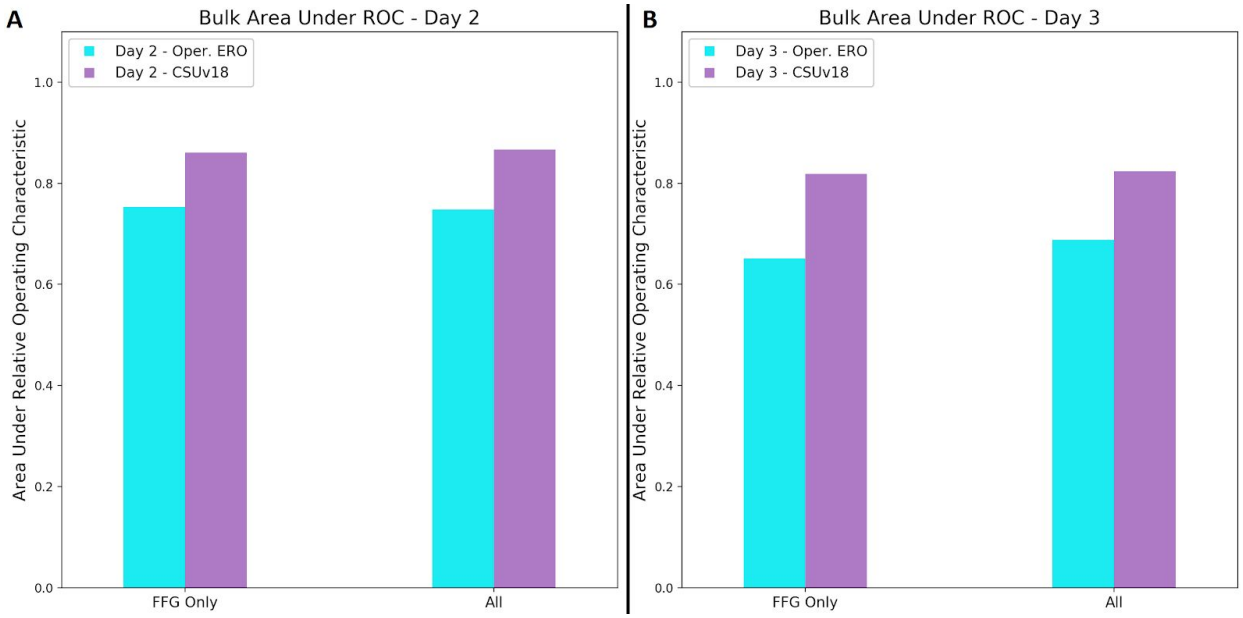


Figure 45. *(A) Bulk area under the relative operating characteristic for the Day 2 CSU-MLP First Guess EROs (blue) and Day 2 Operational EROs (purple) using both FFG exceedance only and the UFV system for verification. (B) Same as previous but for the Day 3 time period.*

Recommendations

Based on positive feedback and the need for a Day 2/3 First Guess Field, WPC-HMT recommends both products be transitioned to operations. Adjustments to the ARI sensitivities are suggested to increase the probabilistic confidence along the Gulf Coast and Southeast United States and decrease the confidence in the upper plains, specifically Montana and the Dakotas. Further development on the Day 1 version that uses NSSL-WRF is recommended to try and establish more consistency between all three days, if possible, considering the different QPF used for Day 1 and Days 2/3.

6. Satellite Guidance Results

CIRA ALPW Difference Field

The CIRA Advected Layered Precipitable Water (ALPW) model difference product was evaluated by collecting comments throughout the experiment. The participants were shown a panel of the CIRA HRRR-driven ALPW difference fields, shown in Figure 46, at four layers in the atmosphere (surface to 850 hPa, 850-700 hPa, 700-500 hPa, and 500-300 hPa) and were asked to assess the dry and moist regions, compare it with the MRMS-GC and HRRR QPF valid over the same time period, and asked to comment on whether or not the data could be useful to the forecast process. Participants focused on the two middle to upper layers at 700-500 hPa and 500-300 hPa.

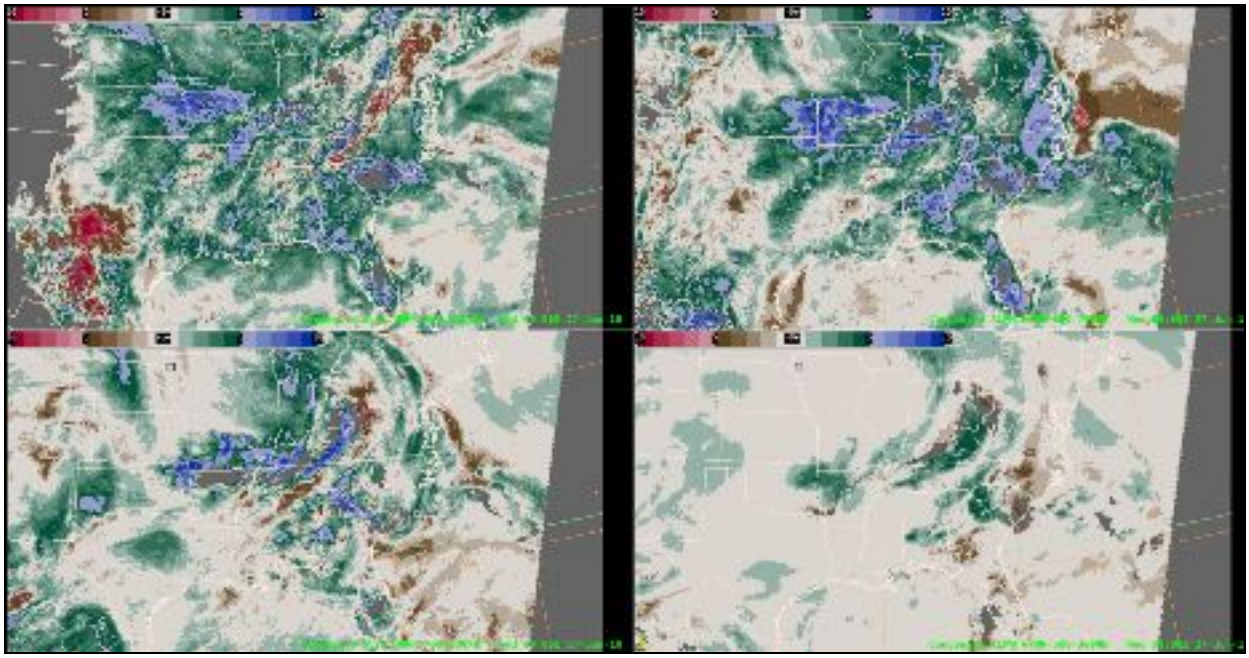


Figure 46. *An example of the CIRA HRRR-driven ALPW difference fields presented to participants during the 2018 FFaIR subjective evaluation, valid 09-12 UTC June 27, 2018. Surface to 850 hPa (top left), 850-700 hPa (top right), 700-500 hPa (bottom left), 500-300 hPa (bottom right).*

Findings

Participants found the highest layers, specifically 700-500 hPa and 500-300 hPa, most useful as they were less noisy, and some questioned the satellite retrieval accuracy at the surface. After some adjustments to understanding the information the product provides, forecasters began to find applications, such as identifying the location of frontal zones and moisture feeder band connections that contribute to convective outbreaks. The difference between model and observed moisture enabled participants to evaluate the model in terms of timing of synoptic features (ie, advancing cold fronts), locations of mesoscale convective systems (MCS), and regions over which the model may be too moist thus producing too much rainfall.

Specific to flash flooding, the ALPW difference field was useful in identifying wet/dry biases in the model QPF that may impact rainfall magnitude, timing, and location of features. In the example shown in Figure 47, the HRRR model is advancing the MCS more quickly and with more moisture than is observed by the QPE and satellite retrievals. Identifying these differences can then be applied when assessing the model, in this case the HRRR, for flash flood forecasting.

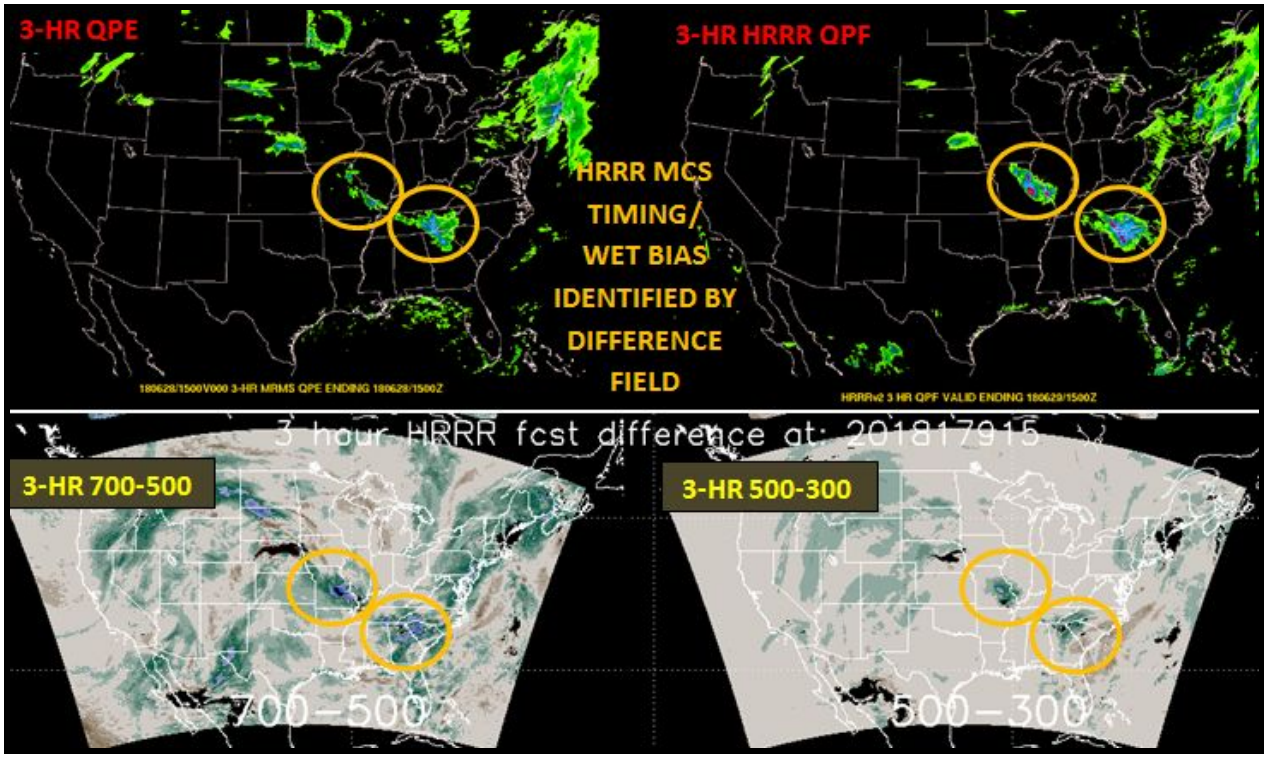


Figure 47. *3 HR MRMS-GC QPE (top left) shown with HRRR QPF (top right) and the ALPW-HRRR difference fields from the 700-500* hPa *(bottom left) and 500-300* hPa *(bottom right) layers, all valid 12-15 UTC June 28, 2018. The orange circles highlight areas in which the HRRR model had more moisture than the satellite retrieval (bottom row) and had higher QPF (top right) than the verification MRMS-GC QPE (top left).*

Recommendations

Extensive forecaster training is recommended to deepen understanding and maximize the utility of the ALPW difference fields. Training should include any known strengths, weaknesses, and known biases (trends) of the product, such as a wet bias in the model over regions of current rainfall. Training should also include the timing of the combined passes as differenced with the model runs to increase forecaster confidence in the product.

Forecasters referred to convection as "contaminating" the data in terms of residual cold pools and a resulting wetter model in areas where it was actually precipitating. Therefore, the desire to somehow mask out convection was mentioned. Aesthetically, participants commented that the brown (drier) and grey (no data) fields were too close and the colors should contrast more.

7. FFaIR Experimental Forecast Activities Results

Experimental Day 1 Excessive Rainfall Outlook

Each day participants evaluated and subjectively scored the 21-hour experimental Day 1 ERO. For verification, the UFV system was used. Please refer to the "Verification" section in section 3

for additional details. Figure 48 shows an example of a FFaIR Day 1 ERO overlaid with UFV reports. Participants assigned scores from 1 (very poor) to 10 (very good).

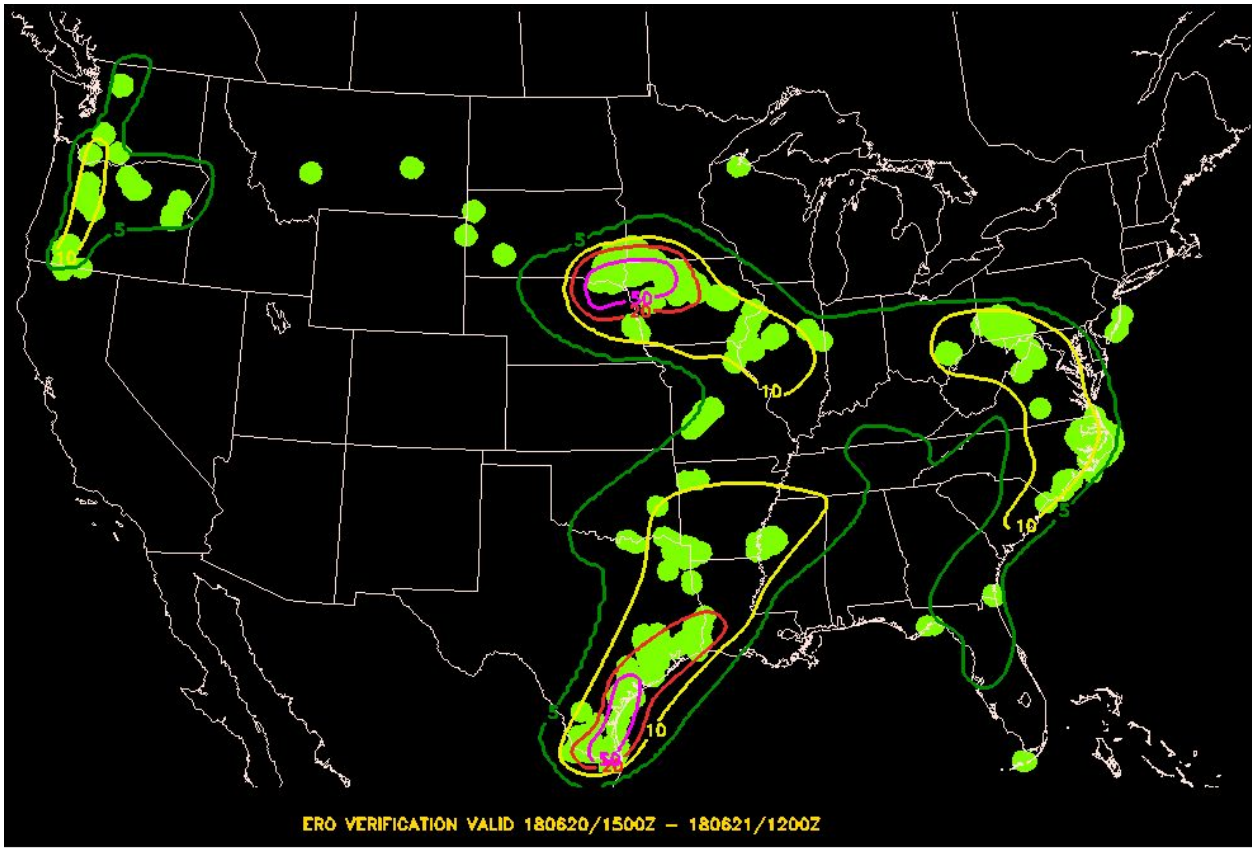


Figure 48. *Experimental Day 1 ERO with probabilities of 5% (marginal/green), 10% (yellow/slight), 20% (moderate/red), and 50% (high/magenta) chance of flooding rains occurring from 15 UTC June 20 - 12 UTC June 21, 2018. The UFV reports are shown by the green circles.*

Figure 49 shows the box plot of the subjective results over the entire experiment for the experimental Day 1 ERO. The average subjective score was 7.28 out of 10 with a standard deviation of 1.00.

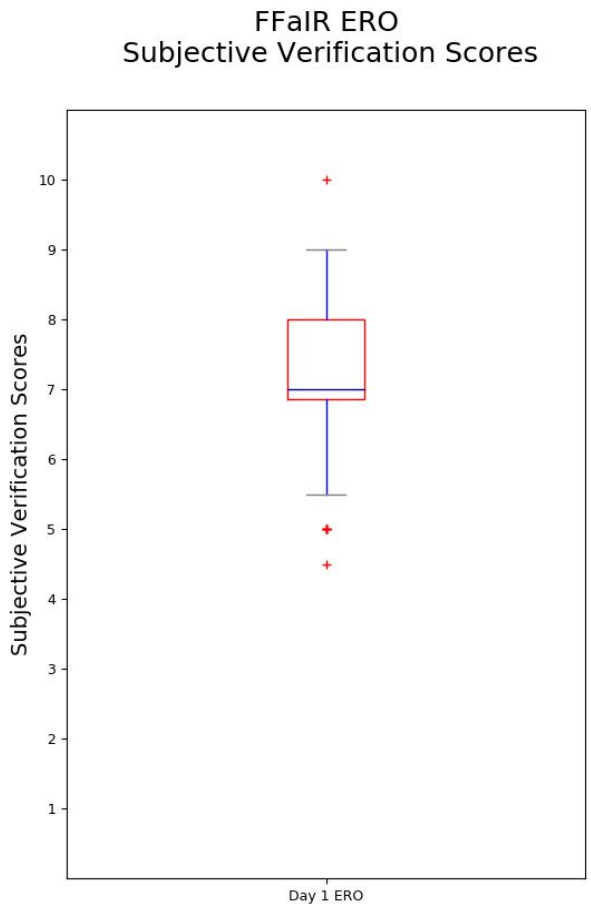


Figure 49. *A box plot of the overall subjective scores for the 21 hour experimental Day 1 ERO over the course of the entire experiment. Red plus symbols denote outliers.*

With an average subjective score of over 7 out of 10, participant feedback on the FFaIR experimental Day 1 EROs were generally positive each day. The marginal contour almost always captured the extent of the reports. However, there were days that reports fell outside of the marginal contour; this most commonly occurred in the Southeast U.S. and the northern Central Plains. There were also times that small scale, high impact flash flooding events occurred within the marginal contour and participants felt that a higher probabilistic threat area would have been warranted. The day with the highest average score (8.83) shown in Figure 50A featured two high risk regions that were well placed and a slight risk in the Mid-Atlantic that captured a small-scale, high impact event in Pittsburgh, PA. The day with the lowest average score (6.08) in the experiment is shown in Figure 50B. On this day, there was a moderate risk across southern Minnesota and into Wisconsin that had no reports directly within the contoured region. Reports fell to the north and east of the region and further southwest where the probabilities were lower. The slight contour in the Southwest was broad and some participants felt a moderate could have been justified in the northwest portion.

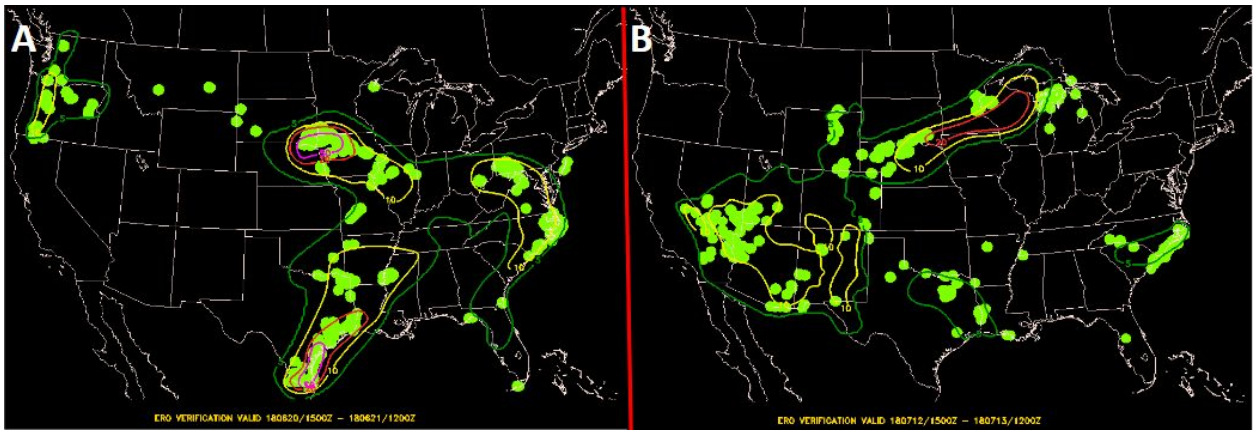


Figure 50. *(A) FFaIR Day 1 experimental ERO (contoured) valid 15 UTC June 20 - 12 UTC June 21, 2018 and (B) FFaIR Day 1 experimental ERO (contoured) valid 15 UTC July 12 - 12 UTC July 13, 2018. The UFV reports are shown by the green circles.*

Comparison of the Day 1 Experimental FFaIR ERO and Day 1 WPC Operational ERO

The Day 1 FFaIR EROs issued over the four weeks of the experiment were compared to the operational WPC EROs issued over the same time period. The 15 UTC issuance of the Day 1 operational ERO was used for comparison. For verification, the EROs were verified against both the UFV system and the exceedance of flash flood guidance only, which is how the operational EROs are verified at WPC. The underlying probabilities of both the experimental FFaIR EROs and the operational WPC EROs were defined at the same 5% (marginal), 10% (slight), 20% (moderate), and 50% (high) thresholds. Figures 51-54 show the probability of being in a "marginal risk" area, a "slight risk" area, a "moderate risk" area, and a "high risk" area from the operational and experimental EROs, respectively, over the four week experiment. From the figures it is immediately apparent that there was lower overall coverage from the operational EROs at all probability thresholds. Zero Day 1 15 UTC high risk operational EROs were issued over the course of the experiment. The area covered by the FFaIR ERO marginal, slight, and moderate contours were 127%, 141%, 451% greater than the operational ERO forecast, respectively. These maps also illustrate the major focus areas during the experiment with higher probabilities in the Southwest associated with the monsoon moisture as well as higher probabilities throughout the central U.S. and into the Southeast.

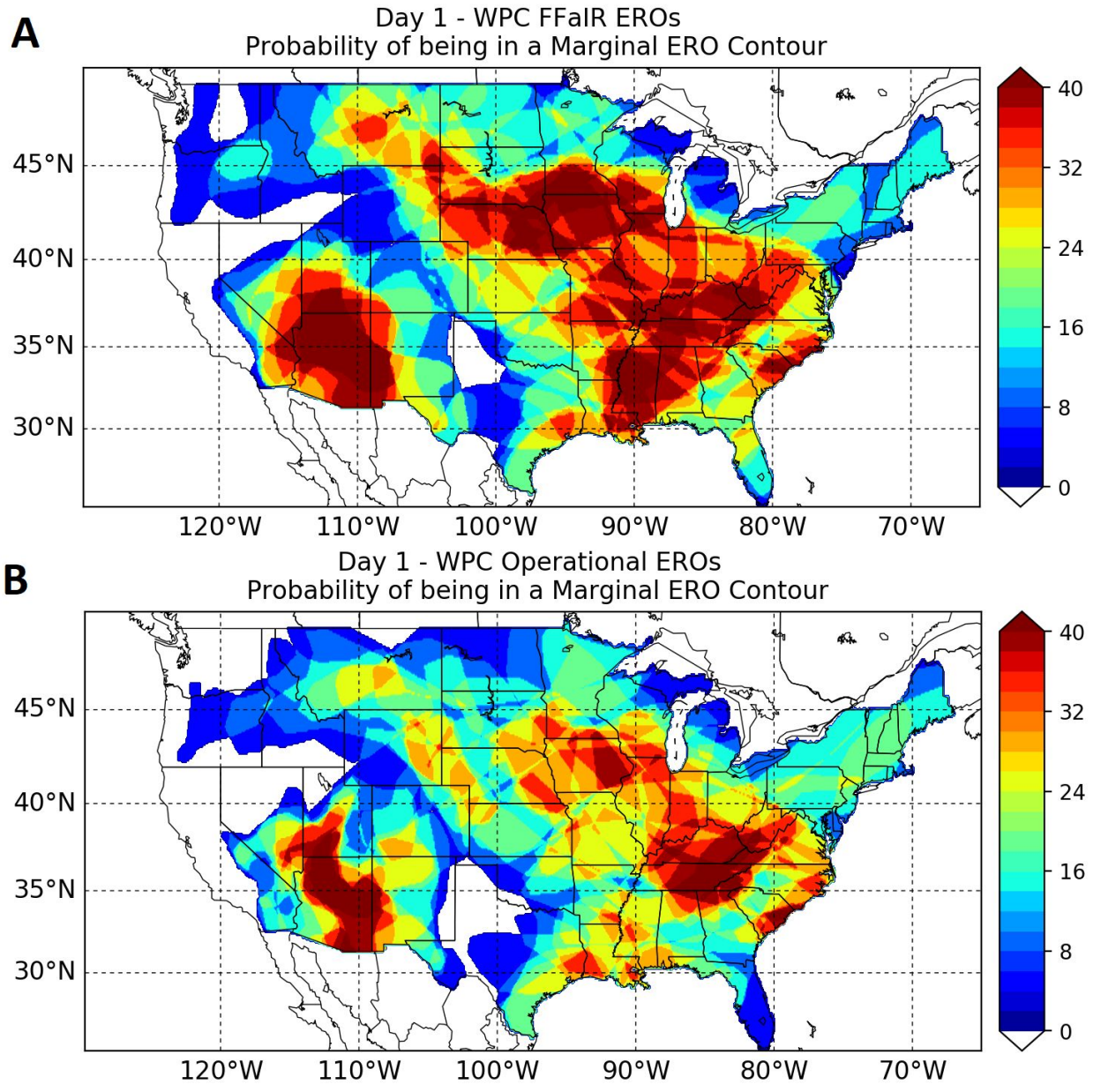


Figure 51. *(A) Probability of being in a "marginal risk" experimental FFaIR ERO contour and (B) "marginal risk" operational ERO contour over the four week experiment.*

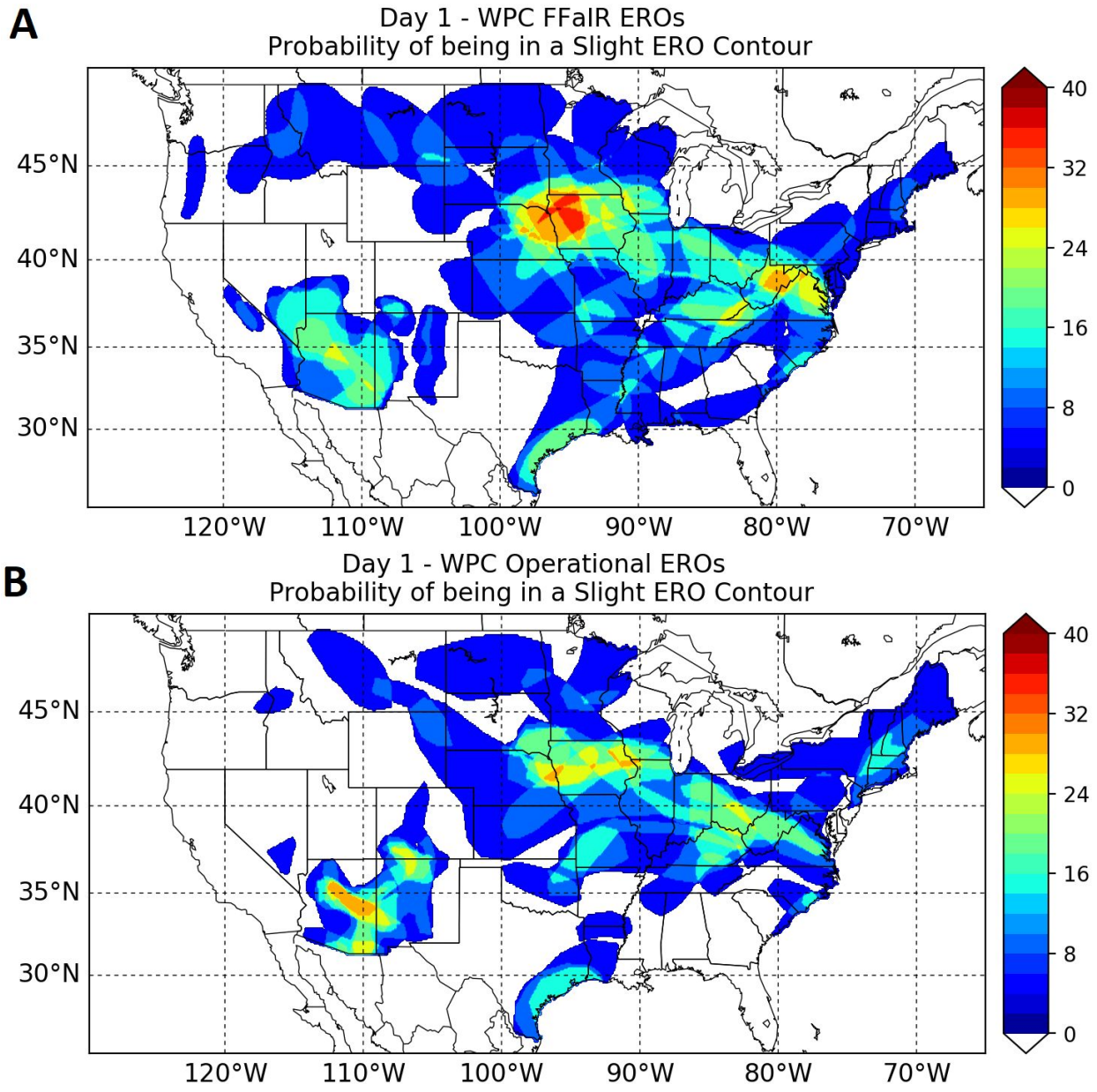


Figure 52. *(A) Probability of being in a "slight risk" experimental FFaIR ERO contour and (B) "slight risk" operational ERO contour over the four week experiment.*

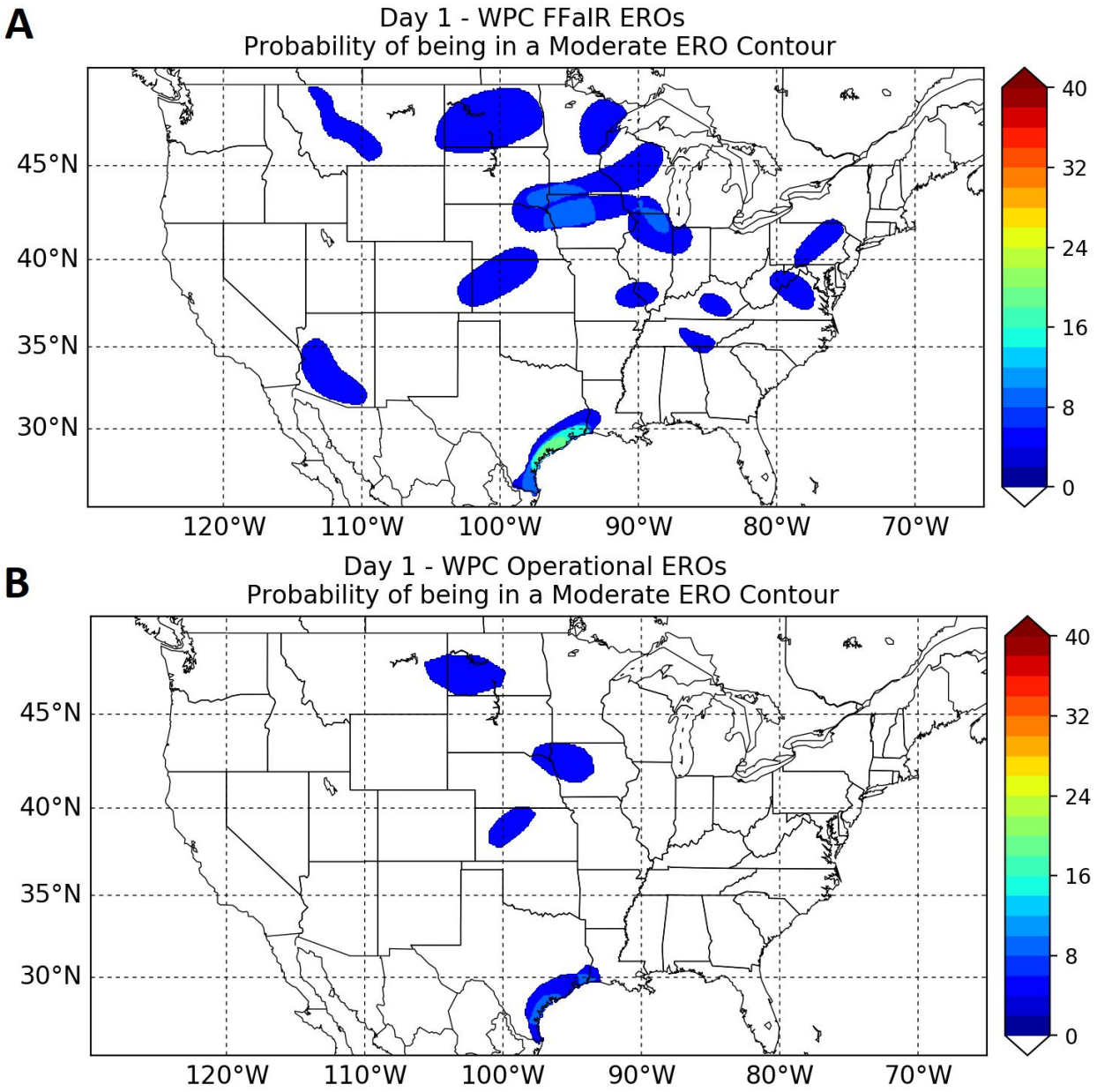


Figure 53. *(A) Probability of being in a "moderate risk" experimental FFaIR ERO contour and (B) "moderate risk" operational ERO contour over the four week experiment.*

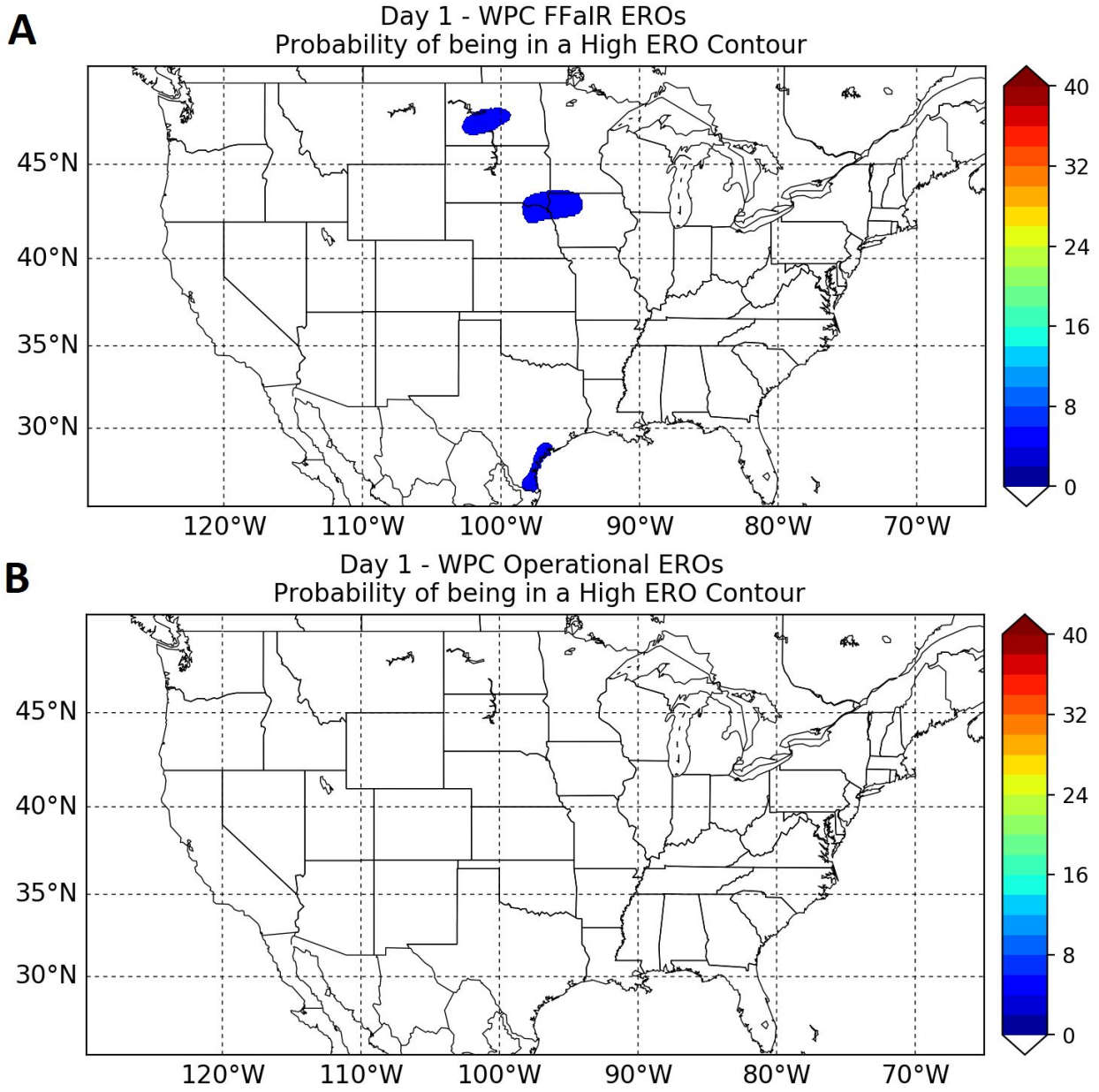


Figure 54. *(A) Probability of being in a "high risk" experimental FFaIR ERO contour and (B) "high risk" operational ERO contour over the four week experiment.*

Figure 55 shows the bulk fractional coverage by threshold of the operational EROs and FFaIR EROs. The green and red horizontal lines represent the lower and higher bound of each probabilistic category. The operational ERO fractional coverage fell above the upper bounds of the slight and moderate categories when using all verification sources (ERO - All/yellow); it was also above the marginal category upper bound. The FFaIR EROs all fell within the categorical definitions except when using all verification sources (FFaIR - All/light purple) in the slight category. The larger areal coverage of the FFaIR EROs led to their average fractional coverage being lower when compared to the operational EROs. Figure 56 shows the daily Brier Skill Score (BSS) referenced against the operational Day 1 ERO throughout the experiment. In the

figure, anywhere with positive values represents instances where the FFaIR ERO performed better than the operational ERO. Verification was done using both the UFV system (all) and FFG exceedance only. Although variable day-to-day, the FFaIR ERO daily BSS was positive for all but one day during the first two weeks of the experiment. During the last two weeks of the experiment, there was an interesting divergence between the two methods of verification. Looking at the FFG exceedance only verification (green), the FFaIR ERO performs worse than the operational ERO most days in the second half of the experiment. However, when using the UFV system and including all verification sources (blue), the Day 1 FFaIR EROs and operational EROs were more similar, with a mix of both positive and negative BSSs.

Lastly, Figure 57 displays the AuROC score for both the FFaIR and operational Day 1 ERO. According to this measure, the FFaIR EROs were slightly more skillful than the operational EROs during the experiment period. The results are similar for both verification methods. Overall, the objective results were mixed. The larger areal coverage of the Day 1 FFaIR EROs led to better calibrated forecasts in terms of fractional coverage compared to the operational ERO. The BSS showed the Day 1 FFaIR ERO performed better than the operational ERO during the first half of the experiment, but results were more mixed in the second half. The AuROC results suggest the experimental FFaIR ERO to be just slightly better than the operational ERO over the course of the experiment.

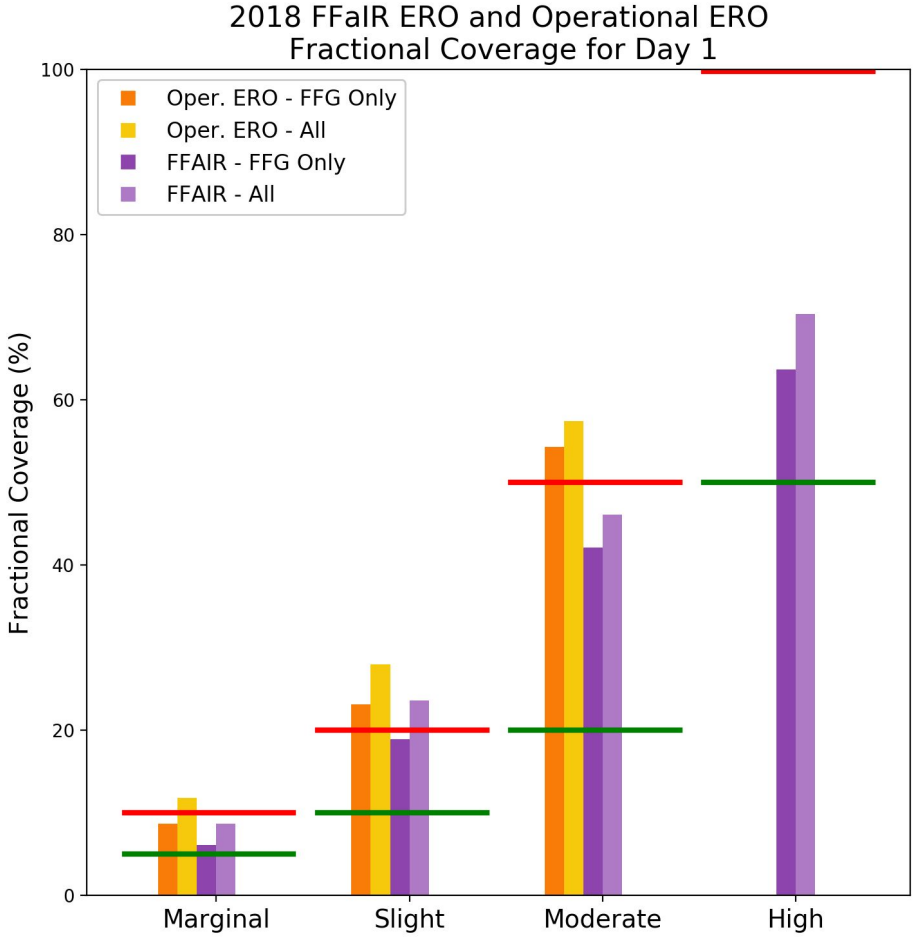


Figure 55. *Fractional coverage of the 2018 Day 1 FFaIR EROs (purple) and operational EROs (orange) issued over the same time period for each probabilistic category. Green horizontal lines represent the lower defined bound for each threshold and red horizontal lines represent the highest defined bound.*

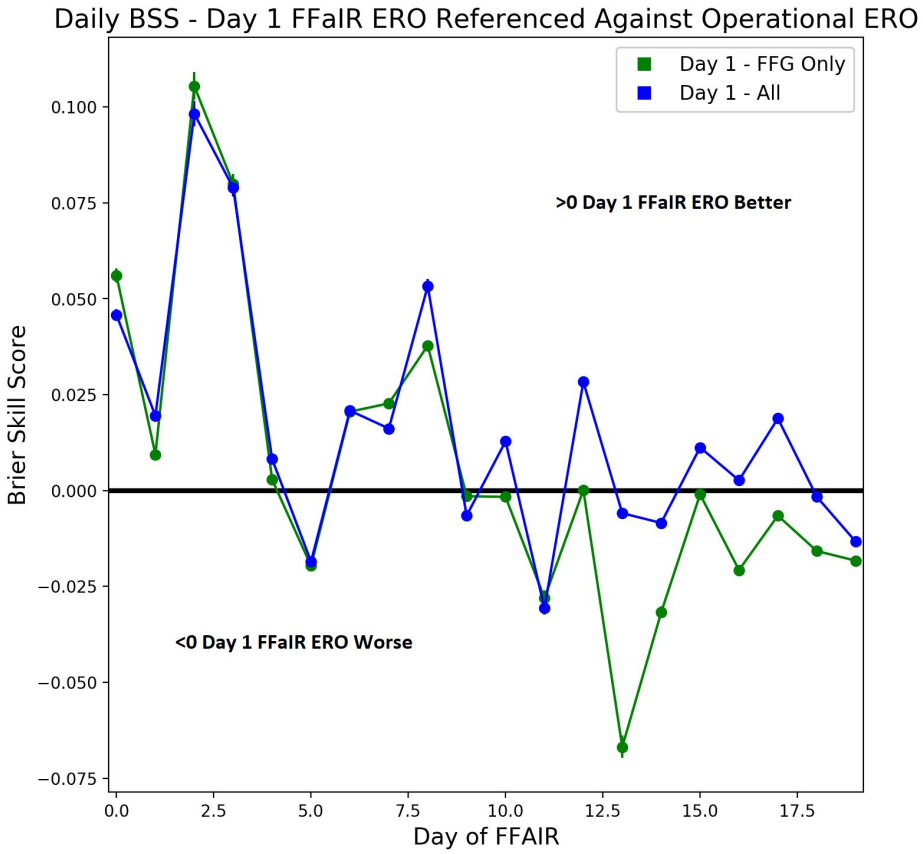


Figure 56*. Daily BSS referenced against operational EROs throughout the entire experiment. Positive values represent days where the FFaIR ERO had better skill than the operational ERO, negative values represent worse skill.*

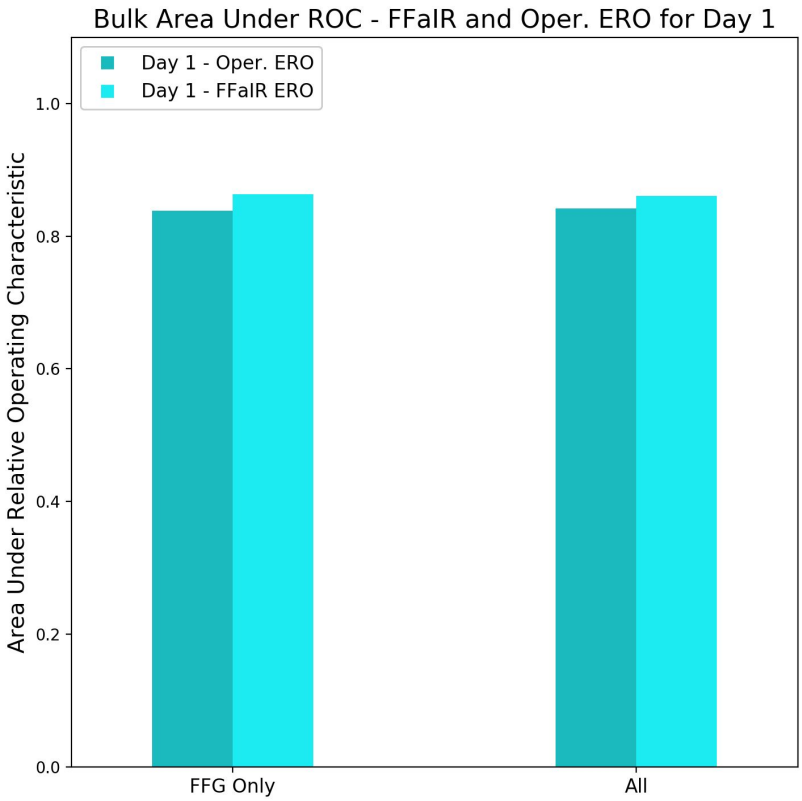


Figure 57. *Bulk area under the relative operating characteristic for the Day 1 operational (dark blue) and FFaIR EROs (light blue) using both FFG exceedance only and the UFV system for verification.*

Experimental Probability of Flash Flood Forecast 1 and 2 Results

Participants subjectively evaluated the PFF1, valid 18-00 UTC, and the PFF2, valid 00-06 UTC, over limited domains each day during the experiment. The UFV system was used to verify both PFFs. The average subjective scores for both the PFF1 and PFF2 were almost equal, with the PFF1 having an average score of 6.30 (standard deviation = 2.23) and the PFF2 had an average score of 6.31 (standard deviation = 1.89). Figure 58 shows the box plot of all the subjective results for both the PFF1 and PFF2. Figure 59A is an example of one of the highest subjectively rated PFF1 forecasts during the experiment. The locations of the contours were well placed and the inclusion and locations of the moderate risk areas were positively rated. On more marginal days, forecasters often debated whether to issue a forecast at all because the minimum probability contour available to forecast started with the 10%/slight category. It was on these days where the participants struggled the most with the placement of the contour(s) and drawing the contour(s) either too broad or too narrow due to the lack of definitive guidance during the more marginal days. Figure 59B is an example of a marginal day showing one of the lowest scored PFF1 forecasts during the experiment. It featured a broad slight risk that did not capture any of the few reports during the 18-00 UTC time period. Despite the longer lead time for the PFF2, no significant differences stood out among either the subjective evaluation scores or the participant feedback when verifying the forecasts.

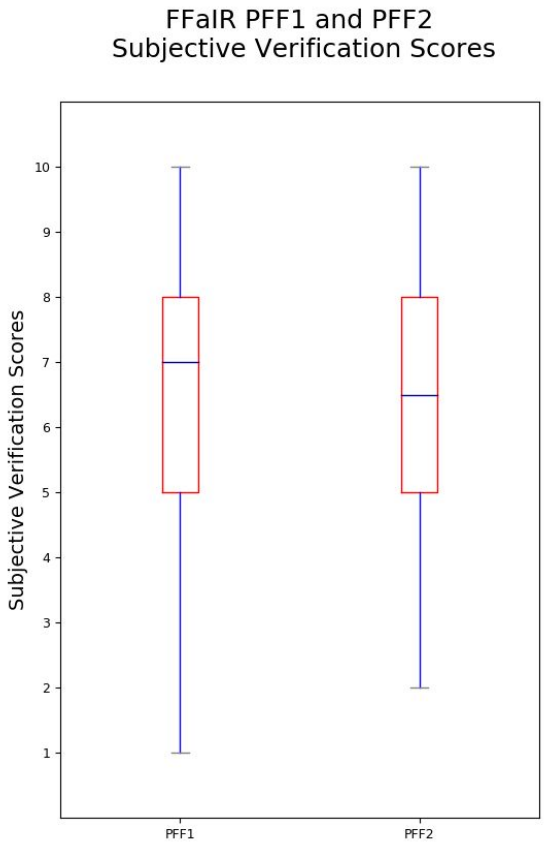


Figure 58. *A box plot of the overall subjective scores for the 6 hour experimental PFF1 (18 - 00 UTC) and PFF2 (00 - 06 UTC) over the course of the entire experiment. Red plus symbols denote outliers.*

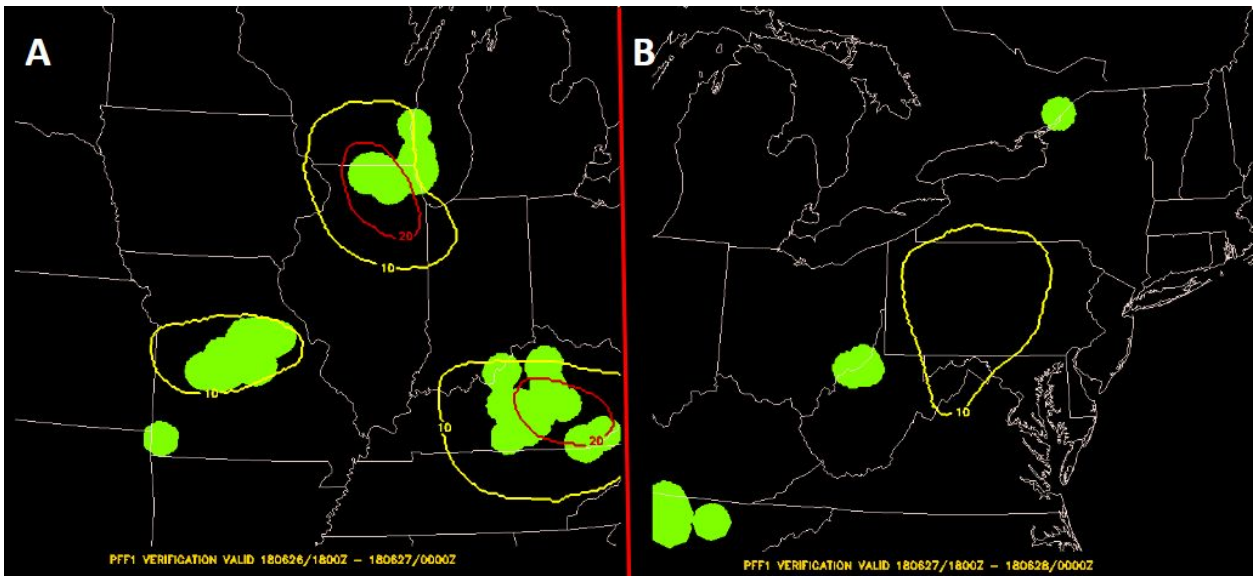


Figure 59. *(A) PFF1 valid 18 UTC June 26 - 00 UTC June 27, 2018 with 10%/slight (yellow) and 20%/moderate (red) probability contours (B) PFF1 valid 18 UTC June 27 - 00 UTC June 28, 2018 with 10%/slight (yellow). The UFV reports are shown by the green circles.*

Figure 60 show the probability of being in a "slight risk" (A) and "moderate risk" (B) experimental FFaIR PFF1 contour valid 18-00 UTC over the four week experiment. Figure 61 shows the same except for the PFF2 valid 00-06 UTC. Each give a sense of where participants were focused for the two shorter term forecasts. The Ohio River Valley, northern Central Plains, and Southwest were most commonly the focus of the two PFF forecasts. Finally, Figure 62 shows the fractional coverage of both the PFF1 and PFF2 forecasts using FFG only and the UFV system (All) as verification. When using all the reports, FFG, and ARI data in the UFV system, the PFF fractional coverage was much higher in each category when compared to using FFG only for verification. The slight and moderate categories for the PFF1 (All) were calibrated according to the probabilistic definitions, but the high category fell under the probabilistic definition. For both the PFF1 and PFF2 forecasts, only two high risk forecasts were issued throughout the experiment leading to a low sample size. All categories for the PFF1 and PFF2 using FFG only for verification, except for the PFF1 moderate, fell below the minimum threshold for each probabilistic category. The PFF2, using all verification sources, was calibrated for each slight, moderate, and high category, but on the lower end of the probabilistic range.

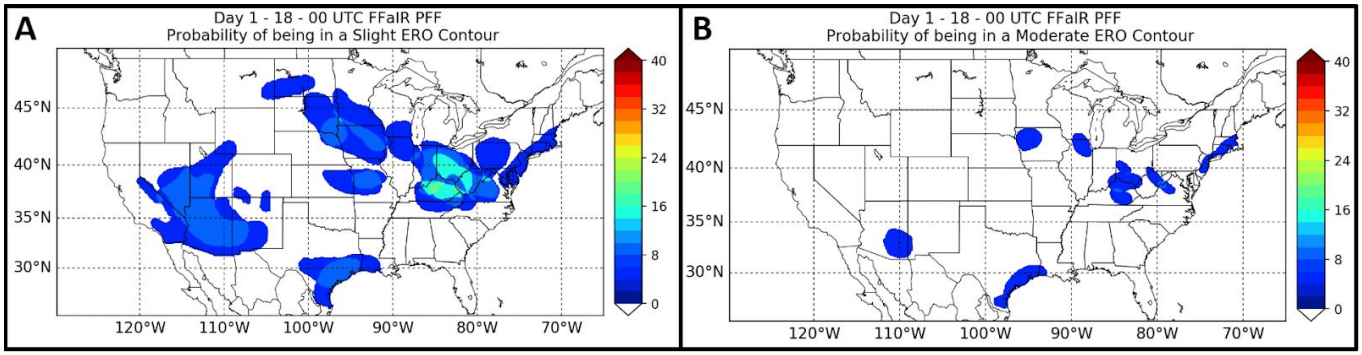


Figure 60. *(A) Probability of being in a "slight risk" and (B) "moderate risk" experimental FFaIR PFF1 (18-00 UTC) contour over the four week experiment.*

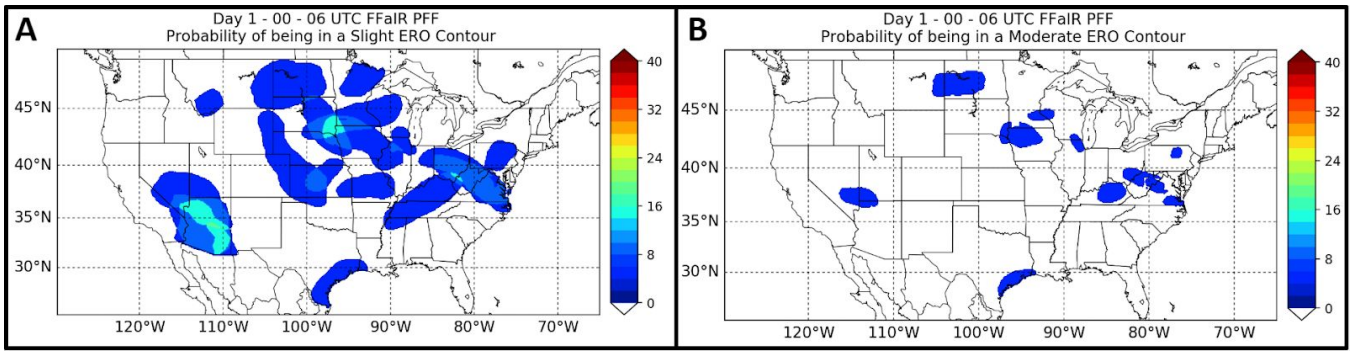


Figure 61. *(A) Probability of being in a "slight risk" and (B) "moderate risk" experimental FFaIR PFF2 (00-06 UTC) contour over the four week experiment.*

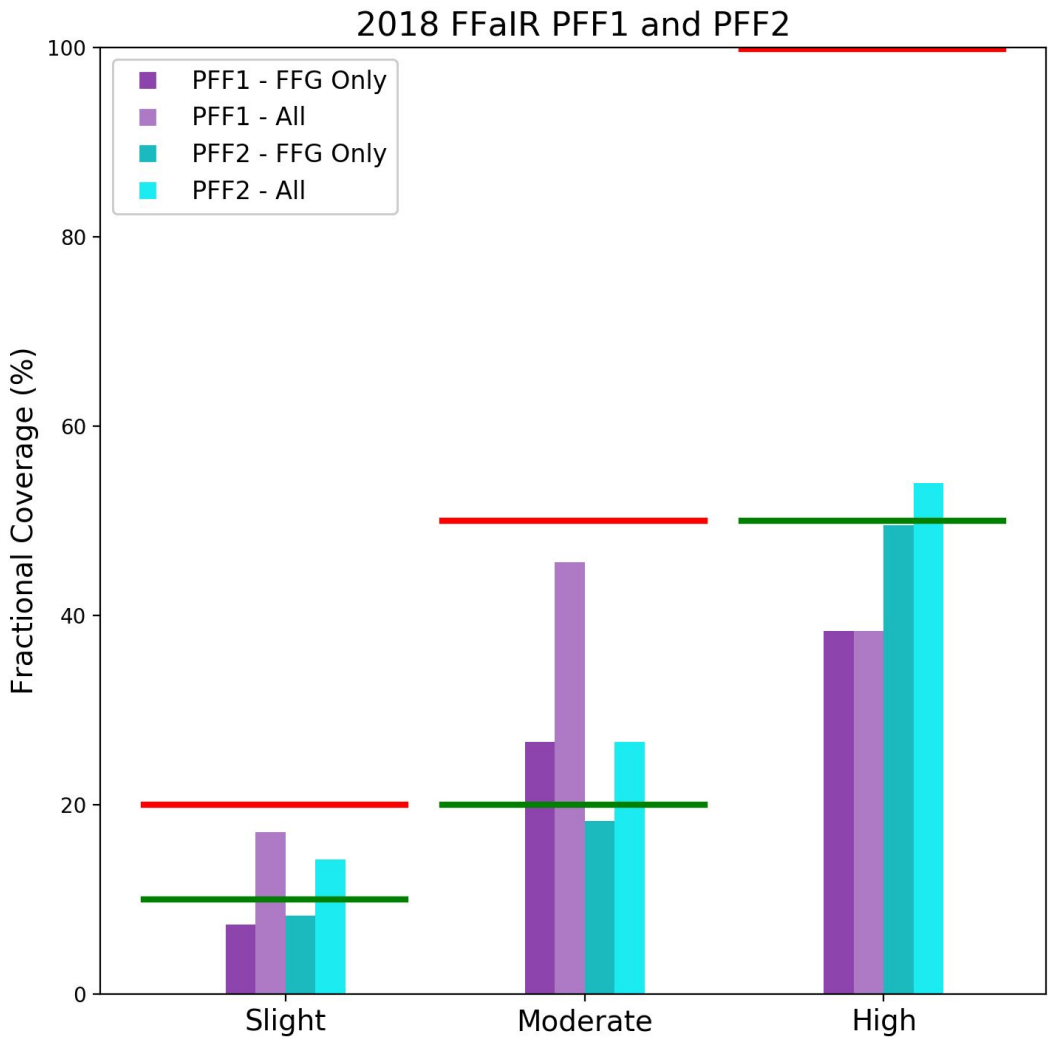


Figure 62. *Fractional coverage of the 2018 FFaIR PFF1 (purple) and PFF2 (blue) issued over the four weeks of the experiment. Green horizontal lines represent the lower defined bound for each threshold and red horizontal lines represent the highest defined bound.*

8. Summary and Research-to-Operations Recommendation

The 6th annual Flash Flood and Intense Rainfall Experiment was conducted within the Hydrometeorology Testbed at WPC (HMT-WPC) from June 18 - July 20, 2018 bringing together NWS meteorologists, hydrologists, and the development and research communities for the advancement of research into WPC and NWS field operations. The FFaIR Experiment focused on synthesizing the use of high resolution, atmospheric guidance, hydrologic guidance, and remotely-sensed tools to improve flash flood prediction in the short range (6-24 hours). Subjective and objective data was successfully collected and analyzed. The conclusions drawn are as follows:

- The WPC-HMT recommends the Day 2 and Day 3 ERO **CSU-MLP First Guess Field** for operations as it showed great potential and was scored well by participants. It is

recommended that the CSU developers work to refine some of the high probabilities in the High Plains and low probabilities in the Southeast and continue to develop and test the Day 1 version.

- The **local probability matched mean** QPF from the SSEFX had the highest average subjective score from participants and was successful in most cases of having a bias closer to 1.00 than the PMM from the SSEFX. HMT-WPC recommends the LPM be tested in additional ensembles in the future based on the improved bias and positive subjective feedback over two years of testing. The LPM method showed promise in reducing the high QPF bias seen in the traditional PMM calculation in all of the ensembles tested in FFaIR.
- Forecasters looked at the **ALPW-HRRR Difference Field** for the first time in the 2018 FFaIR and had positive impressions. Participants noted that when the HRRR was more moist than the satellite retrieval in the 700-500 hPa and 500-300 hPa layers, it often produced higher QPF than was observed. Large frontal zones and dry lines were also able to be detected at times. More extensive training is recommended to deepen overall understanding and maximize the utility of the ALPW difference fields. Training should include any known strengths, weaknesses, biases (trends) of the product, such as a wet bias in the model over regions of current rainfall.
- The experimental **High Flow Potential** product from the National Water Model (NWM) showed promise as a situational awareness tool for forecasters. Participants liked the definition change from climatological-based flow to 1.5 year recurrence flow this year as well as the ability to overlay the QPF from the model. There still was difficulty in understanding how the impacts of high flow potential relate to flash flooding threats. Recommendations include the ability to view hydrographs when clicking, utilizing the new HRRR 36 hour runs to extend the forecast, and a way to view how the forecast has changed from the previous high flow potential forecast.
- The experimental **High Flow Probability** product from the NWM is a great step into probabilistic hydrologic guidance. This product had some utility for the forecasters by probabilistically highlighting regions of hydrologic concern, especially in smaller reaches. There were still problems, however, relating the data to flash flooding impacts. The small 6-8 hour time window when this product was valid was also limiting.
- Experimental **Peak Flow Arrival Time** provided utility in helping forecasters with the timing of high flow situations, especially at smaller scales. Participants felt the dependency on HRRR QPF in this product was especially noticeable and therefore, if they felt the QPF was poor, this product was not looked at as closely. Participants recommended an option to view hydrographs for this product and a way to filter out large mainstem rivers so smaller reaches that may be more susceptible to flooding would be more noticeable.
- The **NEWS-e** was successfully tested during FFaIR. The first four cycles (18/19/20/21 UTC) were evaluated and each newer cycle had an improved CSI value at 0.10 in., 0.25 in., 0.50 in., and 1.0 in. Participants found the NEWS-e struggled to generate QPF associated with lighter, weakly forced precipitation but also produced too heavy QPF in convective cores. Contingent on a full CONUS domain in the HRRRE, WPC-HMT recommends the NEWS-e be tested in the Southwest region during the monsoon as well as evaluate later cycles in future tests.

- Testing **Day 1 deterministic models and NBMv3.1 QPF** in the short term (12-36 hr forecasts) was a primary goal of the 2018 FFaIR Experiment, and results were mixed. The **NBMv3.1** and **FV3-GFS** both did well with areal coverage and had high CSI values at 0.5 in., but very low bias at higher thresholds. The **HRRRv3** had the highest subjective and objective scores of the high-resolution CAMs. It tended to have a low bias in the Southwest U.S. and be too heavy at times in strongly forced patterns. Both **high-resolution FV3 models** had the lowest CSI values at 0.5 in. and low 1.0 in. scores as well. WPC-HMT recommends that work continue on the FV3-Thompson and FV3-NSSL to investigate whether the microphysics schemes were causing the large QPF differences sometimes seen. WPC-HMT also recommends further study to improve HRRRv3 QPF in the Southwest U.S. associated with monsoon moisture.
- The area covered by the Day 1 **FFaIR ERO marginal, slight, and moderate probability contours was larger than the Day 1 Operational ERO**. FFaIR ERO fractional coverage, especially in the slight and moderate categories, was better than the operational ERO, which tended to fall above the upper bound for those categories.
- The Day 1 FFaIR ERO was slightly better than the operational Day 1 ERO based on AuROC results and **better during the first half but similar or worse in the second half** of the experiment than the operational Day 1 ERO based on BSS.

Acknowledgements

The 2018 FFaIR Experiment was only possible through the hard work and dedication of the HMT and WPC staff. Box plots, performance diagrams, objective MODE verification, and comparisons of experimental ERO and operational ERO were generated by **Michael Erickson** (CIRES/WPC). Special thanks to WPC Science Operations Officer, **Mark Klein**, for technical and management support. Special thanks to Branch Chief, **James Nelson**. And many thanks to the operational WPC forecasters that lent their time and expertise to the experiment.

References

Benjamin et al. 2016, A North American Hourly Assimilation and Model Forecast Cycle: The Rapid Refresh. *Mon. Wea. Rev*., **144**, 1669-1694.

Blake, B. T., J. R. Carley, T. I. Alcott, I. Jankov, M. E. Pyle, S. E. Perfater, B. Albright, 2018: An Adaptive Approach for the Calculation of Ensemble Grid-Point Probabilities. *Wea. Forecasting*, submitted.

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

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

Hamill, T.M., E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The U.S. National Blend of Models for Statistical Postprocessing of Probability of Precipitation and Deterministic Precipitation Amount. *Mon. Wea. Rev*., **145**, 3441-3463.

--------, T.M., G.T. Bates, J.S. Whitaker, D.R. Murray, M. Fiorino, T.J. Galarneau, Y. Zhu, and W. Lapenta, 2013: NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. *Bull. Amer. Meteor. Soc*., **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1

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

Herman, G.R., and R.S. Schumacher, 2018a: Money Doesn't Grow on Trees, But Forecasts Do: Forecasting Extreme Precipitation with Random Forests. *Mon. Wea. Rev*, in press; available online at: https://journals.ametsoc.org/doi/abs/10.1175/MWR-D-17-0250.1

---------- and R.S. Schumacher, 2018b: "Dendrology" in Numerical Weather Prediction: What Random Forests and Logistic Regression Tell Us About Forecasting Extreme Precipitation. *Mon. Wea. Rev*, in revisions; available online at: http://schumacher.atmos.colostate.edu/gherman/MDGOT_Diagnostics_Paper_LATEST.pdf

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

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

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

Roberts, N. M. and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev*., **136**, 78-97.

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

Appendix A

FFaIR 2018 -- Forecast Areas and Significant Events

Table 4. *Forecast areas and events for the first week of FFaIR taking place June 18-22, 2018.*

WEEK 1			
Forecast Valid End Date	**Valid Time (UTC) (18 - 00 PFF1) (00-06Z PFF2) (15-12 Day 1 ERO)**	**Forecast Area (Main Threats)**	**Notes**
6/19/2018	18 - 00	N/A	Significant flash flooding in Rockford, IL. Around 15 water rescues were performed. Flash flooding in Port Arthur, TX. Beginning of multi-day event for Texas Gulf Coast
	00 - 06	Iowa/S. Wisconsin/N. Illinois	
	15 - 12	Texas Gulf Coast/Montana/Iowa	
6/20/2018	18 - 00	Texas Gulf Coast	Flash flood emergency issued in Corpus Christi, TX. Flash flooding also occurred in Western Kansas
	00 - 06	Texas Gulf Coast	
	15 - 12	Texas Gulf Coast/Central Plains/Ohio River Valley	
6/21/2018	18 - 00	Iowa/SE South Dakota	Flash flooding continues along Gulf Coast of Texas. Flash flooding in Pittsburgh, PA with multiple water rescues.
	00 - 06	Iowa/SE South Dakota/SW Minnesota/W Illinois	
	15 - 12	Texas Gulf Coast/SE South Dakota/NW Iowa/SW Minnesota	
6/22/2018	18 - 00	Ohio/West Virginia/Virginia	Flash flooding in Richmond, VA with several water rescues NW of town. Richmond International Airport closed for over two hours due to flooded runways.
	00 - 06	Ohio/West Virginia/Virginia	
	15 - 12	Texas Gulf Coast/Illinois/Ohio River Valley/Central Virginia	
6/23/2018	18 - 00	Indiana/Ohio/West Virginia	Widely scattered flash flood reports through northern Indiana, central Ohio, and West Virginia.
	00 - 06	Ohio/West Virginia/Virginia	
	15 - 12	Indiana/Ohio/West Virginia/Virginia/S Arkansas/Montana/NW North Dakota	

Table 5. *Forecast areas and events for the second week of FFaIR taking place June 25-29 2018.*

WEEK 2			
Forecast Valid End Date	**Valid Time (UTC) (18 - 00 PFF1) (00-06Z PFF2) (15-12 Day 1 ERO)**	**Forecast Area (Main Threats)**	**Notes**
6/26/2018	18 - 00	Kentucky/Virginia/North Carolina	Roads closed in NE Nebraska due to heavy rain and resultant flooding.
	00 - 06	North Carolina/South Carolina	
	15 - 12	Iowa/E Nebraska/Kentucky/NE Oklahoma/SW Missouri	
6/27/2018	18 - 00	Missouri/Kentucky/Illinois	Several flash flood reports in the northern and western suburbs of Chicago, IL.
	00 - 06	Missouri/Kansas/NW Illinois	
	15 - 12	S Wisconsin/N Illinois/Kentucky/Missouri	
6/28/2018	18 - 00	Pennsylvania	Widespread reports of flooding throughout Middle Tennessee. Major flooding also reported in Pine Grove, PA.
	00 - 06	Pennsylvania	
	15 - 12	Pennsylvania/West Virginia/Tennessee	
6/29/2018	18 - 00	North Dakota	Flash flooding in Twin Butte, ND and very heavy rainfall throughout northern and central ND.
	00 - 06	North Dakota	
	15 - 12	North Dakota/Tennessee/New England	
6/30/2018	18 - 00	N/A	No significant reports
	00 - 06	South Dakota	
	15 - 12	Northern Plains/SW Texas/Louisiana/Mississippi	

Table 6. *Forecast areas and events for the third week of FFaIR taking place July 9-13 2018.*

WEEK 3			
Forecast Valid End Date	**Valid Time (UTC) (18 - 00 PFF1) (00-06Z PFF2) (15-12 Day 1 ERO)**	**Forecast Area (Main Threats)**	**Notes**
7/10/2018	18 - 00	Arizona	Flash flooding in and around the greater Phoenix, AZ region.
	00 - 06	Texas	
	15 - 12	Southwest states/Central Texas	

7/11/2018	18 - 00	N/A	A couple of flash flood reports in central Arizona.
	00 - 06	Arizona	
	15 - 12	Southwest states/Central Gulf Coast	
7/12/2018	18 - 00	New Mexico/Arizona/SW California/S Nevada/S Utah	Heavy rainfall and flooding with reports of roads washed out near Mora, MN. Several reports in Arizona and around Death Valley, CA as well.
	00 - 06	N Minnesota	
	15 - 12	NW Minnesota/Southwest states	
7/13/2018	18 - 00	New Mexico/Arizona/SW California/S Nevada/S Utah	Flash flooding reported near the Grand Canyon forcing hikers/tourists to higher ground.
	00 - 06	SE South Dakota/S Minnesota/Wisconsin	
	15 - 12	Southwest states/Northern Plains	
7/14/2018	18 - 00	N/A	Southwest monsoon stays persistent with a few widely scattered reports.
	00 - 06	Kansas	
	15 - 12	Southwest states/Central Plains	

Table 7. *Forecast areas and events for the fourth week of FFaIR taking place July 16-20 2018.*

WEEK 4			
Forecast Valid End Date	**Valid Time (UTC) (18 - 00 PFF1) (00-06Z PFF2) (15-12 Day 1 ERO)**	**Forecast Area (Main Threats)**	**Notes**
7/17/2018	18 - 00	West Virginia	No significant reports
	00 - 06	Arizona	
	15 - 12	Southwest states/Appalachian Mountains	
7/18/2018	18 - 00	Mid-Atlantic/New England	Widespread flash flood reports in the Mid-Atlantic and New England, particularly in Massachusetts where many roads were flooded in places like Oxford and Worcester, MA. Water rescues in the Washington D.C. area as well.
	00 - 06	Central Plains	

	15 - 12	Mid-Atlantic/New England/Central Plains/Southeast	
7/19/2018	18 - 00	E Kansas/NE Nebraska	Flash flood reports, road closures, and extremely heavy rainfall (8.87 inches) in Aurora, SD.
	00 - 06	E Kansas/NE Nebraska/W Iowa	
	15 - 12	Southwest states/Northern & Central Plains	
7/20/2018	18 - 00	S Minnesota/NE Iowa	No significant reports
	00 - 06	N Arizona/S Nevada/SW Utah	
	15 - 12	Southwest states/Northern Mississippi River Valley/Southeast	
7/21/2018	18 - 00	Ohio/Indiana/Kentucky	A few flash flood reports in the Ohio River Valley associated with a severe weather outbreak.
	00 - 06	Kentucky/Tennessee	
	15 - 12	Ohio River Valley/South Carolina/Southwest states	

APPENDIX B

Participants

*Denotes participant was an observer

**Denotes participants split/shared the week

Week	WPC Forecaster	WFO/RFC/Other	Research/Academia	EMC
June 18 – 22	Alex Lamers	Belkys Melendez - WFO MHX Jacki Ritzman - WFO MQT Jason Deese - WFO FFC	Brad Diehl - MDL Monica Stone - OWP John Forsythe - CIRA Ming Hu - ESRL GSD Patrick Skinner - NSSL	Eric Aligo
June 25 – 29	Gregg Galina	Tracy McCormick - NERFC Alex DeSmet - WFO PIH Jeremy Buckles - WFO MRX Adrian Wynn - Met Office Flood Centre	Ama Ba - MDL Kate Abshire - OWP Eric James - ESRL GSD Matt Kelsch - UCAR/COMET Jessica Choate - NSSL Tara Jensen - DTC*	Tracey Dorian** Geoff Manikin**
July 9 – 13	Andrew Orrison	Charles Ross - WFO CTP Brett Lutz - WFO MFR Jeremy Wesely - WFO GID	Dave Rudack - MDL** Eric Engle - MDL** Glen Romine - NCAR David Dowell - ESRL GSD Adam Clark - NSSL Eric Loken - OU Phd Student Greg Herman - CSU	Matt Pyle Logan Dawson** Alicia Bentley**
July 16 – 20	Rich Otto	Blair Holloway - WFO CHS Larry Hopper - WFO PSR	Chandra Kondragunta - OAR Jeff Craven - MDL Ryan Sobash - NCAR Jeff Duda - ESRL GSD Lisa Darby - ESRL PSD Kent Knopfmeier - NSSL Keith Brewster - OU/CAPS	Ed Strobach Mallory Row** Ben Blake**

APPENDIX C

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. ()). 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.

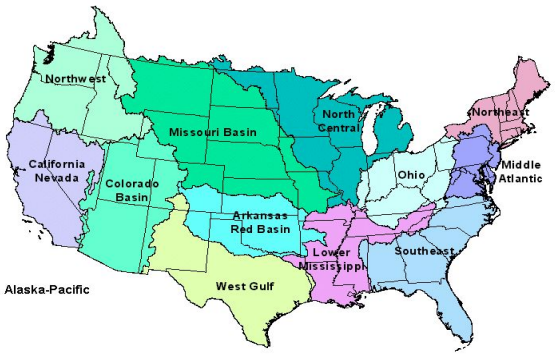


Figure 63. *Showing domain for each NWS River Forecast Center (NOAA/NWS (water.weather.gov)*

Precipitation Average Recurrence Intervals

Precipitation Average Recurrence Intervals (ARIs) are frequency estimates generated mainly from NOAA Atlas-14 Climatology of USGS rain gages. Statistical analyses are applied to the precipitation climatology to generate precipitation amounts representing the approximate frequency of occurrence (e.g. 1 year, 5 years, 100 years, etc.) for various accumulation periods (e.g. 5 minutes, 30 minutes, 3 hours, 24 hours, etc.). ARIs can help to identify how rare a rainfall event is for a given area, alerting forecasters to abnormal or potentially extreme rainfall events. Standard ARIs are available for intervals of 2, 5, 10, 25, 100, 500 and 1000 years, and are measured in inches, and do not account for antecedent conditions. An example of the 6 hour, 100 year recurrence interval is shown in Figure 64.

For 2018 FFaIR, we provided fully-stitched grids (Herman and Schumacher, 2016). Thresholds come from NOAA Atlas 14 for most of CONUS. This includes the New England area which received an Atlas 14 update in autumn 2015, the most recent update to Atlas 14. Two regions of CONUS have not yet received NOAA Atlas 14 updates: Texas and the northwest, which is comprised of Washington, Oregon, Idaho, Montana, and Wyoming. For Texas, thresholds from Technical Paper 40 (TP-40; Hershfield 1961) were used; digital grids of the selected ARIs between 1 and 100 years were included for 6- and 24-hour precipitation accumulations. Therefore no additional processing was required. For the northwest, TP-40 did not provide

coverage; instead, NOAA Atlas 2 (Miller et al. 1973) was used for these thresholds. The only grids that had been digitized were 2- and 100-year ARIs for 6- and 24-hour accumulations. However, the two frequency thresholds at each point for these five states combined with the knowledge that these threshold estimates were originally derived from a (two-parameter) Gumbel distribution, a Gumbel distribution could then be fitted to each point (two equations, two unknowns), and estimates for the 1-, 5-, 10-, 25-, and 50-year ARIs derived from those Gumbel fits. These different threshold estimate sources were then stitched together to form CONUS-wide grids.

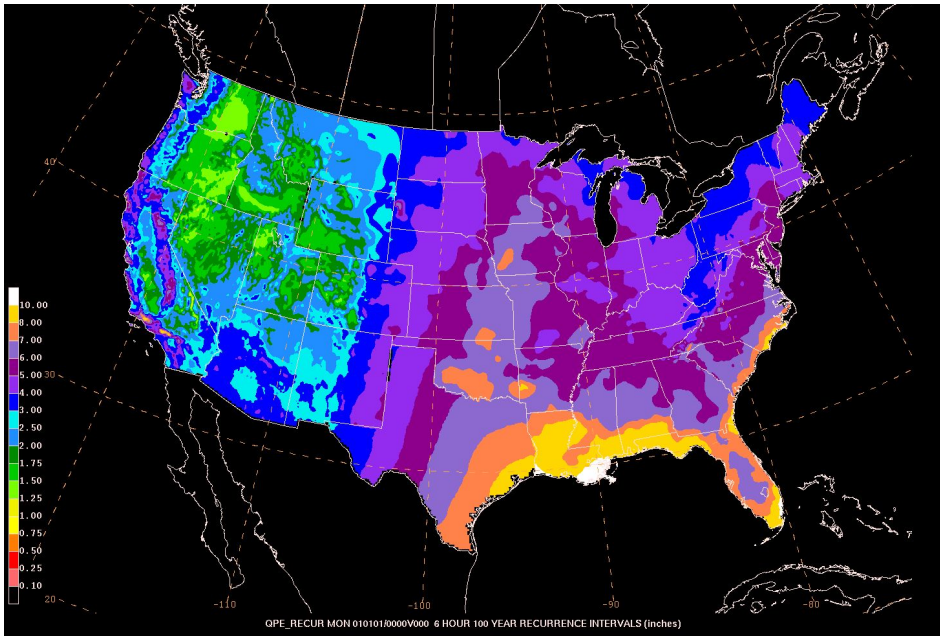


Figure 64. *An example of a full Average Recurrence Interval map (100 year ARI over 6 hours) available to forecasters both operationally at WPC and in the FFaIR Experiment.*

National Water Model (NWM) Experimental Products

The operational NWM runs an hourly uncoupled analysis (simulation of current conditions). Short-range forecasts are executed hourly while medium-range forecasts out to 10 days are produced four times per day. A daily ensemble long range forecast to 30-days is also produced. All model configurations provide streamflow for 2.7 million river reaches and other hydrologic information on 1km and 250m grids. The NWM provides complementary hydrologic guidance at current NWS river forecast locations and significantly expanded guidance coverage and type in underserved locations (Figure 65).

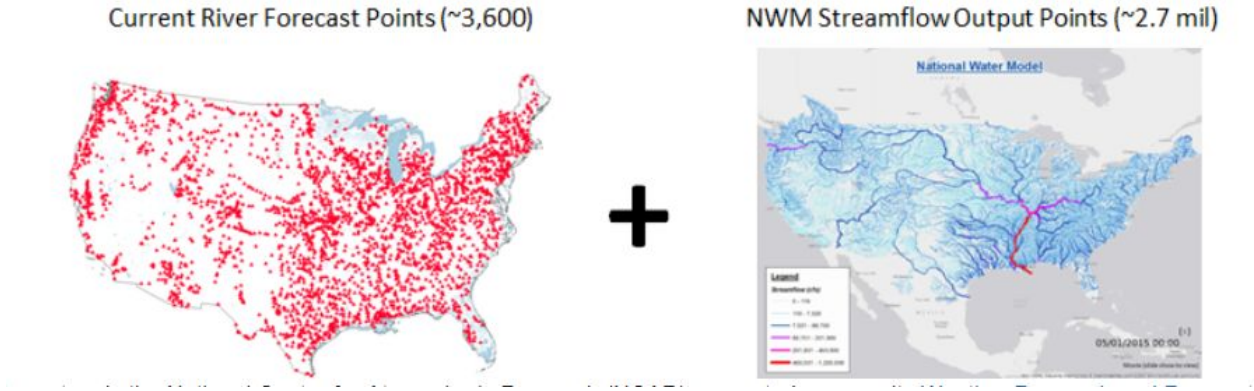


Figure 65. *National Water Model River Forecast Points(left) and Streamflow Output (right).*

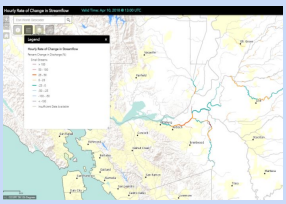
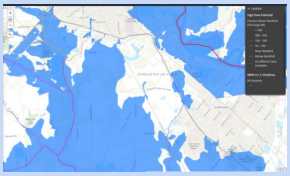
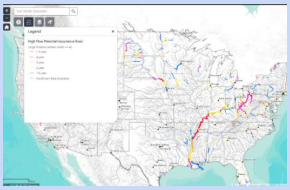
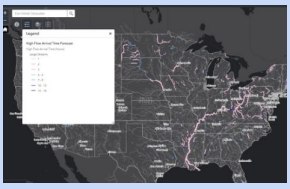
The core of this system is the National Center for Atmospheric Research (NCAR)-supported community Weather Research and Forecasting Hydrologic model (WRF-Hydro). It ingests forcing from a variety of sources including Multi-Radar/Multi-Sensor System (MRMS) radar-gauge observed precipitation data, and High Resolution Rapid Refresh (HRRR), Rapid Refresh (RAP), Global Forecasting System (GFS) and Climate Forecast System (CFS) Numerical Weather Prediction (NWP) forecast data. WRF-Hydro is configured to use the Noah-MP 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 National Hydrography Dataset (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). The system includes an analysis and assimilation configuration along with three forecast configurations. United States Geological Survey (USGS) streamflow observations are assimilated into the analysis and assimilation configuration and all four configurations benefit from the inclusion of 1,260 reservoirs.

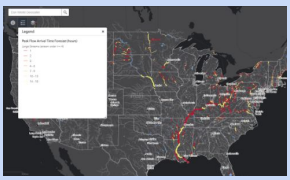
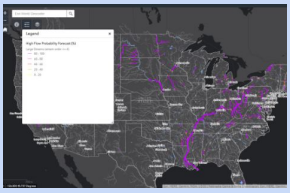
In order to quickly synthesize the spatial and temporal patterns within the NWM output, the Office of Water Prediction (OWP) and NCAR development teams have created experimental post-processed visualizations published to a suite of dynamic map services that allow users to zoom in and and pan around areas of interest. As in operations, this map suite includes streamflow (current discharge in NWM (v1.2) stream reaches), streamflow anomaly (stream reaches compared seasonal average), and soil moisture (current volumetric soil moisture content in the 0 - 40 cm soil layer). The experimental visualizations utilize a dataset of recurrence flows for each stream reach in the NWM (v1.2) hydrologic network. Recurrence flows for each stream reach were derived from a 23-year retrospective analysis of the NWM (v1.0). For a given stream reach, the maximum flows for each year were ranked from smallest to largest.

For each maximum flow value, the corresponding recurrence flow was determined by computing i / (n + 1), where i is the rank of the maximum flow, and n is the total number of years. These (maximum flow, recurrence flow) pairs for a given stream reach were then plotted

and used to construct a curve from which the 1.5-year recurrence flow could be derived. To test experimentally during FFaIR, these NWM products for the 2018 FFaIR Experiment can be found in Table 8.

***Table 8**. A summary of the post-processed NWM products available for evaluation during the 2018 FFaIR Experiment.*

NWM Visualization	Description	Valid Times	Sample
Rate of Change in Streamflow	Depicts rate of change in streamflow over the past hour for all stream reaches in the NWM (v.12) hydrologic network above 75% of their 1.5-year recurrence flow (discharge or vertical water depth).	1 hour	
High Flow Potential	Depicts NWM (v1.2) stream reaches expected to be above their 1.5-year recurrence flow (estimate of bankfull discharge). Will include inundation for the Texas West Gulf River when applicable.	Hourly (based on current conditions derived from the NWM Analysis & Assimilation)	
Maximum High Flow Potential	Depicts NWM (v1.2) maximum recurrence flow expected to be reached by all stream reaches that are predicted to be at or above their 1.5-year recurrence flow (our estimate of bankfull discharge). Will include inundation for the Texas West Gulf River when applicable.	Short Range: Next 18 hrs Medium Range: Next 3 days and Next 5 days	
High Flow Arrival Time	Depicts when NWM (v1.2) stream reaches that are expected to be at or above their 1.5-year recurrence flow.	Short Range: Next 18 hrs Medium Range: Next 10 days	

Peak Flow Arrival Time	Depicts when NWM (v1.2) stream reaches that are expected to be at or above their 1.5-year recurrence flow will be at their peak flow during the forecast period.	Short Range: Next 18 hrs Medium Range: Next 10 days	
High Flow Probability	Depicts the probability that NWM (v1.2) stream reaches will be at or above their 1.5-year recurrence flow.	Short Range: Next 6-8 hrs Medium Range: Next 24-48 hrs, Next 48-72 hrs	

GFDL/EMC FV3-GFS

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

NOAA/NWS selected the Geophysical Fluid Dynamics Laboratory (GFDL) finite volume cubed-sphere dynamical core (FV3) as the Weather Service's Next Generation Global Prediction System (NGGPS). The current operational GFS, which has a spectral dynamical core, is scheduled to be replaced by the proposed GFS with FV3 dynamical core and different microphysics in winter 2018/2019. The implementation of the FV3 dynamical core into the NOAA environmental modeling system infrastructure provides 1) updates to the global data assimilation system to exchange information between the forecast model on the cubic-sphere grids and data assimilation on Gaussian lat-lon grids and, 2) a new workflow system for both research and operation.

This GFS version has a horizontal resolution of 13 km, and has 64 levels in the vertical extending up to 0.2 hPa. It uses the same physics package as the current operational GFS except for 1) the replacement of UTChao-Carr microphysics with the more advanced GFDL microphysics, 2) an updated parameterization of ozone photochemistry with additional production and loss terms, 3) a newly introduced parameterization of middle atmospheric water vapor photochemistry and, 4) a revised bare soil evaporation scheme.

The data assimilation system will be updated to include IASI moisture channels; ATMS all-sky radiances; a fix for an issue with the Near Sea Surface Temperature (NSST) in the Florida Strait; an upgrade to the use of CrIS radiances; addition of NOAA-20 CrIS and ATMS data ; addition of

Megha-Tropiques Saphir data; addition of ASCAT data from MetOp-B; and several additional minor changes. The ensemble part of the hybrid data assimilation will also increase in resolution from 35 km to 25 km. For more information, please see https://docs.google.com/document/d/112GG7yBDMPmEcrNi1R2ISsoLcivj5WPivSnf9Id_lHw/edit.

For the 2018 FFaIR Experiment, the 13 km quasi real-time FV3 GFS forecast was run four times per day, with hourly output up to 120 hours and then 3 hourly output up to 384 hours as provided by EMC.

ESRL High Resolution Rapid Refresh - Experimental → Operational (HRRRv3)

The operational *(prior to July 12, 2018)* HRRR model (version 2) (https://rapidrefresh.noaa.gov/hrrr/HRRR/Welcome.cgi?dsKey=hrrr_ncep_jet) is on a 3 km grid and uses boundary conditions from the hourly updated, radar-DFI-assimilated Rapid Refresh (RAP) model. The HRRR uses GSI hybrid data assimilation (instead of 3D-VAR), is initialized with latest 3-D radar reflectivity and features a WRF-ARW core version 3.6.1, Thompson microphysics, and is fully convection allowing. The operational HRRR is run every hour and produces hourly and sub-hourly forecasts out to 18 hrs.

During the 2018 FFaIR experiment, the HRRR version 3 (HRRRv3; https://rapidrefresh.noaa.gov/hrrr/HRRR), replaced the current version 2 in operations. HRRRv3 runs every hour with output to 18 hrs (01z, 02z, 04z, 05z, ...) or 36 hours (00z, 03z, 06z...) and remains on a 3-km grid with hourly runs that are changed to the forecast lengths listed above. The HRRRv3 is initialized with an hour of 3-D radar reflectivity using a latent-heating specification technique including some refinements in this latent-heating from the parent RAPv4 model (Figure 66). The HRRRv3 uses grid-point statistical interpolation (GSI) hybrid GFS ensemble-variational data assimilation of conventional observations. Building upon the advancements in the operational HRRRv2 at NCEP, HRRRv3 includes assimilation of TAMDAR aircraft observations, refines assimilation of surface observations for improved lower-tropospheric temperature, dewpoint (humidity) winds and cloud base heights and places more weight on the ensemble contribution to the data assimilation. HRRRv3 adds assimilation of lightning flash rates as a complement to radar reflectivity observations through a similar conversion to specified latent heating rates during a one-hour spin-up period in the model. HRRRv3 also contains numerous model changes including an update to WRF-ARW version 3.9 including the Thompson microphysics, transition to a hybrid sigma-pressure vertical coordinate for improved tropospheric temperature, dewpoint and wind forecasts along with a higher resolution (15 second) land use dataset. Physics enhancements have also been made to the MYNN planetary boundary layer (PBL) scheme and RUC land surface model along with additional refinements to shallow cumulus/sub-grid-scale cloud parameterizations including enhanced interactions with the radiation and microphysics schemes for greater retention of cloud features. In addition to the improved PBL and cloud physics, more recent upgrades in the HRRRv3 include improved radar assimilation and hybrid vertical coordinate.

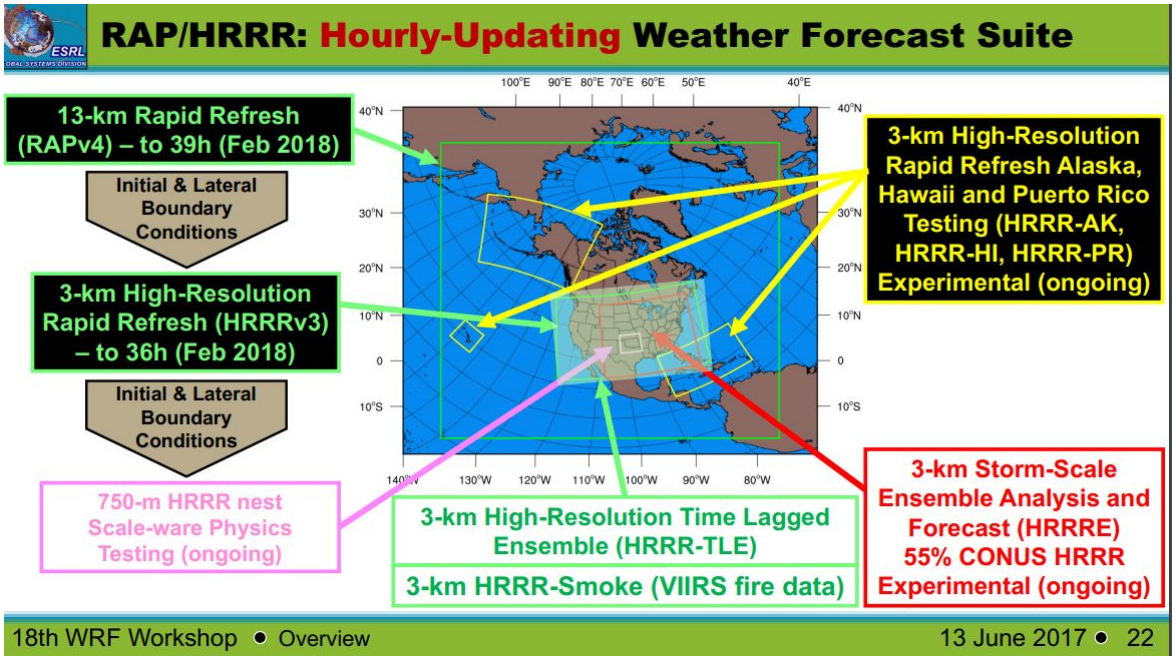


Figure 66. *A snapshot of the interdependencies of the RAP/HRRR hourly-updating forecast suite with associated domains.*

HRRRv3 is run hourly and provides forecasts as follows:

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

3.2 Experimental Ensemble Model Guidance

ESRL/GSD Experimental HRRR Ensemble (HRRR-E)

The HRRRE configuration is identical to the HRRRv3 (described above) with the WRF-ARW version 3.8+ core, combining elements of versions 3.8 and 3.9 plus other GSD-specific features. Differing from the HRRRv3 is the 3-km domain which covers the central and eastern US only (60% of HRRR domain as shown in Figure 67), and a standard vertical coordinate is used instead of a hybrid coordinate. The physics in the HRRRE are as described in (Benjamin et al. 2016), MWR http://journals.ametsoc.org/doi/abs/10.1175/MWR-D-15-0242.1, except that a deep convection parameterization is not used on the 3-km domain.

Nested 15-km and 3-km domains

36 members initialized daily at 0300 UTC

- Initial mean from GFS (atmos.) and RAP-HRRR (soil)
- Atmospheric perturbations from GFS ensemble (GDAS)
- Random soil-moisture perturbations

Hourly cycling with EnKF DA

- 0300 – 0000 UTC (21 hours)
- Conventional observations both domains
- Reflectivity observations 3-km domain only
- Analysis variables: U, V, PH, T, MU, QVAPOR, QCLOUD, QICE, QRAIN, QSNOW
- BC perturbations, posterior inflation

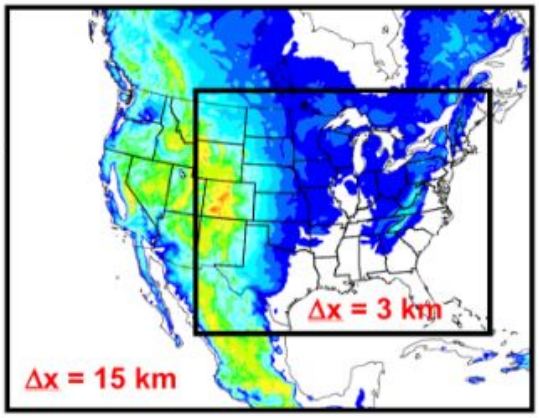


Figure 67*. Graphic and description of the HRRRE 3km limited CONUS domain.*

The deterministic HRRR currently assimilates observations with a hybrid ensemble-variational (EnVAR) method, and the background ensemble for this assimilation is the 80-member GDAS (GFS) ensemble. During the spring, GSD tested the use of a higher-resolution, convective-allowing ensemble instead for assimilation in the HRRRE. 36 members initialized daily at 0300 UTC with the initial mean from GFS (atmos.) and RAP-HRRR (soil). Atmospheric perturbations came from the GFS ensemble (GDAS), and featured random soil-moisture perturbations. A background ensemble with explicit convection enabled direct assimilation of high-resolution observations such as radar reflectivity in convective storms. This high-resolution, ensemble-based assimilation could lead to improved forecasts, particularly in the 0-12 h range.

Longer forecasts require more attention to model error. Soil-moisture perturbations and stochastic parameter perturbations to the MYNN PBL scheme have both been tested as ways to introduce realistic growth of ensemble spread during the 0-36 h forecast. Stochastic parameter perturbations to other parameterization schemes will also be tested in the coming year. HRRRE forecasts with stochastic physics will be evaluated internally in 2018 before becoming a candidate ensemble for the 2019 Spring Experiment.

New for the 2018 FFaIR experiment is the coordinated design of the NCAR Ensemble and HRRRE. The 2018 NCAR Ensemble and HRRRE share features such as hourly cycling, a large outer analysis grid with 15-km grid spacing, and a nested grid with 3-km grid spacing. Forecast-model and data-assimilation codes will also be made as similar as possible. One primary difference between the two systems is continuous cycling in the NCAR Ensemble versus once-daily partial cycling in the HRRRE. Table 9 shows the differences in configurations.

Table 9. *Membership spread and attribute differences between the HRRRE and NCAR Ensemble featured in the 2018 FFaIR Experiment.*

Attribute	HRRRE	NCAR Ensemble
Dynamic Core	ARW-only	ARW-only
Physics	RAP/HRRR Suite	RAP/HRRR Suite
Initial Condition Atmos	GFS + GDAS partial cycle start	Continuous cycle
Initial Condition Soil	RAP/HRRR soil perturbations	RAP/HRRR soil perturbations
Hourly DA Cycling	EnKF w/posterior inflation	EnKF w/posterior inflation
Lateral Boundary	Perturbations	Perturbations
Stochastic Physics	LSM, PBL, MP, Cu	None
Number of DA Members	36	80
Number of Fcst Members	9	9

During the 2018 FFaIR experiment, the HRRRE analyses and forecasts provided initial and boundary conditions for a prototype Warn-on-Forecast system run at the National Severe Storms Laboratory. The Warn-on-Forecast project is developing on-demand, regional, high-resolution, ensemble-based, 0-6 h numerical weather prediction capabilities to support warnings of severe convective storms and flash flooding.

Ensemble Forecast Details

- 3-km horizontal grid spacing
- 9-member, half-CONUS, 48-h forecasts initialized from first 9 members of data-assimilation ensemble (nested 15-km and 3-km analyses) at **1200 UTC**
    - *24-h forecast typically available at 1600 UTC*
    - *48-h forecast typically available at 1800 UTC*
- 9-member, half-CONUS, 18-h forecasts initialized from first 9 members of data-assimilation ensemble (nested 15-km and 3-km analyses) at **1800 UTC**
- 9-member, half-CONUS, 18-h forecasts initialized from first 9 members of data-assimilation ensemble (nested 15-km and 3-km analyses) at **2100 UTC**
- 9-member, full-CONUS, 36-h forecasts initialized from first 9 members of data-assimilation ensemble (15-km analyses) at **0000 UTC**
    - Control version (real-time distribution): WRF 3.8 without stochastic physics
    - Parallel version (internal evaluation only): WRF 3.9 with stochastic physics
- Random perturbations to MU, U, V, T, and QVAPOR added to boundary conditions of each ensemble member
- Post processing: An ensemble post-processing system is applied to the nine HRRRE forecast members to produce all-season weather hazard probabilities including heavy rainfall as is done with the time-lagged HRRR. HRRRE probabilities are the fraction of members that exceed a given threshold, or predict a given precipitation type, at a point.

The final probability field (100*(n/total)) is smoothed using a Gaussian filter of width 25 km.

NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e)

The NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e) is a 36-member WRF-based ensemble data assimilation system used to produce very short-range (0-6 h) probabilistic forecasts of supercell thunderstorm, rotational characteristics, rainfall, hail and high winds. The starting point for each day's experiment will be the experimental HRRRE provided by ESRL/GSD. The HRRRE is initialized at 0300 UTC and updated by hourly EnKF assimilation of conventional observations and Multi-Radar/MultiSensor (MRMS) radar reflectivity for 21 h to 0000 UTC. A 48-hr ensemble forecast launched from the 1200 UTC HRRRE analysis is used to provide initial and boundary conditions for the NEWS-e system for the period 1600 UTC Day 1 – 0400 UTC Day 2.

The daily NEWS-e domain location targeted the primary region where heavy rainfall was anticipated and covered a ~900-km wide region. All ensemble members utilize the NSSL 2-moment microphysics parameterization and the RAP land-surface model, but the PBL and radiation physics options are varied amongst the ensemble members to address uncertainties in model physics. Multi-Radar/Multi-Sensor (MRMS) radar reflectivity and Level II radial velocity data, cloud water path retrievals from the GOES-16 imager, and NCEP prepbufr will be assimilated every 15 minutes using an EnKF approach, beginning at 1600 UTC each day. The first forecast of the day will be initialized at 1800 UTC and another forecast at 1830 UTC. The NEWS-e ensemble forecasts will output WRF history files every 5 minutes.

Table 10 shows the differences in model specifications between HRRRE and NEWS-e, and Figure 68 shows an example of a WPC Day 1 Excessive Rainfall Outlook and corresponding NEWS-e grid with WSR-88D radars used for data assimilation overlaid.

Table 10. *HRRRE and NEWS-e configuration comparison.*

	HRRRE	**NEWS-e**
Model Version	WRF-ARW v3.8+	WRF-ARW v3.8+
Grid Points	1150 × 960 × 50	300 × 300 × 50
Grid Spacing	3 km	3 km
EnKF Cycling	36 mem w/ **GSI-EnKF every 1 h**	36 mem w/ **GSI-EnKF every 15 min**
Observations	**Conventional obs:** *T*, q_v, *u*, *v*, and *p* from rawinsonde, aircraft, surface (land and marine), profiler; **MRMS reflectivity**	**Conventional obs:** *T*, q_v, *u*, *v*, and *p* from rawinsonde, aircraft, surface (land and marine), profiler; **Doppler velocity from WSR-88D sites** **MRMS reflectivity** **Cloud-water path (GOES-16)** **AERI and Doppler Lidar**
Radiation LW/SW	RRTMG/RRTMG	Dudhia/RRTM or RRTMG/RRTMG
Microphysics	Thompson (aerosol aware)	NSSL 2-moment

Cumulus Param.	none	none
PBL	MYNN	YSU, MYJ, or MYNN
LSM	RUC (Smirnova)	RUC (Smirnova)

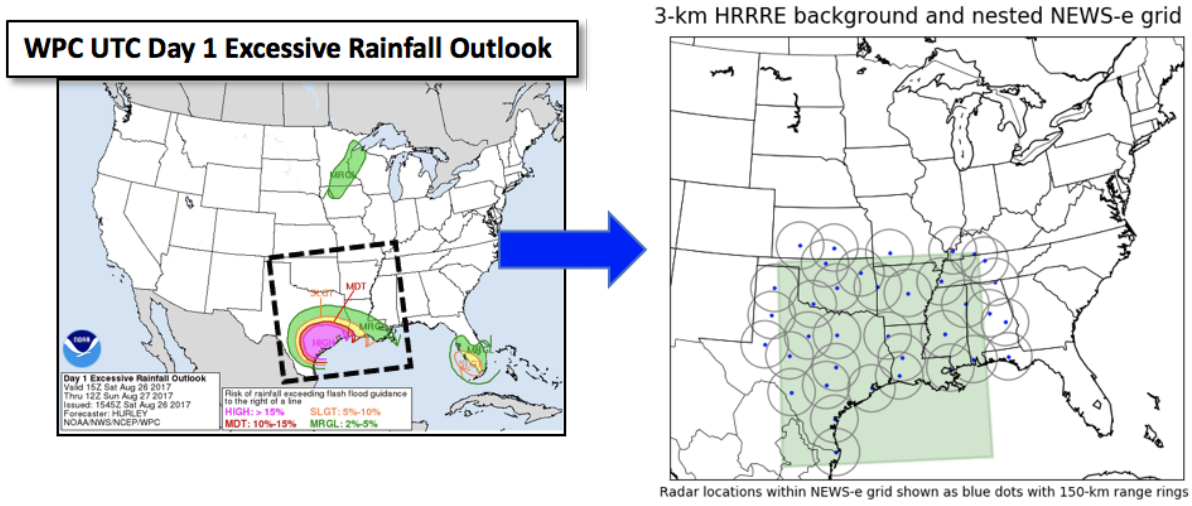


***Figure 68**. WPC Day 1 Excessive Rainfall Outlook (left) and corresponding NEWS-e grid (right).*

Table 11 shows the NEWS-e products that were made available during the 2018 FFaIR Experiment. These forecasts were viewable using the web-based NEWS-e Forecast Viewer (https://www.nssl.noaa.gov/projects/wof/news-e/wpc/).

***Table 11.** Probabilities available from the NEW-e for FFaIR 2018*

NEWS-e Probabilistic QPF	Probability of 6-h precip > 0.5, 1.0, 2.0, 3.0, 6.0 in Probability of 3-h precip > 0.5, 1.0, 2.0, 3.0 in Probability of 1-h precip > 0.5, 1.0, 2.0 in
NEWS-e Probabilistic ARI Exceedance	Probability of 6-h precip exceeding 1, 2, 5, 10, 25, 50 and 100-yr ARI

EMC Experimental High Resolution Ensemble Forecast (HREFv2.1)

A unique version of the HREFv2 was run for the 2018 FFaIR Experiment. The operational HREFv2 is an ensemble product generator utilizing multiple cycles of operational convective allowing models of ~3 km horizontal scale: namely the High-Resolution Window (HiresW; both the Weather Research and Forecast (WRF) Advanced Research WRF (ARW) and Non-hydrostatic Multiscale Model on the B-grid (NMMB) members) and the NAM CONUS nest (Figure 69). In the experimental HREFv2.1 provided for this experiment, the 36 h long High Resolution Rapid Refresh (HRRR) runs are utilized as well. With the addition of the HRRR, the parallel HREFv2.1 system has ten members: the two most recent runs of the NAM CONUS nest, the HRRR, the HiresW NMMB, and of two different HiresW ARW members. The time-lagged NAM and HRRR members are 6 h old, while the time-lagged HiresW members are 12 h old. These time-lagged

members are given less weight than the current cycle members: the 6 h old cycles are given 87.5% weighting while the 12 h old cycles are given 75% weighting.

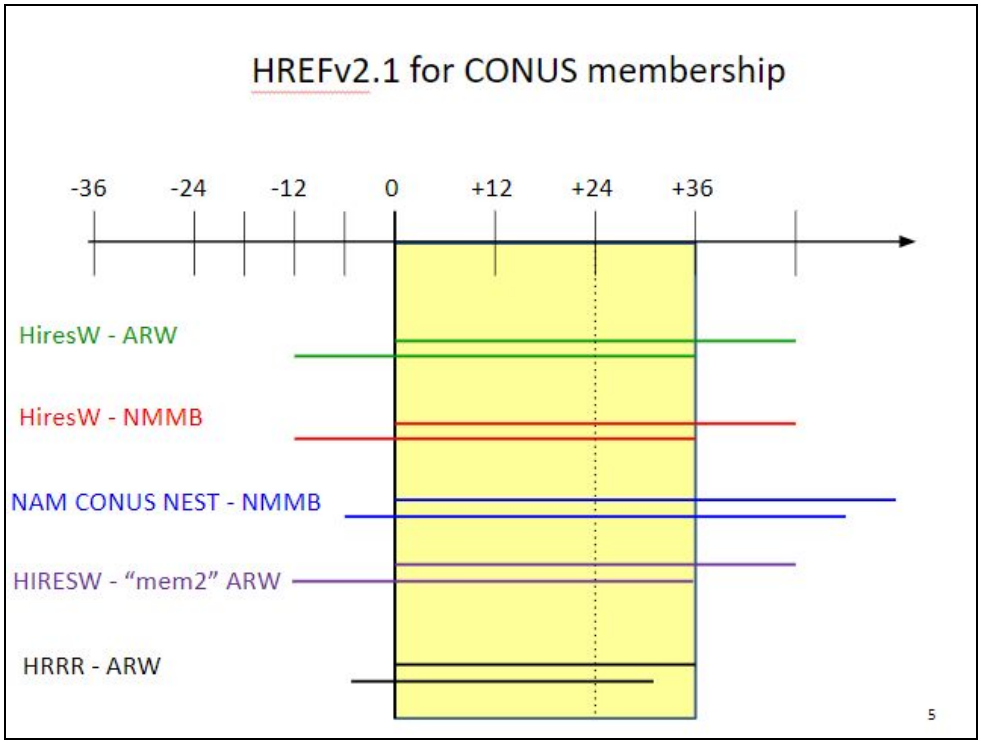


Figure 69. *The HREFv2.1 ensemble membership including one real-time and one time-lagged ARW, one real-time and one time-lagged NMMB, one real time and one time-lagged NAM CONUS Nest, one real-time and one time-lagged "mem2" ARW and one real-time and one time-lagged HRRR-ARW.*

Probabilistic guidance includes neighborhood probabilities (Harless et al. 2010) and Gaussian smoothing (Silverman 1986) of probabilities for select fields, including for QPF. This experimental system also includes examples of the Ensemble Agreement Scale (EAS)* Probabilities (Blake et al. 2018; Roberts and Lean 2008) for QPF. An example of these variations can be seen in Figure 70. Probabilities of precipitation exceeding flash flood guidance (FFG) and average recurrence interval (ARI) values also are generated by this experimental version.

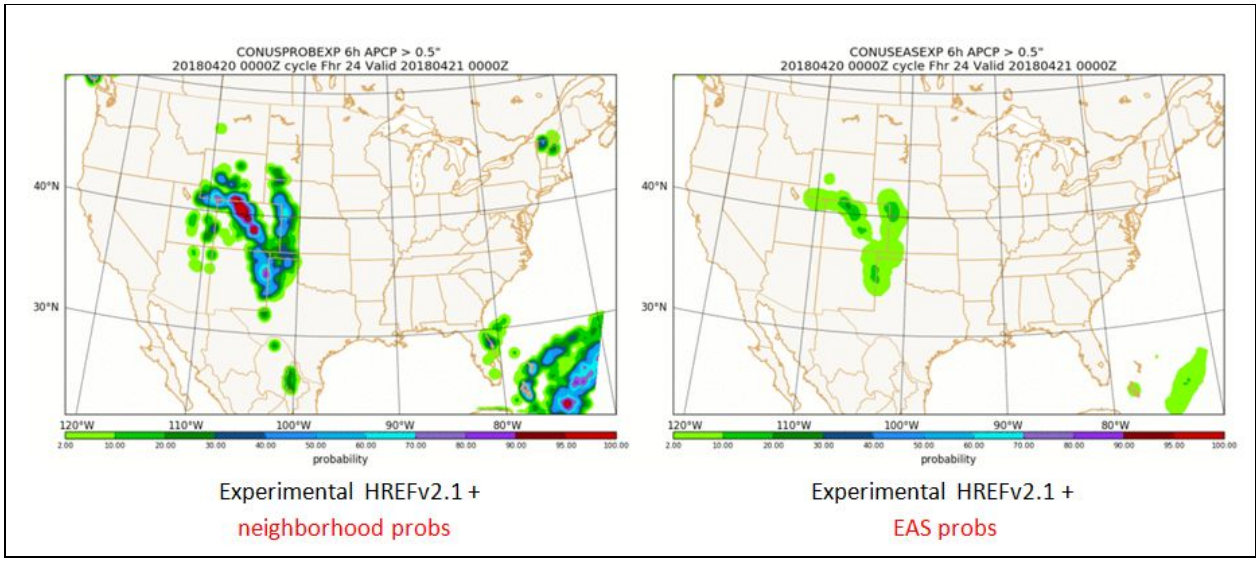


Figure 70. *Probability of 6-hour total accumulation exceeding 0.5" valid 00 UTC 4/21/2018 as calculated by the parallel HREF with the neighborhood filter (left), and parallel HREF with EAS technique applied.*

This experimental HREF was run for the 00 UTC, 06 UTC, 12 UTC, and 18 UTC cycles, generating output to 36 hours from the cycle time. A description of the products evaluated in the FFaIR experiment:

- Probability matched mean precipitation. in three forms (as a conventional mean, as a PM mean (Ebert, 2001), and a 50/50 blend of the conventional and PM means).
- Probability of exceedance of QPF at various fixed thresholds over several duration periods (e.g., percentage of the ensemble exceeding 3" over a 6 h period). In the parallel HREF, the lowest thresholds (0.25" or less) are being provided as point probabilities and EAS probabilities. Heavier thresholds (0.5" and higher) are provided as neighborhood probabilities and EAS probabilities. The neighborhood probabilities are computed over a ~40 km radius neighborhood, and also are Gaussian smoothed.

Additional fields available to view:

- Ensemble mean precipitation and a 50/50 blend of the mean precipitation using the average of the conventional and PM means.
- Probabilities of QPF exceeding ARI (e.g., percentage of ensemble exceeding the 50 year recurrence interval value for a 6 h accumulation period) and FFG guidance (1 h, 3 h, and 6 h periods) values also are provided by this version, and are produced as smoothed neighborhood probabilities.

In addition to a suite of fields relative to precipitation forecasting, derived probabilities will be available from a special HREFv2.1 run for the 2018 FFaIR Experiment (Table 12).

Table 12. *Probabilities available from the special HREFv2.1 for FFaIR 2018*

HREFv2 Probabilistic QPF	1-h, 3-h QPF > 0.5, 1, 2 inches 6-h, 12-h, 24-h QPF > 1, 2, 3 inches

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

***With high-resolution ensembles, i.e. those where convection is not parameterized, spatial displacement at the grid point level is large. This issue has motivated the development of methods which account for this spatial uncertainty. As part of the USWRP project, fractional coverage approaches for the generation of point probabilities were developed (e.g. "neighborhood methods"). The EAS method attempts to account for the fact that a uniform radius is not always appropriate, i.e. orographically forced precipitation. In such cases, the traditional fractional coverage approach can reduce the probabilities of these often well handled events. Therefore, a variable radius approach has been developed based upon ensemble agreement scale (EAS) similarity criteria outlined in Dey et al. (2016). This approach varies the neighborhood radius size according to member-member similarity criteria. In this method, the radius sizes range from 10-km, for member forecasts that are in good agreement (e.g. lake effect, complex terrain, very short forecasts, etc.), to 100-km.*

OU/CAPS WRF-ARW+FV3 SSEFX

The experimental Storm-Scale Ensemble Forecast (SSEFX) is generated with the Weather Research and Forecast (WRF) modeling system (Version 3.9.1.1), with the Advanced Research WRF (ARW) core, and the experimental GFDL FV3. CAPS will produce 15 (13 ARW and 2 FV3) members to support the FFaIR Experiment. The 3-km FV3 will feature Thompson microphysics. Membership details can be found in Table 13. Major features for 2018 include:

- **3-km** horizontal grid spacing over the CONUS domain (1620×1120)
- WRF version **3.9.1.1** is used for 2018 season (coupled with ARPS v5.4)
- 00 UTC 60-hour forecast (one deterministic member may go out to 84 h)
- ARPS 3DVAR analysis of radar data

Table 13. *Membership characteristics of the SSEFX for the 2018 FFaIR Experiment. NAMa and NAMf refer to 12 km NAM analysis and forecast, respectively. ARPSa refers to ARPS 3DVAR and cloud analysis. * For all ARW members: ra_lw_physics= RRTMG; ra_sw_physics=RRTMG; cu_physics=none*

Member	IC	BC	Radar	Microphy	LSM	PBL
Member	IC	BC	Radar data	Microphy	LSM	PBL
arw_cn	00 UTC ARPSa	00 UTC NAMf	yes	Thompson	Noah	MYJ

arw_m2	arw_cn + arw-p1_pert	21 UTC SREF arw-p1	yes	NSSL	Noah	YSU
arw_m3	arw_cn + arw-n1_pert	21 UTC SREF arw-n1	yes	NSSL	Noah	MYNN
arw_m4	arw_cn + arw-p2_pert	21 UTC SREF arw-p2	yes	NSSL	Noah	MYJ
arw_m5	arw_cn + arw-n2_pert	21 UTC SREF arw-n2	yes	Morrison	Noah	YSU
arw_m6	arw_cn + nmmb-p1_pert	21 UTC SREF nmmb-p1	yes	Morrison	Noah	MYJ
arw_m7	arw_cn + nmmb-n1_pert	21 UTC SREF nmmb-n1	yes	P3	Noah	YSU
arw_m8	arw_cn + nmmb-p2_pert	21 UTC SREF nmmb-p2	yes	P3	Noah	MYNN
arw_m9	arw_cn + nmmb-n2_pert	21 UTC SREF nmmb-n2	yes	Thompson	Noah	MYNN
arw_m10	arw_cn + arw-n3_pert	21 UTC SREF arw-n3	yes	Thompson	Noah	MYJ
arw_m11	00 UTC ARPSa	00 UTC NAMf	yes	Morrison	Noah	MYJ
arw_m12	00 UTC ARPSa	00 UTC NAMf	yes	P3	Noah	MYJ
arw_m13	00 UTC ARPSa	00 UTC NAMf	yes	NSSL	Noah	MYJ
fv3_m14	GFS	-	yes	Thompson	GFDL	GFDL

fv3_m15	GFS	-	yes	NSSL	GFDL	GFDL

Building upon the 2017 FFaIR Experiment, the SSEFX again produced an experimental Localized Probability Matched Mean. The localized probability-matched mean (LPM) calculates the probability-matched mean over small patches (typically 10×10 or smaller) of the domain, using calculation regions with substantial overlap (typically around 60×60 grid points for each 10×10 patch), and then smooths the resulting field with a Gaussian smoother. The result is a forecast field that provides many of the advantages of the probability-matched mean (PM) while retaining small-scale structures in the resulting LPM field that may be informative or of meteorological interest. The LPM also does not suffer from potential errors resulting from considering all precipitation from a full CONUS domain for each smaller patch, limiting the influence to the nearest 100 km or so, ensuring that values used are from local storms and the local near-storm environment.

An example is shown in Figure 71, for a 3-hour rainfall forecast where rain was present over much of the southern Great Plains. The PM field (Fig. ()a) exhibits a typical highly-smoothed distribution of rainfall amounts; this is typical of PM forecasts. In contrast, the LPM field (Fig. ()b) retains much more small-scale structure, particularly for storms in Kansas, and the predicted rainfall in Wisconsin.

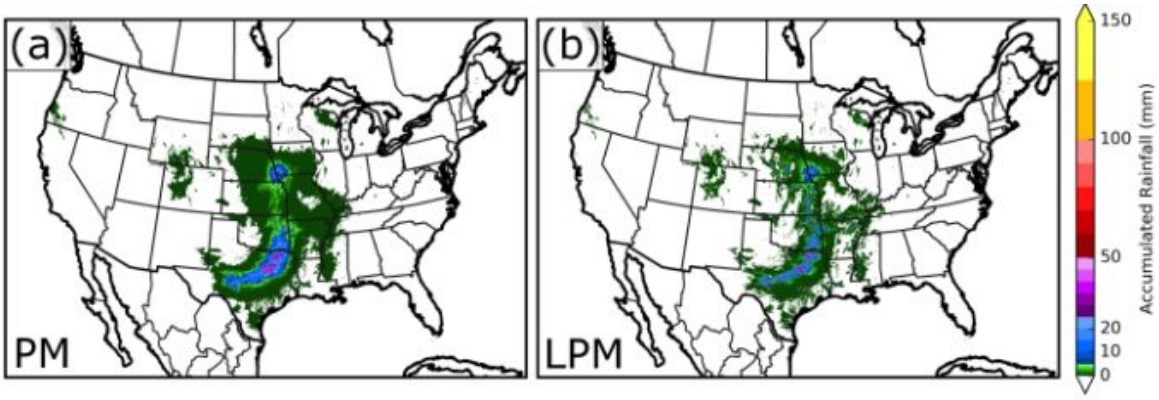


Figure 71*. Sample PM (a) and LPM (b) means from the SSEFX for 3-hour accumulated rainfall forecasts.*

Various QPF exceedance probabilities were also available in the SSEFX and shown in (Table 14).

Table 14. *Probabilities available from the SSEFX during FFaIR 2018*

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

***National Blend of Models Version 3 (NBMv3)*:**
The National Blend of Models (NBM) is a nationally consistent and skillful suite of calibrated forecast guidance based on a blend of both NWS and non-NWS numerical weather prediction model data and post-processed model guidance. The goal of the NBM is to create a highly accurate, skillful and consistent starting point for the gridded forecast. This new way to produce NDFD grids will be helpful by providing forecasters with a suite of information to use for their forecasts.

Probability of Precipitation (PoP) forecasts the probability that a threshold amount (usually the equivalent of 0.01 inches) of precipitation will fall during a time period, while the Quantitative Precipitation Forecast (QPF) predicts the amount of liquid-equivalent precipitation expected to fall over a time period. For NBM products, this time period is a number of hours before the specified forecast time. (i.e. PoP12 valid at 12 UTC is the probability there will be precipitation from 00:00 UTC-11:59 UTC of that day; PoP12 valid at 00 UTC is the probability there will be precipitation from 12:00 UTC-23:59 UTC of the previous day. Note that for the NBM text products all dates/times are in UTC time.)

Forecasts of 6-hourly quantitative precipitation (QPF06) are produced as follows: 1) Form a grand ensemble mean, interpolated to ⅛°. 2) Quantile map the mean forecast using CDFs of the ensemble mean and analyzed distributions. 3) Spatially smooth the field (Hamill, 2017).

Building upon the major upgrades in the NBMv3, the 2018 FFaIR Experiment will feature the **experimental NBMv3.1**, which is currently in a parallel evaluation and scheduled to become operational in August, 2018**.** The Hamill method described above is applied to the GFS, GEFS (0.5 deg), GDPS (0.25 deg), GEPS (0.50 deg), NAMNest, RDPS (10km), REPS (15km), SREF (16km), NAVGEMD (0.5 deg), NAVGEME (1.0 deg), ECMWFD (0.25 deg), ECMWFE (1.0 deg). The training ground truth is 2.5km RFCQPE/URMA using 60 day archive and is now updated 4 times a day at the 00z, 07z, 12z, and 19z cycle times. Models included in the NBMv3.1 are found in Table 15.

***Table 15**. Data Dependencies for the NBMv3.1, which will become operational in Fall 2018.*

QPF06 Weights (%) V3.1 (Operational August 2018)								
Projections 6-12		**Projection 18**		**Projections 24-36**		**Projections 42-48**		**54 +**
Model	Weight	Model	Weight	Model	Weight	Model	Weight	
QMD (Hamill)	12	QMD	15	QMD	20	QMD	50	100
HRRR	40	HRRR		HRRR		HRRR		
HRRRX	10	HRRRX	27	HRRRX	20	HRRRX		
RAP	5	RAP	5	RAP		RAP		
RAPX	3	RAPX	3	RAPX	5	RAPX		

HiResW NMMB	5	HiResW NMMB	10	HiResW NMMB	10	HiResW NMMB		
HiResW WRF	5	HiResW WRF	10	HiResW WRF	10	HiResW WRF		
HiResW Mem2	5	HiResW Mem2	10	HiResW Mem2	10	HiResW Mem2		
NAM	5	NAM	5	NAM	5	NAM	5	
NAMNest	10	NAMNest	15	NAMNest	20	NAMNest	45	
Total	**100**	**Total**	**100**	**Total**	**100**	**Total**	**100**	

3.3 Other Experimental Forecast Tools

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

Developed by Greg Herman and Russ Schumacher of Colorado State University, this first-guess field is a prediction system using the random forest machine learning algorithm. The model is trained to probabilistically predict 1- and 10-year NOAA Atlas 14-based ARI exceedances across CONUS for Days 1-3 based on an extended record of comparisons between historical model forecasts and precipitation observations. The model synthesizes many different aspects of forecast information, including the local precipitation climatology (as quantified by the ARIs), an ensemble of model QPFs, simulated environmental conditions and convective ingredients (CAPE, CIN, PWAT, surface temps/winds, etc.), and the spatiotemporal evolution of those fields throughout the forecast period to produce a single starting point from which to base an ERO or similar product. Unique random forests are trained for eight different regions of CONUS, partitioned based on geography and hydrometeorological characteristics. Forecasts are produced for each point within each forecast region, and then blended together to form a single, cohesive CONUS-wide product which predicts the probability of a 1- or 10-year ARI exceedance occurring within 40 km of the given forecast point over the forecast period. An example forecast for Day 2 1-year ARI exceedances for 24-hour 12-12 UTC precipitation accumulations is depicted in Figure 72. For more detail on how these products are produced, see Herman and Schumacher (2018 a,b).

Model forecasts are produced for several different lead times and accumulation intervals. For short term, Day 1 forecasts, three ARI exceedance probability products are issued: 1) the 24-hour 12-12 UTC period, 2) the 6-hour 18-00 UTC period, and 3) the 6-hour 00-06 UTC period. For each of these fields, the First-Guess Field is based on the 00 UTC initialization of the HRW-NSSL member of the operational convection-allowing HREFV2 ensemble. The random forest statistical models for these three fields are trained using historical daily 00 UTC NSSL-WRF forecasts from June 2009-September 2016. For Days 2 and 3, single 12-12 UTC ARI exceedance probability fields are issued for each forecast day. The model at these lead times is trained using 11 years (January 2003 - August 2013) of Day 2 and 3 GEFS Reforecast (GEFS/R; Hamill et al. 2013) data. Two different versions of these models are being run quasi-operationally. One is based on the 00 UTC initialization of the GEFS/R, while the other has the same underlying statistical model (trained on the GEFS/R), but uses inputs from the

operational GEFS as model predictors; this model runs on both the 00 and 12 UTC version of the GEFS.

The first-guess field for the ERO was evaluated in the 2018 FFaIR Experiment at Days 1, 2 and 3 for its utility in predicting the slight, moderate and high probabilities of flash flooding.

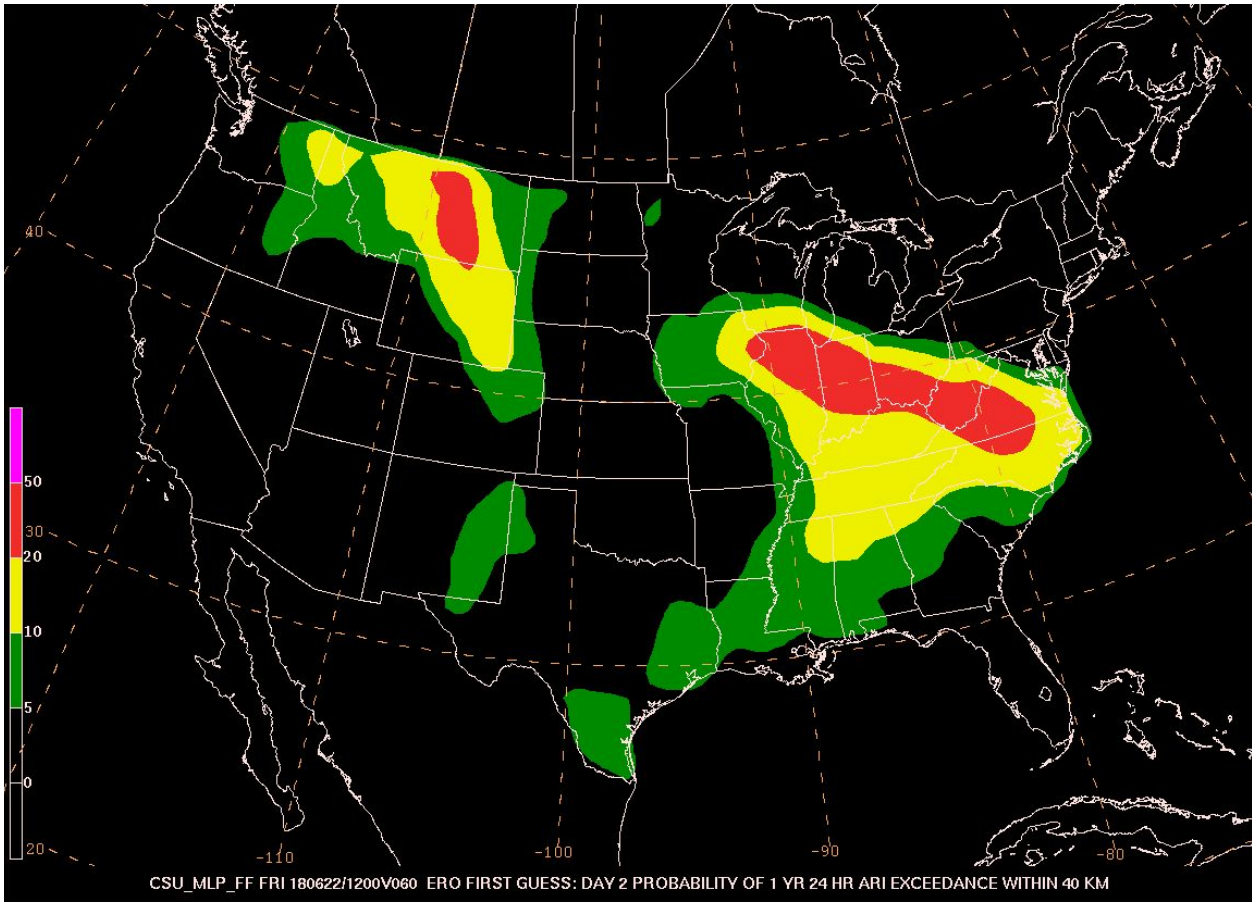


Figure 72. *An example of the CSU-MLP ARI exceedance product to be tested as a first-guess field for the experimental Day 1 Excessive Rainfall Outlooks. Here shown is the Day 1 probability of 1 year 24 hour ARI exceedance within 40 km of a point.*

Experimental Remotely-Sensed Products

CIRA Layered Precipitable Water and Model Difference Products:

CIRA has developed an experimental four-layer layer (sfc-850, 850-700, 700-500 and 500-300 mb) Advected Layer Precipitable Water (ALPW) product which is distributed to WPC, NHC and select WFO's every three hours in AWIPS-II format, with the assistance of NASA SPoRT. This research is supported by the JPSS Proving Ground and Risk Reduction initiative. Passive microwave water vapor profile soundings from seven spacecraft (Suomi-NPP, NOAA-18/19, Metop-A/B, and DMSP F-17/18) are combined with GFS model winds to depict layered precipitable water amounts over CONUS and adjacent oceans. The retrievals are produced by the NOAA operational Microwave Integrated Retrieval System (MiRS). This product has been used widely by WPC in Mesoscale Precipitation Discussions to detect long-distance tropical connections of water vapor for flash flooding. ALPW allows forecasters to determine the depth

of water vapor and in particular whether mid- and upper-level moisture is available to enhance precipitation. It complements the operational blended TPW product by adding a vertical dimension. ALPW was evaluated in the 2018 FFaIR experiment. An example is shown in Figure 73. Animations of ALPW are available here:
http://cat.cira.colostate.edu/sport/layered/advected/LPW_alt.htm

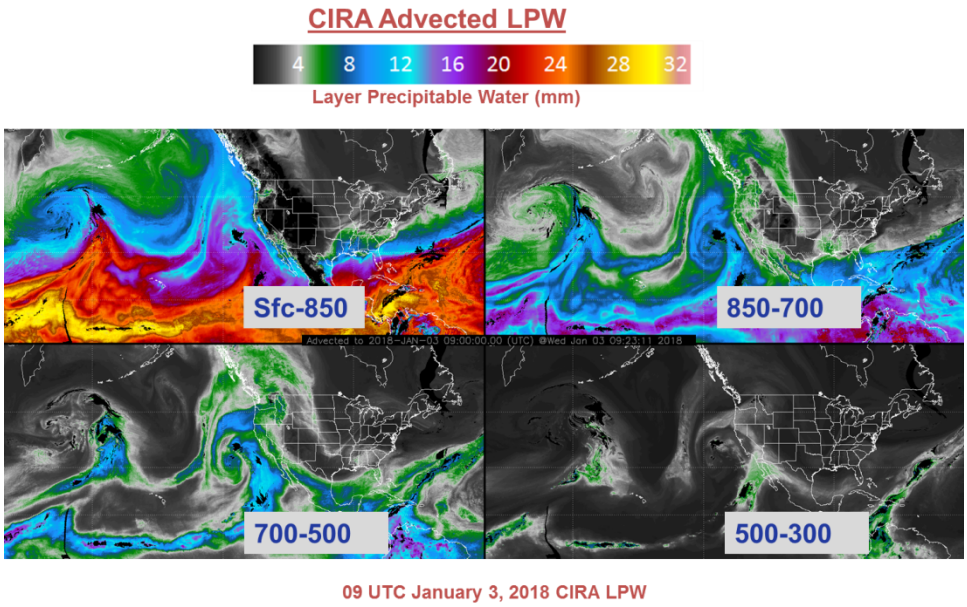


Figure 73. *CIRA Advected LPW for four layers at 09 UTC January 3, 2018.*

Since the ALPW water vapor retrievals are independent of NWP models, the question arises as to how well ALPW and NWP water vapor fields agree. ALPW uses all of the passive microwave data available, while NWP data assimilation approaches may thin or reject valid data, especially in cloudy areas. Toward that end, a model water vapor difference product has been created under the NOAA HMT. LPW is derived from the water vapor profiles in the GFS and HRRR models, and difference maps are created for the four ALPW layers. The LPW derived from the GFS 3 hour forecast for the time matching Figure 73 is shown in Figure 74, and their difference for the four layers is shown in Figure 75.

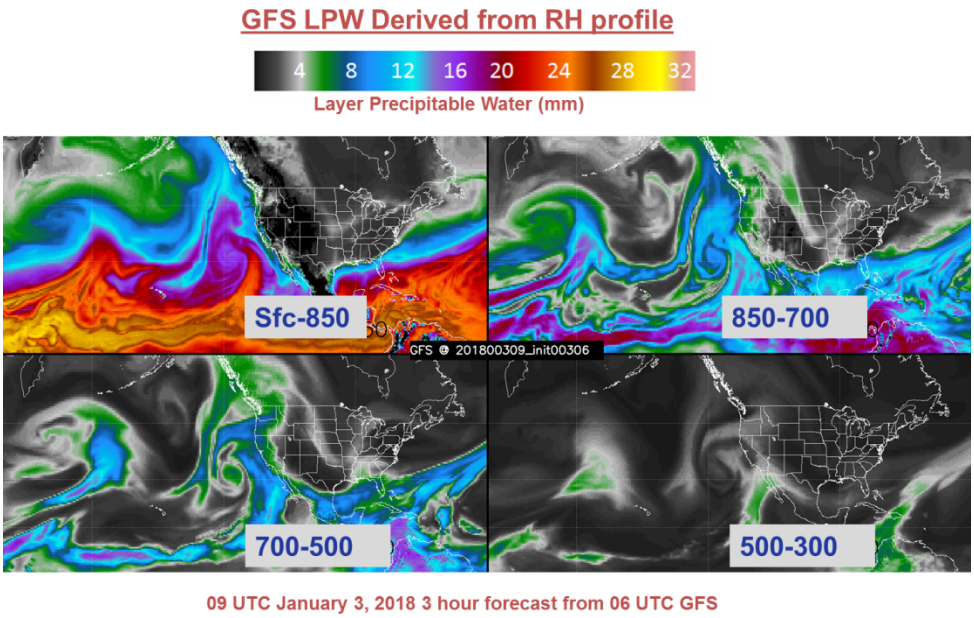


Figure 74. *LPW derived from GFS for 09 UTC January 3, 2018.*

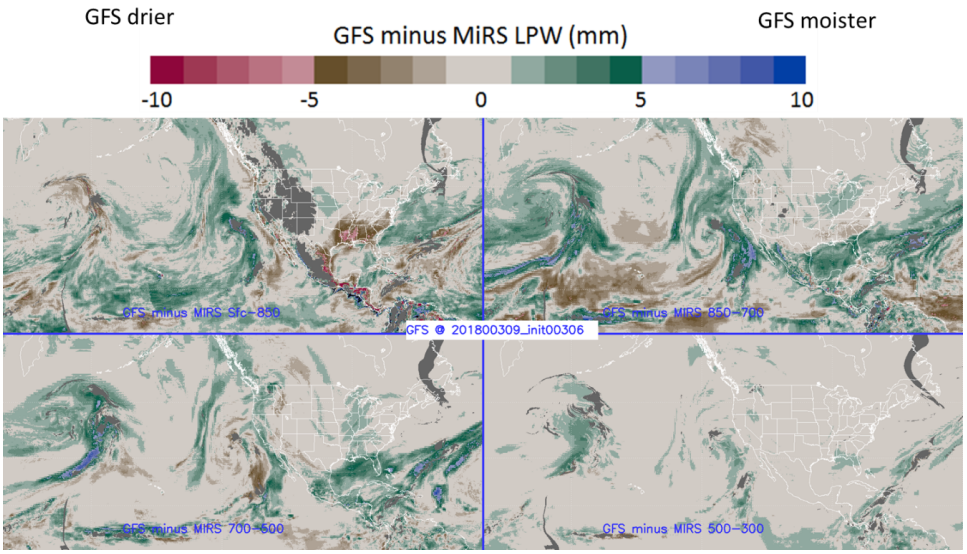


Figure 75. *LPW difference product for the four layers shown in Fig. 1 for 09 UTC January 3, 2018.*

The difference maps for the GFS are available here, and are updated every six hours: http://cat.cira.colostate.edu/hmt/hmt_main.htm

GOES-16 Products for Evaluation:

Derived Stability Indices

The Derived Stability Indices such as Convective Available Potential Energy (CAPE), Lifted Index (LI), Totals Total (TT), Showalter Index (SI), and the K-Index (KI) are computed from the

retrieved atmospheric moisture and temperature profiles. These indices aided forecasters in nowcasting severe weather by providing them with a plan view of these atmospheric stability parameters. Forecasters used this information to monitor rapid changes in atmospheric stability over time at various geographic locations, thus improving their situational awareness in pre-convective environments for potential watch/warning scenarios.

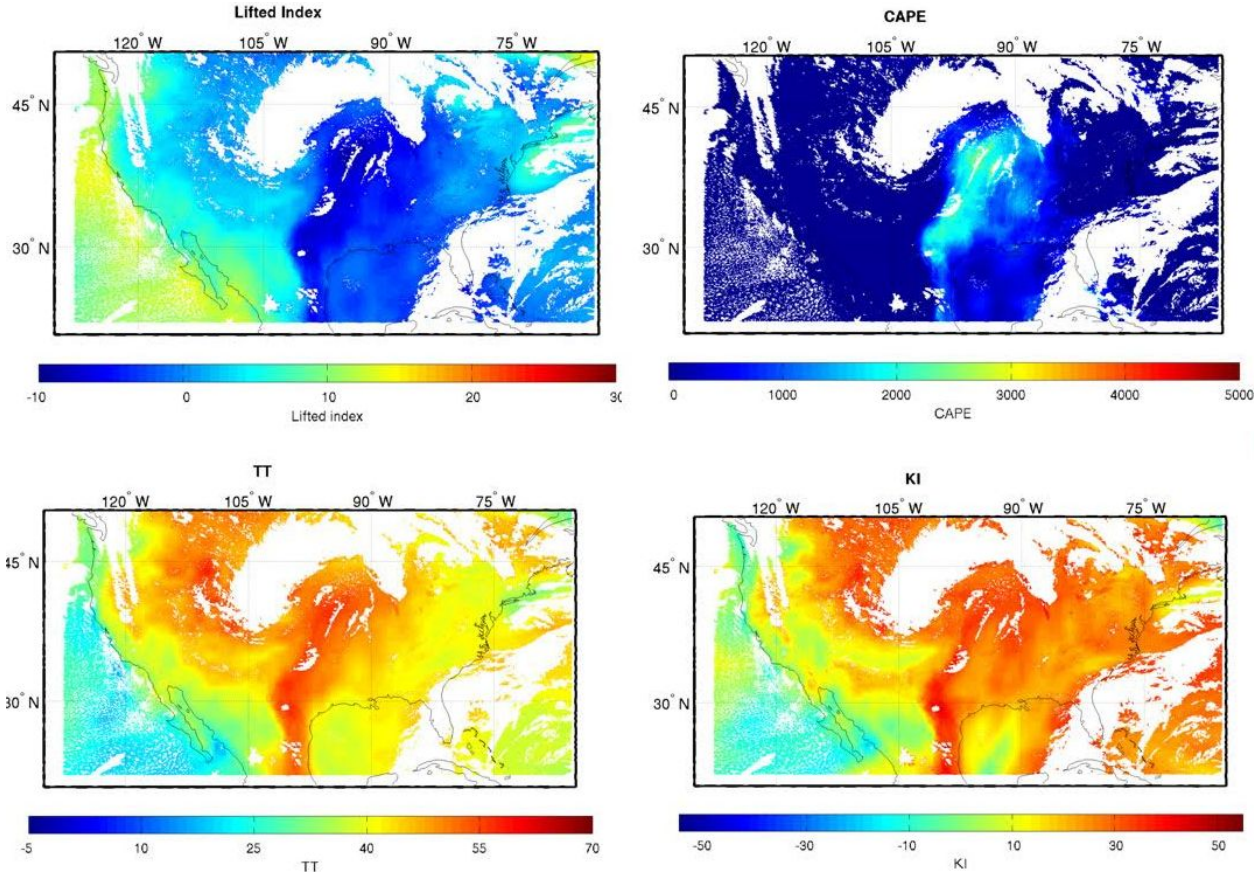


Figure 76. *An example of the satellite-derived instability indices from GOES-16.*

TPW

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

APPENDIX D

WPC MODE Settings for Objective Verification

- 36 HR & 24 HR QPF verified against Stage IV QPE
  - 00 UTC forecast cycle used
  - Both QPF and QPE re-gridded to a common 5km lat/lon grid
  - CONUS mask applied to common grid
  - Thresholds of 0.5”, 1.0”, 2.0”, 4.0” and 6.0” investigated
- MODE
  - Grid stats harvested from MODE CTS
  - Circular convolution radius of 3 grid squares used
  - Double thresholding technique applied
- MODE Analysis
  - Summary of all forecasted vs. observed shapes throughout experiment
  - Describes centroid distance, angle, and interest